Ca' Foscari
University
of Venice

Master's Degree programme

in International Management

Final Thesis

HR Analytics driving business outcomes:

A descriptive and predictive Case of Analysis

Supervisor
Ch. Prof. Pontiggia Andrea

Graduand

Matteo Miani

Matriculation Number 862080

Academic Year
2020 / 2021

To Bea, for always believing in me like nobody else;

To Elio, for the spontaneous moments together;

To Matteo, for unleashing my hidden powers;

To Mum and Dad, for giving me this opportunity.

Table of contents

Introduction

In today's volatile economic settings, competition has greatly intensified. Among the ways to achieve long-term success, organizations should shift the HR role from an operational partner with administrative tasks to a core center aligned with their overall business strategy and impacting the business performance. In addition, in an environment in which analytical processes and capabilities are rapidly reshaping industry competition, the importance of digital technologies has increased even inside HR departments, leading companies to find better ways to quantifying, monitoring and forecasting the most important employee behaviors influencing the overall performance. In particular, this master thesis will then examine the discipline of "Human Resource Analytics" (HRA) defined as the use of data analytics in the field of human resource development. Though its rise in popularity is accompanied by skepticism about the ability of HR professionals to effectively utilize employee data to reap organizational benefits, in this work it will be emphasized how implementing such HR Analytics practices would help companies to address the primary HR challenges of recent times (such as high turnover rate, low levels of individual performance, ineffective talent retention programs and many more). More precisely, in the first chapter, it will be provided a review of the main issues around the debate on HR analytics, both supportive and critical to this discipline, reporting its main premises and objectives: synthetically, making operational HR processes more data-driven and result-oriented, changing company leaders' approaches to manage their employees by enabling them to make more informed workforce decisions. Initially, it will be explored the concept of HR Analytics in general, along with the most impacting tecnhnologicial forces that have driven its rise as an unfolding HR trend, finally discovering the reasons why is so urgent and beneficial to for organizations, when dealing with workforce problems, to promptly embrace such data-driven approach and supplement HR intuition with objectivity given by a fact-based decision making; then it will be presented some barriers to implement HRA practices and the current limits like the worrying ethical debate generated around it regarding employee data privacy and the algortihms' perpetuation of discriminatory biases. Over the second chapter, it will be first outlined how the HR data ecosystem, in this ever-increasing digitalized world, has radically changed over the last decades with the advent of “Big Data”, that have allowed organizations to leverage massive amounts of data from a growing number of sources. Then, it will be investigated the traditional overall framework of HR procedures and protocols used to treat and interpret available employee data: this scenario has been enriched with new complex free-decision mechanisms techniques that have changed the way HR reads and analyzes personnel data. Accordingly, it will be stated that many companies do not take advantage of the availability of more sophisticated computational methods: HR departments should be much more evidence-based and much less Tayloristic as, nowadays, simple employee workloads are not the dimensions HR is most interested to analyze to gain valuable insights on its labor force. Other than that, a specific section will be dedicated to address the fascinating world of machine learning, by explaining in detail the most widespread algorithms that I will further adopt for my analyses. Importantly, most of the literature that deals with this topic gives little hints about how to translate HRA concepts into practice. For this reason, as this thesis’s goal is to make readers conscious of the potential of HR Analytics process in the management and development of human resources, in the third chapter, it will be directly illustrated the implementation of quantitative descriptive and predictive techniques, by providing empirical cases on the field, applied to real-world company's problems. Regarding the descriptive part of analyses, intuitive HR dashboards will be developed along with some clustering techniques, while, in the more advanced predictive analyses, with the use of the programming language of Python, it will directly analyzed a dataset, realistically simulated, through intelligent machine-learning algorithms on the front-line areas of interest of HR: employee engagement, turnover risk, performance appraisal and son on. Overall, the analyses will suggest how this type of data-driven approach, which helps identify factors deeply affecting employees behavior, facilitates the creation of a sustained and high-performance ecosystem within an organization, thus increasing the productivity of the employees and in turn increasing revenue generation. Eventually, the very last section will regard the interpretation and the limits of this study as well as the HRA implications, investigating how companies could adjust their HR initiatives according to their analysis's findings.

CHAPTER ONE

UNDERSTANDING THE HRA PHENOMENON

1. HRA concept and purposes

Throughout this first section, it will be explored the world of HR Analytics, mainly providing a brief description of the HR setting in which its practices are being implemented and then explaining the phenomenon by starting from its various definitions to winding up with its ambitious purposes.

1.1.1 HR as a strategic business partner

Human Resource Management has always played a decisive role in any organization, as it is a function specifically devised to capitalize on the employee's performance to reach the company's strategic goals. Ever since the 1980s, thanks to advances of new technologies, the role of HRM has progressively shifted from administrator of obligatory HR activities to a more strategic position, necessarily aligned with overall business strategy, by acting as a strategic business partner (Ulrich and Dulebohn, 2015). Also, nowadays, along with the prevailing economic set-up characterized by globalization trends and changing business dynamics, it has become increasingly crucial to focus on how to best improve and use an agile and highly-competent workforce for generating returns for the business, while preserving cost efficiency. As it's apparent that reaching a great level of efficiency within the HR department and having a consolidated HR structure is no longer enough to meet the current business challenges, organizations are getting aware of their need to maximize the return on human capital, the organization's most valuable asset, through innovative methods. For this reason, the Human Resources department, responsible for dealing with all things related to people, has been called to manage its workforce in a way that reflects the strategic purposes of the business. More precisely, the evolution path identified regarding the HR departments aims to elevate it to the centralizing role of personnel management and key support the business processes, by creating the best conditions to enhance their people leveraging data available inside firms. Within this ambitious scope, a rising solution to this challenge is given by the practice of exploiting the complex interaction between all factors concerning the internal workforce at disposal: staffing levels, personnel profile, competencies, compensation benefit structures, training programs and so on. The most critical issues being addressed by HR today are in fact such as high turnover and absenteeism rate, low employee performance, and many more.

1.1.2 Defining HRA

Human Resource Analytics (HRA), also called 'People Analytics', 'Talent Analytics', or 'Workforce Analytics', is a relatively new term that first appeared in the academic literature in 2004 (Marler and Boudreau, 2017) and since then, it has been interpreted in a variety of ways (Bassi, 2011). One first approach highlights the distinction between HR Analytics and HR metrics, arguing that analytics represents statistical and experimental techniques applied to demonstrate the effect of HR activities on the performance of a company (Lawler et al., 2004). Later in time, the definitions became more general, describing the term as a process that either focuses on analysis or on decision-making. Differently, Harris et al. (2011), and Falletta (2014) directly provide a definition hinged on different types of analytical processes and HR practices whose research was based on. Certainly, there was an ambiguous lack of agreed definition that is quite surprising, given that the discipline has been developing for over a decade. Nonetheless, it logically derives from the definition of analytics as to the "intersection of computer science, decision-making, and quantitative methods to organize, analyze and explain the increasing amount of data generated by modern society" (Mortensen, Doherty, and Robinson, 2015). Even if some of the researches (e.g. Angrave and Pape, 2016) only provided a general definition of business analytics without a specific link to HR, attaching aside the HR component to analytics, it refers more accurately to the identification and use of analytical techniques and statistical methods such as data mining, predictive and contextual analytics to interpret people-related data and HR systems and processes. HR analytics, then, represents a powerful pool of activities that offer the chance of enhancing the effectiveness and efficiency of every aspect related to the HR department through logical and numerical explanations (H.H.D.P.J. Opatha, 2020).

Another interesting comprehensive definition, provided by the school of management "Politecnico Di Milano", claims that the term "HR Analytics & Big Data" indicates the set of skills, analysis and visualization tools and information sources that allow to capitalize, manage and analyze data relating to personnel with the aim of improving the impact of people strategy on the business through greater decision-making and strategic support in terms of acquisition, management, development and retention of people. From this latter definition, it becomes crystal clear how HR managers are facilitated to make better workforce-related data-driven decisions by leveraging an array of tools and technologies enabled by information technology. Thus they are given a great chance of managing their employees in a more efficient way and of relieving the solutions to business problems. In fact, this evidence-based and technology-driven approach has the power to transform the HR department with full automation, releasing it of its administrative burden and driving a bigger contribution at its strategy side. Having Big Data finally reached HR, the outdated perception of this function, overwhelmed by the too many responsibilities, is being overtaken and left to the past by organizations, who are discovering the new technologies elaborated according to the principles of People Analytics. In particular, by leveraging the raw HR data that provides insightful information to the strategy formulation, HR managers not only are enabled to take fact-based decisions but are enabled also to justify the investments made to the human resource projects, to see if they're paying off and also to predict future outcomes. Therefore, capitalizing on the insights acquired, organizations move forward and stay ahead of the competition, on the promise of building competitive advantage by delivering actionable business intelligence outcomes on the people's sides, thus attracting, retaining, and improving their talents and preserve the success in the long run (Reena et al., 2019). In addition, as HR Analytics assists firms to define the future with accurate predictive analyses, organizations are developing a more proactive role in driving business strategy, getting firms to align their People Analytics offices with strategic organization areas that were once considered far away from the HR domain. We are referring to areas such as, for instance, sales effectiveness, culture, and risk, conduct and compliance. In fact, HR is now impacting the business results and adding value as it is capable, with these analytical instruments, of making workforce decisions that lead to a reduction of costs, identification of viable revenue streams, mitigation of risks and sheer formulation of corporate strategy, thanks to decisions based on evidence instead of intuition or personal experience (Rasmussen and Ulrich, 2015). What's more, aside from the basic purpose of People Analytics policies and processes of allowing major firm stakeholders to measure and report key workforce trends, such as employee well-being, there's more. Other than to recruitment & selection, performance management and employee engagement, these practices have been increasingly giving different importance to supplementary topics that have grown a business-centric vision and approach in the last decade: these HR specific fields correspond to organizational design, talent retention, staff turnover, training & development. and diversity & inclusion.

1.1.2 A new mandate: from HR reporting to analytics

Importantly, if we are to discuss the realm of analytics in the human resource world, it's misleading to not have clear in mind the natural distinctions between reporting, metrics and analytics (Fink and Sturman, 2017). While HR reporting, being considered the simplest level, regards capturing elementary facts at the current state about an organization or team aiding managers and leaders in tracking and managing their workforces, HR metrics make a step further in the analysis. Such metrics, including, for instance, compensation, turnover and employee engagement ratios, are essential for HR employees to evaluate how efficient, effective and impactful the HR programs and processes put in place. At the greatest level of complexity and sophistication, we find the analytics branch, whose underlying goal consists of identifying high-end forecasting patterns (e.g., "what-if" scenarios that predict the consequences of changing conditions) that can inform a more conscious decision-making process. We will go much more in-depth throughout the reading in this issue to gain a fundamental understanding of how differently these three different layers of analysis work and why and when HR practitioners should rely on them. For now, explaining it plainly with a diversity and inclusion's practical example could give a sort of helpful preview. Taking an organization, where, as a first step, is reported the proportion of women promoted at each level of seniority within a specific period of time, in order to understand the wider gender picture within the organization. At this initial phase, Talent Analytics' purpose is to carefully identify the problems in the business - what is going on inside the company - and formulate the strategic solutions to cope with them. Afterward, only having these data arranged at hand, indeed is possible to apply and design a precise array of metrics for managers to closely track the effects of subsequent HR programs, purposely implemented to increase the diversity rate inside the workforce, based on the visualization of data reported before, and to check whether they work or not. Eventually, analytics should complete the process by contributing to the ultimate purpose of improving diversity through the use of predictive modeling. These techniques give a direction to follow depending on the future scenarios concerning the diversity situation of the company, so as for managers to know in advance how to envision better solutions. As a matter of fact, comparing the available HR data (like payroll, performance appraisal, attitudinal survey and absences) with demographic and gender data, it is possible, for example, to identify the biased fallacies of the promotion process for women, along with the underpinning and hidden factors leading females to leave the organization (e.g., not getting a promotion in the first two years). This greatly helps HR senior executives in proactively driving changes in the actual policies adopted and, above all, in choosing which employees deserve a promotion the most.

To conclude, despite the rising recognition of the HR Analytics potential to create value, very few organizations actually rely on a decision-making process based on unbiased facts or objective deliberations (Fitz-enz & Mattox, 2014), and we are going to investigate

1.2 The rise in HRA adoption

HR Analytics represents a fast-growing discipline in HR that has been developed over the last two decades, a time in which the organization landscape has greatly changed, and it doesn’t yet seem to stop. Mainly, the most influencing transformations have corresponded to the recent incoming waves of globalization, the new perception of human capital, and the spread of the newly sophisticated technologies. These represent also the main issues addressed in this section, which will be accompanied with some figures regarding the HR Analytics rising trends.

1.2.1 Driving factors behind the rise of HRA

Even though the first time that a relationship between an organization’s performance and its investment in human workforce was taken into consideration dates back to more than 50 years ago (K. King, 2016), the widespread of data analytics practices associated with the HR field has taken off solely in the last two decades. More specifically, as anticipated, from the early 2000s on, the outcome of these joining forces has been associated with the key recognition of the worth of today's company's intangible assets - including human capital - that lines up to more than 70% of its total value. To understand this major change, it’s worth mentioning the difference with the 1970s, when tangible holdings reached plainly the tune of 95% of a company’s assets value. Therefore, this new increased perception attributed to human capital has proven to be a first factor for the rise of HR analytics, whose main promise is to better invest the limited resources available and to add value to the process of managing a workforce. Furthermore, the overall scenario for treating HR data is radically changed, as will be extensively explained over the next chapter, so have the technological equipment that companies were found to rely on. In fact, though lots of underpinning statistical techniques have been around for decades, the radical shift to analytics was favored, essentially, by the rise of the so-called “Big Data'': nowadays, as the volume of data available has changed, so has the processing power of the analytical instruments at disposal of the companies. Likewise, analysts could now rely on a much wider variety, availability and affordability of user-friendly analytical tools that are capable of storing, accessing and analyzing both structured and unstructured data. In addition, if it is true that over the last decade, the adoption of technology has passed from static and secondary business-supportive HR management solutions to more dynamic, cloud-based tools and platforms, this has occurred also because corporations have been grasping strong evidence of the efficacy and efficiency of predictive modeling techniques as it was highlighted before. Overall, from recruiting to hiring to performance evaluations, HR executives have been investing in tech-driven data analysis to make better people decisions. Making an example, in the recruitment area, building a strong pipeline of best candidates, and arranging actions to subsequently retain the ones successfully hired, has proven to be decisive to remain competitive in the future. Hence, boards and executives are surging their demand and acceptance for an HR Analytics department and team spaces within their firms, leading HR software vendors to enhance their investments in providing the most sophisticated analytics platforms. However, to gain a competitive advantage inside such an increasingly competitive landscape, relying on a promising talent pool is not enough. Identifying future organizational needs in terms of size, structure and future people necessary to achieve the corporate mission and the financial objectives is key. There is indeed a deep need for businesses to predict where they may navigate in 10-15 years by leveraging analytics and to get their HR function more quantitative in nature. To conclude, in today's times, in spite of the pandemic crisis and international economic uncertainty, it's clear the need to maintain momentum following the HR Analytics mandate as a potential source of long-term success.

1.2.2 The three waves of analytics

Noticeably, growth in technology and innovation represent the two most critical factors behind the escalation of Workforce Analytics. Over the last decade, technology adoption in the second half of the 2010s was firstly marked by the pioneering analytics experiences developed by few experts and data scientists. Thanks to the use and development of more advanced tools to answer common questions like "what has been the whole revenue for Q1?", they paved the way for the 'First Wave of Analytics', which presented a focus on cloud-based core HR systems. Newer levels of analysis and the curiosity to delve deeper into data with the purpose of answering different types of questions using analytical instruments, required companies to retain analytics experts and support their formation with upskilling training programs. Cloud-based IT solutions also favored the initially slow 'democratization' of these analytics practices to a broader community of people, accelerating with the rapid pace of business change. After the step was set for the 'Second Wave', characterized by the introduction of advanced business intelligence technology for Workforce Analytics, analytical knowledge passed thus from few expert people's hands to a larger set of reporting specialists with power-user tools, with the rise of the "Data Scientist" role. In this framework, while an average analyst would be able to master the less sophisticated and complex tools without being considered a highly-trained expert, there was still a significant gap in expertise with the normal business employee that, lacking the technical necessary background, was not in power to perform analyses or reports on his own. This had the effect of tremendously slowing down the process of advancement in analytics, as the person who requested the data insights oftentimes didn't know what to do with the information at hand: increasingly, it became apparent the need for all employees to get insights into real-time, by accessing the evolved analytics capabilities quickly and easily at the point of work. Thus, HR Analytics was perceived as a tool of an elite of few companies with the necessary pool of resources to build complex systems inside their departments and able to invest in the technology and expertise required to support analytics (Falletta 2014). As of today, non-technical business persons are now empowered to take decisions by themselves as insights are embedded directly within business processes and are way more accessible. This was made possible thanks to the advent of the 'Third Wave of Analytics' that was characterized by a massive proliferation of the latest generation of HR systems, which is making it simpler for HR professionals to analyze data and present the findings in a visual, comprehensible way to executives. Analytics is then acting as a game-changer for HR, ensuring HR professionals they're not being misled by superficial patterns but they base their HR policies and decisions on reliable foundations and on an effective and strategic measurement system of the HR data that is strongly connected to the business results (P . Reddy & P. Lakshmikeerthi, 2017).

1.2.3 Exploring recent HRA trends

According to research by the Corporate Research Forum, 69% of organizations with 10,000 employees or more now have a People Analytics team. The reality is perhaps not so rosy, as in my experience, many of these teams are still essentially restricted to reporting, and not really doing analytics. In fact, a trend is recognizable among diverse levels of business: large companies do rely more on HR analytics. According to the figures of an MHR Analytics report, if only 6% of companies surveyed with more than 250 employees confessed to not performing any reporting or analytics and to not have data analysis roles within HR, firms having less than 50 employees accounted for the beauty of 48%. This doesn't necessarily guarantee, for what concerns larger organizations, that analytical practices are being adopted across the whole business as an important strategic priority since most of the executives interviewed declared that they were unaware of their HR functions having specifically a focus in analytics; still, these numbers put them in an advantageous position in contrast to smaller realities. In addition, from a research run by Andrew McAfee and Erik Brynjolfsson of the MIT (Massachusetts Institute of Technology), it has been found that companies with a solid data-driven vision have shown more successful results than those not equipped of these characteristics, figuring, on average, 6% more profitable and 5% more productive than their industry-like competitors. It's not a case then if the best examples in the HR analytics field typically come from industries with a pronounced technical, scientific or data orientation, such as retail, hi-tech and biotechnology and other high-tech industries. Another important finding from an IBM's 2016 survey, shows that, over the previous two years, the number of CHROs reliant on predictive analysis for addressing their decision-making process has raised by approximately 40%, whereas, according to a 2015 Deloitte's survey, of about 3,300 business leaders and HR executives from 106 countries across the world, the vast majority of respondents (75%) were confident of the profit-generation potential of HR Analytics to the overall collective and individual performance level and considered the function of significance importance. Having said that, this third wave of analytics progress, accompanied with the driving factors discussed above, is testifying statistically how technology investment in the People Analytics departments is steadily increasing. 54% of organizations surveyed in 2020 by a research of "Insight 222" have declared their intention to augment their expense on technology, making it much easier to scale the use of analytics to multiple job roles, functions, and capabilities; rather, 37% stated to currently implement specialist technologies for particular People Analytics solutions.

In this direction, the proliferation of analytics in various other functions, in comparison to the laggard HR, has brought to a wider understanding of what could be learnt from different kinds of data, leading to deeper connections among offices, which is essential to reach an informed decision-making. It's worth mentioning the central use of analytics in areas such as marketing, finance and customer service, for instance, which HR is currently playing to catch up. It's Interesting also to notice, from the MHR Analytics report, that only 23% of respondents stated to use insights across other business areas to prevent HR from making stand-alone decisions. The finance and accounting industries are the most virtuous examples, committed to conduct strategic or financial planning along with HR. In parallel, other than the figure of “Business Consultant”, inside the HR Analytics departments are emerging and are being incorporated into teams new roles like “Data Scientist”, which presents the highest predicted growth. According to Insight 222’s survey, 57% of companies interviewed forecasted a surge in headcount, suggesting a switch to undertaking more advanced and sophisticated analytics. Also, the “Data Architect” and “Dashboard Developer” roles seem constant in their positive trend, with the perception for the great majority of companies (72%) reporting either no growth or a drop in demand for this role, besides the 58% of the organizations surveyed affirming that they already have this role inside their organization’s walls. Even People Analytics team size is jumping, with 60% of firms found either to have plans to increment their HR Analytics teams. Most importantly, the fundamental shift of organizations’ investments in operational analytics towards a more strategic-oriented form is witnessed by the voluntary migration of basic data systems and related statistical endeavor elsewhere across the firm or by the decision to render it fully automated, supported by 42% of companies declaring to not have yet this Data Scientist role.

Despite the widespread of these expert roles in the corporate realm, and the expectation based on which analytics will progressively become an essential part of the HR function for organizations wishing to stay competitive in the surrounding data-driven world, there is still substantial variability in the degree of maturity of organizations in some parts of the world regarding their use of analytics to nurture predictive insights about what will happen in the future. This confirms their being stuck on intuition and guesswork when making workforce decisions, highlighting an intense gap from their current state to the appropriate use of analytics. A KPMG’s research, in 2016, revealed that, for example, within the UAE boundaries, a very low percentage of respondents (7.5%) have the intention of near-future heavy investments in advanced intelligent data analytics and payroll systems. We will understand further the underlying causes of this slowdown. To conclude, Statistics MRC Global Market Research Reports Company has estimated the monetary value of the global market for people analytics to have touched $439m in 2015 and forecasted it would grow at 16.7% a year over the subsequent seven years, reaching $1.29 billion by 2022.

1.3 HR’s journey towards an evidence-based approach

We have emphasized, until this point, how organizations, in their highly competitive landscape, are in need to make sound, ponderate and judicious decisions when dealing with their workforce, the primary source of value, for ensuring valuable employee recruitment, development and engagement practices. Over the last two decades, however, an important debate has been rising among scholars and HR practitioners regarding firstly the balance of reliance on intuition versus data, and secondly the type of analyses to adopt when it’s time to make such people-related decisions.

1.3.1 From intuition to a data-driven decision-making

Historically, the qualitative nature of the HR department, as opposed to other corporate functions such as finance, supply chain management, risk management and sales and marketing, has caused most HR managers decision-making to be, for quite a long time, predominantly and preferably associated with the use of intuition, anecdotes, gut feelings and instincts. Fortunately, this trend has lately started to change, with organizations, above all the larger ones, realizing the hidden potential of analytics when combined with HR issues. Advocates of this type of algorithm-based decision-making had the idea to radically change the corporate environment “from a culture that largely depends on heuristics in decision-making to a culture that is much more objective and data-driven and embraces the power of data and technology” (McKinsey Global Institute, 2016). Innovative managers that have given credit to this vision are witnessing how Analytics application to the HR field leads to data-driven, quantitative and objective decision-making. Implementing such analytical tools results in a greater overall organizational performance thanks to the increase of efficiency and rationality and a decrease of human errors (Leicht-Deobald, 2019). More precisely, firms are going toward a more evidence-based management that calls for managerial decisions based on hard facts to avoid “dangerous half-truths and total nonsense” (K. King, 2016) that derived from over-reliance on gut feelings, past experiences or generally accepted beliefs. In this way, data insights arising from the meticulous and explicit use of the best available evidence from multiple HR data sources, turn out to be a precious type of information that can consequently inform data-driven workforce managerial practices.

In fact, managers are often limited in their information process and are inevitably threatened by their biases. Two main instances are the well-known confirmation bias and stability bias. Concerning the first mentioned, it's apparent how decision-makers are inclined to give credit to data points that support their original hypothesis in spite of reflectively weighing possible alternatives that are opposed to their working theory. As victims of the stability bias, on the other hand, people tend to prioritize data points that back up to certain agendas and maintain the status quo, ignoring data outcomes that are considered as a source of risks (Falletta, 2014; Rasmussen & Ulrich, 2015). Eventually, along with the proliferation and growth of increasingly rich and diverse data sources, data insights derived by a data-conscious decision-making approach are proved to lead to more accurate, consistent, transparent, and faster key decisions. Algorithms react automatically, providing real-time responses, and are also very accurate, as predictive analyses offer a wider and clearer view of the future; not only that, as algorithms are far more reliable than human intuition, because they don't miss data points, or look differently at the same information based on the contingency, and are even transparent, grating the chance to review the decisions later in time for improvement. Yet, on the debate, also Davenport (2006) made his position clear, explaining that there still needs to be a balance in the degree of use of numbers over instinct in every area of management, since research confirms how the majority of managers are still capable of taking fast, judicious and accurate decisions related to their employees' personality and character under some circumstances. Human judgment is hence believed to be still valuable, since HR analytics provides input for management discussions that can enrich the decision quality, linking data findings to the "story" behind organizational circumstances. In addition, when it comes to people interaction, without being able to consider socio-cultural trends and the psychological status of each worker, statistics can't determine, alone, all sorts of strategic business choices that are impacted by ambiguity and other unpredictable people behavior and actions. This is due to the fact that data cannot speak for themselves, then it's a human job to make sense of analytics results, first contextualizing and interpreting them, and then figuring out the likely consequences of the algorithmic outcomes (Leicht-Deobald, 2019). Besides, it's rare to draw a straight line from data and analyses to action. In the end, not all decisions should be entirely based on analytics when dealing with human capital; in contrast, they should rather equilibrate the use of both wisely, without stubbornly looking for limitless answers and shortcuts theoretically provided by numbers. This combination of critical thinking and best available evidence would eventually favor better conclusions to come out.

1.3.2 Unleashing the power of predictive analytics

Taking into account that we are going to explore much deeper in the next chapter the distinctive features of the three main levels of HR analysis - descriptive, predictive and prescriptive (Fred & Kinange, 2015) -, it's essential to highlight that, for the first time, HR owns the analytical tools it requires to elevate to the next level. and gain long-standing effects for the entirety of the organization. In order to do so, it's crucial to alert HR managers that they urgently need to realize the importance of wide-ranging and advanced predictive analytics techniques, in particular. This is because most HR personnel are currently found to spend considerable effort and time creating mere descriptive reports to deliver organizational insights that deploy only a snapshot picture of what is happening within their firms at that precise time of reporting (M.R. Edwards, K.Edwards, 2019). With this limited mindset, they waste enormous amounts of time in monitoring the differences between these reports, confronting them across various time-span periods, and lastly in updating them on a continuous basis. All of which without getting the little value those descriptive reports eventually provide: they represent tools for later-in-time evaluations of company data but they are not a suitable solution for deeply understanding how to change what is going on inside an organization. What's more, other than being unable to test the level of accuracy and validity of the data analyzed, when running these reports, analysts inevitably miss to fully interrogate the HR data at disposal for grasping possible explanatory reasons and factors that put lights into the deeper questions at the roots of the problem. Then, basically, it should be clear now how HR practitioners have put much more effort into gaining a summary of what has happened, rather than assessing what could happen (Levenson, 2018). Rather, in order to solve deeply-rooted workforce problems, it's necessary to use predictive techniques that possess a high volume of available historical data sets to drive patterns and trends. This way, managers could identify the actions most appropriate for managing their workforce (Wang & Hajli, 2017).

Being more precise, such predictive HR analytics refers to all the sophisticated statistics and quantitative analysis and techniques - modeling, machine learning, data mining and AI - that scientists use to make forecasts about unknown events that are applied to the workforce-related information gathered by each organization. Only this way HR professionals can be in power to make better strategic decisions about the workforce issues and challenges they face day in day out in their work. Namely, grasping the causal factors for why things have happened and manifested in evidence. By analyzing historical and current data to predict future data, predictive analysis is being used by HR practitioners in firms to foresee human behavior and optimize performance. For instance, finding which reasons drive certain behaviors' patterns inside an organization (e.g. what may cause an employee to leave the company, high employee and team productivity's driving factors or great engagement level), managers can test statistical models and would most certainly end up designing higher-quality strategic actions and HR policies. For all these benefits, HR predictive analytics is making the HR function way more evidence-based, since it has the potential to achieve 100% accuracy in decision making for people-related matters. Anyway, predictive analytics is recognized in many other different contexts as a method to discover the drivers of tangible business outcomes, understanding the past and present to predict the future. To bring some examples, financial agencies conduct cost/benefit analysis daily, banks assess consumer and commercial credit risk using facts and data with the application of predictive models, while buying behaviors could be forecasted by using customer demographics metrics in other sectors. However, descriptive analytics is the most well-understood type of analytics and the most commonly used by organizations (Evans, 2016). Speaking of which, a recent 2019 study run by OrgVue, discovered that the large majority of organizations were utilizing analytics office software such as spreadsheets and power points presentations for routine tasks and cost analyses, with only 10% of respondents stating to use HR Predictive Analytics software extensively. By contrast, the trend is changing, as shown by a report by KPMG, in 2019, ensuring that most companies were planning investments in predictive analytics over the next couple of years. That being said, the risk for HR Analytics to become a "management fad" (Rasmussen and Ulrich, 2015) is still present whether, during the next years, this trend wouldn't go as forecasted, claiming Workforce Analytics to fail the passage from descriptive to predictive analytics that would increment HR's value, stopping HR from gaining a seat at the strategy table. Organizations should account for the fact that human capital future assessments are crucial to gain competitive advantage. In the recruitment area, for example, predictive techniques carry out quantitative evaluations about a candidate's talent and automatically double-check if the candidate's profile matches the skill requirements. Leveraging all the personal information given by the Big Data sets of each company, not only predictive analytics would help in acquiring the best talent from the market, but would also develop insightful algorithms that tell executives how to motivate their workforce, how to designate their careers and how to possibly retain them for a long time. All these aids will lead HR Analytics to align with the organizations' strategies. To sum up, it's important to stress how HR should step away from the mere signs of progress in operational reporting and descriptive-analytical deliverables to hit the road of predictive techniques with significant investments and embracing new technologies. Understanding and interpreting large volumes of people-related data, the main benefit of an HR department that relies on predictive analytics is to know in advance the needs of an organization and of its employees. Tangible deliverables benefiting shareholders, customers and employees themselves will follow. Ultimately, that subjectivity in the decision-making process, typical of an untrustworthy non-evidence-based management, would be finally set aside, leaving room for transparency and a more conscious endeavor for redirecting money on more beneficial employee initiatives that impact critical business metrics.

1.4 Barriers to overcome for implementing HRA

We have already mentioned how most organizations' cultures worldwide don't reflect a data-driven business environment. If the HR function wants to become a key strategic player and not a vanishing fad, in the first place it should take steps to develop a strong focus for analytics aligned with the wider business strategy (Rasmussen and Ulrich, 2015). For this to come true, it's required to follow along a path full of threatening obstacles. Throughout this section, the intention is to describe a narrative about these hurdles and provide solutions to overcome them and make real progress into the Workforce Analytics domain.

1.4.1 The lack of analytical skills and confidence

Even if increasingly seen as a “must-have” capability for the HR profession, is to consider that the most visible and diagnosed barrier to effective HR practices lies in the insufficient technical background present in the HR professionals’ careers (Angrave et al., 2016; Marler and Boudreau, 2017). Generally, it’s understandable that HR practitioners may not have the same quantitative rigor compared to other more number-oriented departments. Sure enough, in 2019, according to a report released by OrgVue, 62% of CEOs deem the shortage of availability of key skills and the difficulty to acquire them to be the single biggest threat to their business for initiating an effective HR Analytics project. These HR analytical skills, in the literature, according to Levenson (2011), are divided into those related to statistical techniques (i.e. basic data analysis, intermediate data analysis, basic multivariate models and advance multivariate models) and other analytic competencies (i.e. data preparation, data visualization, research design, survey design and qualitative data collection) (V. Fernandez and E. Gallardo-Gallardo, 2020). Lately, also Machine Learning and AI are growing in importance as value-driver techniques for HR, but also here the barriers are high as is testified by a recent study delivered by KPMG, that found that only 36% of HR functions, in 2019, have started to introduce AI into the HR function. By the way, as HR analytics will be seen as a source of competitive advantage, it will be increasingly fundamental for organizations to have in-house professionals who know how to interpret the workforce data provided by predictive analysis insightfully, in order to make impactful changes affecting business performance. This means HR practitioners should also be able to identify problems, assess potential solutions and meticulously test solutions through a defined research design and methodology.
Additionally, software technology complications only aggravate the situation as HR professionals find themselves reluctant to use HR predictive analytics software extensively because of their little technical knowledge to successfully adapt their HR systems to the standard models arranged inside the vendor’s software (Angrave, 2016). Sometimes, even the leading-edge software firms themselves often can misinterpret the specifics of the HR organizations’ context, worsening the process of software adoption. That is why it’s crucial to choose the right analytics technology provider that acts as a strategic partner, able to suit the present and future requirements that are likely to

change, adapting to the business needs. As Boudreau (2017) suggests in his research, more user-friendly interfaces in the software packages for analytics beginners are needed. All these issues are ascertained by an OrgVue report where it has emerged that, in 2019, 55% of decision-makers consider software technology as a jeopardizing pitfall when it comes to conducting effective Workforce Analytics. However, is to be reminded that People Analytics represents a new and open research area for organizations, so special guide materials are strongly required for building experienced teams and for understanding how to get the most out of these new technologies. In addition, a change in the approach for what concerns the level of familiarity with analytical tools becomes fundamental also because, even if not expert, sometimes the right technology may reduce the complexity at play-acting as a bridge between the absence of skills by one side and the desired outcome. Technology itself is also favoring the integration of functional analytics thanks to the advancement and diffusion of cloud, real-time data, and cross-functional/line of business "enterprise" platforms that permit HR departments to integrate employee data with other business criteria. By contrast, historically, these data platforms were constrained to each functions' boundaries, resulting in many different non-aligned reporting activities. If a workforce team is not familiar with the technologies and instruments at disposal, they may not be conscious of the analytical possibilities they may thrive. For all these reasons, and above all for the business to keep pace with the market, getting the workforce taking part in extensive data science training may turn out to be fruitful, also because getting caught up with technical skills in the world of analytics is not such a deed. It's then recommended that an organization's HR tool kit be renowned to make sure that HR serves as the engine room for exploiting available data. Nonetheless, despite a growing desire to acquire analytical capabilities, the majority of HR professionals are not that attracted to analytics training, mainly because of their employees' little opportunity to work with data as part of their current role. Efforts should be directed to get them to experience firsthand how analytics supports their activities in delivering beneficial business outcomes. Lastly, even if advanced computational tools are being introduced on the market more and more to increase the portability of all companies, the huge costs that these models cause to the company must be considered, both for the high price of the software that support these analyses, both for the costs of maintenance of the tools, as well as for the remuneration to be guaranteed to the new business analyst figures who support the entire process.

1.4.2 Delving into irrelevant workforce questions

What's more, even when HR professionals do own some level of skills, they most certainly might not know how to effectively consult available HR data sets to ask the right questions and gain the valuable insights expected (CIPD, 2013). As applying a classical HR perspective based on which Workforce Analytics takes care of ''doing things right'', HR practitioners may fall into the error of elaborating the right type of questions. They could stubbornly get stuck with questions such as, for instance, "do we use the right compensation system? Is the staffing level adequate? Did the training program improve performance? How efficient is our recruitment process?". Until not long ago, these questions were answered fragmentarily through internal HR systems that often gather individual performance data deduced indirectly from after-the-fact and often subjective appraisal ratings by supervisors (Angrave et al., 2016). By contrast, adopting a different perspective based on the promise that HR 'does the right things'', regarded also as ''outside-in'' thinking by the literature (Rasmussen and Ulrich, 2015), they would be able to grasp the real drivers of the overall business performance from which extract meaningful insights for managing the current and future employees. Such questions could then be modified as follows: "how could we project the compensation system so as to see the workforce's level of productivity maximized? How can we rethink our recruitment process to get the most talented people? How can we increase the likelihood of retaining our employees by offering that specific training program? How does diversity affect team performance? How can we facilitate learning and develop talents in the organization?".

Moreover, such an "outside thinking" of HR enables people's matters to get a closer connection to the key strategic organizational objectives, capturing the strategic linkage between human capital and profitability and tying HR issues to metrics like productivity, revenue, and growth. Therefore, it's recommended that analytics be actionable, meaning that it should resist the temptation to continuously chase many smaller and less value-adding issues that are not core for a business issue. For this to be effective, different sets of more focused, long-term oriented and comprehensive questions could be ‘‘what is our business calling in terms of biggest challenges over the next 3-5 years, and how can HR add value to the business on the same?”. Secondly, a wise action may consist of encouraging HR analytics to release “smart” actionable data to prove the value of HR analytics to shareholders. These are known to appreciate employees able to tell a story about their data when it comes to building a compelling business case. To catch their attention, it’s crucial to blend quantitative results of advanced models with an attractive narrative, so as to be more persuasive. Displaying findings using strictly data language, like only mentioning “p” values, doesn't provide high-level figures solutions to support envisaged business initiatives. The competence of “selling a story” includes, for instance, visualization, communication and captivating storytelling skills that allow analysts to achieve a great deal of quantitative self-efficacy and gain a developed analytical mindset. In line with that, another prerogative for scaling a HRA function to its full potential, corresponds to assembling teams that own a range of varied competencies, as it is unlikely to find all individuals with all the abilities needed (V. Fernandez and E. Gallardo-Gallardo, 2020). Lastly, another frequent pitfall on this matter is represented by the C-level managers’ tendency to refuse data insights that bring into question existing beliefs for which they previously have invested time and efforts or which they consider great ideas or projects of their own. In other words, fresh HR Analytics findings might doubt the value of these initiatives, so senior executives could be emotionally pushed to refrain from their activism in a prompt intervention on policies. This is the power of people analytics: not just validating existing knowledge at hand, but adding value to the decision-making process to make an impact on business success. Furthermore, another obstacle is represented by the lack of support and direction from the same chief human resources officers, which generates considerable practical and bureaucratic barriers to implement effective analyses.

1.4.3 HR data not integrated with other business functions

From the literature, it seems also evident that many organizations may not be capable of gathering the right, necessary, readily available data to transform that information into results (Fitz-enz & Mattox, 2014). In fact, it is not always a matter of the amount of data, but about acceptable quality data for informed decision-making. If the data are not reliable or appropriate, employees will not certainly be able to interpret the results and implications the right way, accordingly. If any meaningful insight is to be gained, improving the quality of data is fundamental and requires a huge effort that starts with the creation of an environment in which people's mindset inextricably recognizes the value of producing data. An understanding of the data at disposal and of the context under which that data has been collected will also assist managers to determine, by one side, the resources that are needed and, by another side, the form of the subsequent analysis that will be run (Fitz-enz & Mattox, 2014). Then, data has to be treated and regarded as an asset to success in working with it, indeed, and some standards need to be imposed. Beyond that, generally, the data at hand cannot be defined as integrated data, meaning that not all companies own data systems that enable an enterprise-level strategic debate. This kind of debate should emerge thanks to the evident connections between diversified insights from the different HR areas and the different corporate functions, like the financial, engineering and IT departments. In fact, merging these data is helpful to consider actions affecting business core operations. As lack of integration in HR corresponds to an obstacle toward more efficient and effective analytics, HR employees should work with the high-ranked data officer to discuss whether to rethink the relationship between HR and business data in the context of strategic organizational objectives. In sum, it's not enough to execute the HR processes, rather, all the parts of HR should come together to take a more holistic approach with the intent of solving business problems without a preconceived approach. Also, under this enlarged perspective, HR Analytics must transcend HR functional boundaries to aim to be part of the rest of existing cross-functional business analytics, since in most companies data is not integrated across functions, geographies or divisions (Douthitt and Mondore, 2014). Silos mentalities are known to hinder people-related data link with other determinants of productivity and performance associated by other functions (V. Fernandez and E. Gallardo-Gallardo, 2020); within this setting, it turns out to be difficult to build effective analytical models that study the impact of HR-related factors while taking into account other relevant factors at play (Angrave, 2016). Eventually, as already stressed in the first place, in order to undertake advanced HR Analytics & Big Data solutions in a pervasive and effective way, it is then necessary to undertake a growth path within the HR Department not only from a technological point of view, but also from an organizational and strategic point of view. This would lead to building a People

Analytics culture in which business strategy is aligned with HR strategy and really incorporates analytics, avoiding insisting obstinately to envision an analytics strategy that is apart from the business core and that takes care only of specific business areas. Furthermore, in order to start this more strategy-oriented path, HR practitioners should be empowered to generate deeper insights into a few very important critical areas of improvement, considered as priorities, where analytics may have a great impact. For instance, aiming at optimizing the recruitment process or reducing the turnover rate of younger employees. Otherwise, without a clear and narrow focus but multiple requirements solicited at the same time, the risk is to not favor the success of the overall organization strategy.

As a conclusion, barriers to delivering analytics are inevitable and those that aren't aware of them aren't just implementing analytics in a meaningful way or their path is still in an early mature stage.

1.5 The ethical debate around HR analytics

According to an HR executives interview by the HR Innovation Practice Observatory, among the reasons that compromise the spread and realization of HR analytics projects, there are also fears related to the actual use of data and a lack of an extensive legal framework to consult when data ethics concern emerges. In this section these issues will be analyzed, presenting also some dysfunctional limits of automatic algorithms influencing manager decisions.

1.5.1 Employee private data concern

In the context of HR, much of the employees' data is theoretically owned by organizations, as oftentimes employees have voluntarily agreed through contracts that the company can access and utilize this data in the course of the business. However, the personal workforce data storage is still highly sensitive and quite personal (eg. age, gender, address, religion, bank account records, salary level, marital status, etc) and shall be protected. Therefore, companies need to exercise great care in deciding what data to collect and what to do with it (Cappelli 2017). In fact, considering the large volume of data that can be collected on people and that can be matched and linked to identifying undiscovered patterns, people may worry that their data can be used to control, arising grey areas on the use of Big Data in the context of the HR function. On this issue, it is important for employers to clarify for what purposes such workforce data is collected, that is to include it in HR Analytics activities and projects, with the aim of obtaining information that leads the company to a better management and development of its people. For example, monitoring social interactions between colleagues can allow an executive to identify the focal points of competence within the organization's workforce, regardless of its hierarchical roles; the monitoring of the level of attention of the employees during the different online training methods, combined with the evaluations of the same, allows to understand which are the most effective formats; finally, correlating performance data with certain characteristics of the person, may provide useful insights to the personnel selection process in order to include people more suitable to the future high potential corporate culture.

Apart from this, it should be apparent that the only way to better manage the "psychological" aspect linked to the use of data within the company is transparency in communicating how such data will be used to avoid employees' discomfort. Companies need to question also where the surveillance boundaries should lie, in order to distinguish what information of personal nature should be taken into account for analytical purposes and which not. Many questions arise, for example, with regard to companies accessing employees' food habits at the staffing canteens: many organizations were found to link these data with employee performance without having their employees even aware. Another issue has emerged with companies checking workers' social media activities, to see if they were making disparaging remarks about the organization on Facebook, Twitter and other sites. Importantly, it's reasonable to argue that companies cannot scrutinize employee lifestyle choices as they wish only to come up with algorithms that tell them how they can be most productive: some trade-offs between efficiency and ethics are to be done to avoid invading employee personal space. For this reason, issues such as privacy, acceptability and data security are of paramount importance for companies, that must both collect and use data relating to personnel following the regulations in force, and ensure the privacy of their employees for the processing of the same data. It's the HR analyst's role and responsibility to abide by and to treat data according to the specific geographical legislative guidelines. They should guarantee that some data don't fall outside what has been formally settled. Nowadays, greater and greater amounts of data are available from sensitive sources such as employees' wearable technologies and mobile phones records (e.g. people's emails, social media channels, websites), putting a massive responsibility on employers to separate discrete actions for monitoring from intrusive practices (CIPD, 2013). In addition, since HR data is pretty different from data used by other analytics teams, most of the personal HR data is even more sensitive and private. Also, in the context of a workplace where data is gathered through a contractual relationship, employees may not always choose which information to share and not share or may not even be consciously aware the precise data are being collected. Furthermore, at some point employees could feel as if they are reduced to just numbers other than individuals whose company cares about.

1.5.2 Developing a governance framework for an ethical usage of data

Surrounding all these factors that seem to cross the boundary of acceptable ethical practices using employee data, the real concern emerges when HR functions still miss an appropriate privacy-preserving strategy to balance employee private data collection with the business value gained through the insights derived from the analyses on these same data. In order to find this compromise, firstly, companies have to consider that legal standards differ across countries, when determining data policies. More than that, consulting with employee representatives for shedding light on the purpose of gathering data and the potential benefits for employees to do so, is becoming strictly necessary in some countries. Nowadays it is then required the creation of a robust governance framework for the ethical use of personal data on employees (CRF Research, 2017). Privacy protection only concentrates on securing personally-identifying information, not protecting workers from the abuse of data within the work context. While, so far, governance has focused primarily on data issues, such as creating data standards and policies for dealing with sensitive personal data, the final outcome of this renewed organizational framework should have key business stakeholders committed to defining the purpose and the boundaries of the workforce analytics, setting some small decision-making bodies. Alongside these information governance committees, also new organizational roles such as Chief Data Officer, Chief Information Governance Officer or Chief Privacy Officer, are considered as ways of protecting employee privacy while remaining in line with corporate objectives. To enforce employees' trust in executives, for instance, the renowned bank JP Morgan has established a People Analytics governance committee consisting of group-level HR operating committee members. What's more, the link between Big Data and HR cannot be managed and sustained in any single place; to create a greater sense of responsibility, other parties, as well, should be involved in the discussion, like user communities and data suppliers, who may help to reach more collective ethics, bringing points of view of participants who are not the companies and "usual" stakeholders. Reflecting on interactions and negotiations, organizations might reach awareness about the incremental need for a more inclusive and shared space of action in HR Analytics.

As long as legal compliance, a building block of all HR data policies is concerned, there is still a lack of robust legal protections in diverse parts of the world and uncertainty over their regulatory framework. This could make organizations put their workforce analytics projects on hold. At least, the GDPR (General Data Protection Regulation), entered into force on 28th May 2018, is seen as a significant advancement of employee rights in the digital era. It only applies to European Union citizens, even though all organizations processing personal data overseas of individuals located in EU member nations must adhere; GDPR is a strict set of new rules which entails intimidating fines for non-compliance. It moderates the capability of any employer, not affecting only big tec giants, to utilize personal data about their employees for means not previously settled at the moment of data collection. Ultimately, when personal data is needed no more, it should be deleted from the server, meaning that HR databases should be checked periodically. Besides compliance, for HR analysts, the use of aggregated, non-identifying data is suggested where possible, to show employees that the analytical projects' only goal is to capture wider organizational trends, attributing non-single responses to a precise employee (Kumar, 2018).

Overall, then, as HR Analytics is an emerging but growing field, it needs to draw a careful line between behaving as an employee representative and driving better business performance, though it is acknowledged that it is not always an easy thing to distinguish what is personal and what is job-related, especially where data are collected from cell phones or notebooks of employees (Bersin, 2019). It's then recommended that companies prepare and publish clear guidance in the form of ethical charters, potentially in collaboration with other organizations. This way, analysts could hold on to some ethical principles that potentially guide HR analytics activities.

1.5.3 Algorithms replicating human biases

Thanks to their automated nature, artificial intelligence algorithms, typically used within HR Analytics frameworks, lower the human energy needed to perform specific repetitive tasks and their associated costs, standardizing underlying business processes and freeing up time for higher-value activities (M. Heric, 2018). However, recent literature has contested the promises of their total objectivity (Bilić, P., 2016; Thelwall, M., 2018). Despite the widespread belief that human judgment in the HR context is strongly shaped by the presence of some biases, mostly related to demographics, in some cases mathematical algorithms may unexpectedly raise problems of discriminatory practices.

In fact, these algorithms could be still impaired by racial and gender biases and turn out to be backward-looking (Leicht-Deobald, 2019). This happens because these artificial intelligence systems could be still guided by people subjected to unconscious prejudice: in this case the software risks replicating the inherent human biases, thus ending up being ineffective. For this reason, HR critics stress analytics teams to assure that their activities are not propagating biased decision making, fostering the need for the model to be appropriately built (R. Hamilton, W. Sodeman, 2020). Predictive algorithms risk outlining a stereotype picture of the employees based on variable categories identified in the analyses, that may support a simplified view of the world based on prejudicial perceptions. Taking as a reference a hiring algorithm created for improving the selection process, for instance, a model might result in faulty discriminatory outcomes, disproportionately selecting solely white males because in the past white males accounted for most of those rated as high performers, and thus reifying and propagating the original gender and/or diversity bias (P. Tambe, P. Cappelli, and V. Yakubovich, 2019). In this scenario, the result of the application of such an algorithm would pinpoint toward an advised decision that a male would be a better hire than a female one, merely because this trend was favored in the past for that organization (M.R. Edwards, K.Edwards, 2019). For all these counter effects of predictive algorithms, senior HR managers should offer guidance to their data scientists to be more careful in their analyses, as it’s a high risk of keeping exacerbating possible discriminatory patterns or confirming implicit bias (R.H. Hamilton, William A. Sodeman, 2020). When attempting to predict future outcomes through algorithms, HR practitioners and expert HR analysts should standardize the application of criteria by which they run models and remove information that is irrelevant for the specific manager decisions, like the race and sex of candidates for hiring decisions. In addition, it's recommended to somehow demonstrate that no other process for making decisions would produce a prediction at least as accurate or with less adverse impact, and that the final outcome is considered good as desired (P. Tambe, P. Cappelli, and V. Yakubovich, 2019). However, employers are still not completely held accountable for such inconvenient cases: rarely they are forced by the legal framework to provide a sort of "explainability" for having made those decisions in a fair manner that was not directed to provide such discriminative outcomes. Importantly, they should necessarily get aware of which key attributes actually drove the decisions to not make that happen again.

CHAPTER TWO
READING HR DATA IN THE MODERN ANALYTICAL SCENARIO

2.1 The technological revolution behind the new HR data ecosystem

Intending to realize today's technological potential, which, as has been demonstrated so far, has a significant impact on the effective implementation of HR practices, it is necessary to retrace the main historical phases and the most influencing factors which contributed to progressively change the way of using and processing company staff data.

2.1.1 Radical changes from Taylorism to 2000s

In the first place, it can be argued that the intellectual foundations of modern data HR vision can be traced back to the concept of "Scientific Management", developed by the management theorists Frederick Taylor during the years between 1910 and the end of the Second World War. His main premise was that non-observed workers are inefficient and non fully productive, then, organizations, to enhance their performance, should constantly measure and monitor any employees' move (U. Leicht-Deobald et All, 2019). This new approach to management, along with the impact of the Technological Revolution's innovations on the industrial processes, brought about unprecedented innovations in human engineering, with the application of the assembly line by Henry Ford, and novel patterns of interactions among different disciplines and society, paving the way for a first evidence-based HR on applied psychology.

Subsequently, after the end of the war and throughout the mid-1960s to early-1970s, another revolutionary period characterized by a mindset evolution in history was represented by the development, in businesses, of Operation Research and Management Science and by the advent and growth of management information systems (MIS) (M. Mortenson, N. Doherty, S. Robinson, 2014). In particular, thanks to the development of standalone software packages and the commercial applications of computers, at the beginning of the 1980s, were disclosed the first real form of managerial strategic support with the early MISs, which move beyond the mere procedures of routinary tasks regarding employee records, such as payroll processing and simple transaction processing systems (TPS). These systems gave new functionalities concerning automating distinct HR functions, and also turned raw transactional data into meaningful information for managers for the first time (J. H. Dulebohn & R. D. Johnson, 2013). With these new tools, the idea of a decision-making HR culture, which should focus on measuring data to gain inspirational insights benefiting the overall business, grew further and the first metrics on measuring cost, quantity, and time of workforce appeared at the end of the 1970s, including also benchmarking during the 1980s. Some years later also balanced scorecards and dashboards technologies were elaborated as new intuitive instruments for the organizations to better address operational targets and strategic goals. HR, which at that time looked like a back-office function with almost no analytics involved and lots of paperwork, was changing its role toward a sort of a service provider to the business. Yet, the computers' dispersion and the MISs were still limited in that decade, making the process of change toward a more analytical HR function still slow. Rather, the 1990s marked another step ahead as they were characterized by stronger efforts to integrate developments in technology and quantitative methods into business processes and decision making (M. Mortenson, N. Doherty, S. Robinson, 2014). In this very decade, we assisted, above all the other innovations, to the proliferation of enterprise resource planning (ERP) software, decision support systems, and many other web-based architecture systems, that were finally capable of managing diverse HR functions and could equip managers with more sophisticated reporting and analytical features. Later on, from the mid-2000s on, as was raising a solid prominence of analytics concern, companies started realizing the connected relevance of the data storage processes to preserve the reliability of the outcomes from data analyses. Therefore, data storage became far more efficient and consistent, and, helped by the growth of cloud-based computing platforms, led to the spread of Business Intelligence (BI) architectures ad new forms of relational databases that companies use nowadays (J. H. Dulebohn & R. D. Johnson, 2013).

2.1.2 The impact of "Big Data"

Nowadays, in this ever-increasing digitalized world, in response to the vast increase of data available from close-to-infinite sources, the wide field of Big Data and advanced analytics, including Machine Learning - which I'll be discussing over the next sections - is making rapid advances, as more data require greater computational power and infrastructure to analyze and turn this flood of raw data into rigorous and relevant insights (McKinsey Global Institute, 2016). More precisely, the so-called "Big Data", widely considered the most significant technological disruption in business, refers to "large volumes of data generated and made available online in digital media ecosystems" (Pappas et All., 2018) and are characterized by five key features: Volume, Variety, Veracity, Variability, and Velocity. These sets of data are so large that traditional ways of storing, analyzing, and sharing data can be inadequate. In fact, during the last two decades, the development of the Internet has exponentially scaled up the continuous generation of massive amounts of data from a growing number of sources (referring to handheld devices, laptops, machines, and so on). For this reason, databases of different types and of growing size were developed to favor the constant gathering of any possibly helpful information within companies, that was progressively made accessible even to small firms having their cloud-based data warehouses and analytics services outsourced, furthermore greatly simplifying their data architecture and IT requirements. HR information systems developers, such as Oracle, IBM, and SAP released their software packages to store data in one place, which made the collection of information from a range of existing databases easier and faster (Fitz-enz & Mattox, 2014). Along with this trend, as was anticipated before, new effective tools have been offered by combining statistics, SQL, and data mining (M. Mortenson, N. Doherty, S. Robinson, 2014). In addition, with the further creation of non-relational databases architectures, like NoSQL databases, in order to store the increasing types of unstructured data (such as images, text, audio, and other media records), collateral developments in text mining, network analysis, and natural language processing followed accordingly (McKinsey Global Institute, 2016).

However, the most-influencing factor inside the HR ecosystem has been the integration of multiple sources, by which organizations today can not only leverage data more readable, but also link disparate data sources together from a variety of data streams, either internal and external to the company, to produce more fine-grained measures to evaluate employees' recruitment, retention, performance, namely HR Analytics' fundamentals (U. Leicht-Deobald et All, 2019). For example, datasets from other organizational resource planning software units, that were traditionally kept separate, like customer relationship management, supply chains, accounting&finance are now associated with HR data (Angrave et al. 2016).

In conclusion, it's essential to stress how the HR departments, nowadays, in order to drive higher productivity and enhance overall employee management, should leverage cloud solutions and cutting-edge technology that allow expanding information and link together different types of data. Along with the new combined set of data, and some professional competencies in using rigorous statistical techniques, HR managers are enabled, in all HR operations, from recruitment to performance appraisal and development, to work with the best type of data from which they can base objective decisions, predict workforce trends, and investigate on areas of concern (J. Baier et All, Report BCG, 2021). That's the reason why, over the long term, increasing the analytical complexities, the data infrastructure will pose a challenge to scalability and migration of data to cloud services (E. Ledet et All, 2020); this way, the value will likely accrue to providers of analytics and data platform owners who will be able to tap the potential of disruptive technologies, leverage advanced data analytics and make important correlations between people data and business data.

2.2 HR data treatment and specifics

In this section, dedicated to the data entering the HR departments, I first explain some of their specific characteristics, providing a couple of fundamental classifications to divide them, and in a second place I try to explain in what HR managers are interested in nowadays when they analyze their workforce data. Ultimately, I will give a panoramic of the main stages that HR data have to pass through before being actually utilized in the analyses.

2.2.1 Classifications of HR data

Regarding HR data in its specifics, it's first of all important to underline how the data sources that are used to populate HR Analytics & Big Data tools can be classified according to two main information: the origin of the data and the type of data itself (Osservatorio HRIP, 2016). With respect to their origin, data are divided into internal company data or external data; internal data come, almost exclusively, from the internal databases of the HR function itself, from other business processes, especially from the company management systems such as ERP, and are extrapolated, in the most advanced cases, in the form of unstructured data, from contributions such as intranets, internal communities, social collaboration tools.

On the contrary, external datasets, characterized by less diffusion, represent data obtained from sources outside the company, mainly such as salary benchmark data or external studies, inter-company datasets of suppliers based on cloud platforms, and lastly data from the world of social media, the latter almost completely confined to the search and selection of personnel, such as for the Linkedin case (Osservatorio HRIP, 2016). Once again, the cloud-based platforms of some providers of human resource management solutions represent an important opportunity for accessing external data, that in this way are analyzed by computational algorithms that work on the entire customer database, leading to more reliable simulations and forecasts of workforce trends.

On the other hand, as for their typology, they are distinguished in structured and unstructured data. The former, usually organized within databases, are easily manipulated and interpreted. The latter is characterized, on the contrary, by containing schemaless information, not adaptable to a relational database or for which there is an irregular or partial structure not sufficient to allow its storage and management through the traditional databases. This refers to images, video content, text files (such as paragraphs of natural language text about the performance appraisals of employees), and so on (CIPD, 2020).

Going deeper, data held in HR Information Systems, which is typically composed of information on the workers who are hired (and oftentimes on candidates that were not hired) may be quantitative or qualitative. Quantitative data can be measured and visualized through numbers, so they are defined as objective, unbiased and measurable (examples are the number of women, an employee's age, or the remuneration levels); qualitative data are those that can't be measured and are often result of subjective assessments, whose analysis was based on words or ideas, representing an individual's view of something. Therefore, they are subjective to interpretation and far less reliable than numbers, and are so defined as value-laden and biased, like could be an employee's opinion on an engagement survey or a performance appraisals review (CIPD, 2020). To name some, the main examples of employee's data in the HR realm are the following: attendance, assessments, performance and measures of individual output, skills and competencies, engagement, demographic information, job status, job type, training, team, diversity, job location, payroll level, employment history, compensation/benefits, educational history and qualifications.

2.2.2 A new approach to treat HR data

Within the analytical framework of the HR department, an important issue regarding the HR data focuses on how these data are treated by professionals to reach out for some relevant insights on the entire company population. Precisely, the overall scenario with respect to the techniques adopted to work and read the HR data has changed in recent years, allowing analyzes that, compared to the past, no longer use free-decision sorting techniques. In this sense, with the advent of information technology (IT) and computational tools, and with the cost of storing data falling drastically with the technology for the production of data getting cheaper and cheaper, companies are realizing how "leveraging Big Data in HR has become a source of competitive advantage that radically transform their decision-making process" (V. Fernandez & E. Gallardo-Gallardo, 2020). In fact, thanks to this phenomenon as well as to the progressive digitization of personnel management processes, data and information available to the HR Department is developing fastly. Additionally, the exponential amount of data certainly creates opportunities to better support the decision-making processes of the business, but it also requires an effort by the HR Department in expanding its skills, both in terms of competencies of the technologies available and of the methods and cases of use of these tools to support the business (Osservatorio HRIP, 2016). State-of-the-art tech companies are firstly approaching either relational and non-relational database software to store and organize data from different sources, creating consolidated profiles of employee data, and secondly are exploring new techniques for representing, understanding, and analyzing data (D. Angrave & All, 2016). In particular, the concept of data sharing and of platforms' standardization seems to be a priority for companies' executive managers both in the long and short-run: if the components of precise data needed for a workforce analysis belong to different parts of the organization, the HR function could not possibly make objective decisions regarding employees (McKinsey Global Institute, 2016).

More than that, HR analytics teams now investigate beyond basic employee data to harness nuanced insights on individual attitudes within the company's walls (J. Baier et All, Report BCG, 2021). This issue of data selection is not a simple procedure as HR data could correspond to information about any aspect of employees and still not all data types are accessible by companies that are left behind technologically (CIPD, 2020). In fact, as was mentioned before, data could be combined with 'bigger' data on how a worker interacts and communicates inside the workplace context with his colleagues. These are regarded as relational data and are captured by organizations to gain a better understanding of their employees' social networks and measure their job contextual performance.

In this sense, other than controlling job-related behaviors, task performance, and compliance with organization rules, HR novel computational algorithms that analyze people data are more strategic, according to a vision for an HR more evidence-based: they can reveal to HR professionals patterns on contextual performance, such as employee engagement levels, attitudes in project groups and employees' moods at the workplace (affecting the turnover rate), that have an impact on the overall organizational performance and could be adjusted designing the right HR practices (U. Leicht-Deobald et All, 2019). All this is due to companies' ability to exploit new completely different unconventional data inputs, which describe what an employee does at work, in order to identify insightful patterns in areas as turnover, retention, and performance, and not just focusing on employees' workloads like happened in the past. With these new data, we are referring to, for example, internet browser histories, email contents, electronic calendars, internet-connected operating software transactions and other information from wearable devices, like tracking the geolocation and the time away from work, in conformity to the privacy boundaries (Angrave et al. 2016). Also, companies are interested in evaluating their employee's communication dynamics and job engagement trends by perceiving their mood, sentiments, and morale through internal community platforms, phone records, online collaborative tools, and social media interactions (tweets, posts, likes, comments, etc.) (D. Angrave et All, 2016). All these efforts because, along with workers' attributes, the relationships between employees has proven to be strongly interrelated with their workplace performance, and HR professionals can analytically predict some key performance indicators, such as the level of team collaboration and a worker's openness to diversity, by looking at their workforce structural signatures (P. Leonardi & N. Contractor, 2018).

2.2.3 HR Data processing stages

Another fundamental issue to discuss is the complicated and multiple-step path that today's raw data gets through before being actually used and analyzed (McKinsey Global Institute, 2016). Within the data ecosystems, firstly, data need to pass over the first step, called "data collection and generation", where data is simply stored and captured from the different sources at disposal of the organization and made available to the various HR applications. At this level, data are fragmented, in the sense that process data are not integrated and are still in raw format, are predominantly of poor quality, and provide information that is often contradictory and inconsistent. Moreover, along the way, data might be heterogeneous, incomplete, might be made of inconsistent data combinations or might even contain out-of-range or missing values. For this reason, data then need to be turned in an understandable and workable format, being manipulated through advanced techniques and time-consuming processes of data transformation (S. Patel, 2017). In particular, irregular or inconsistent data would most likely lead to inaccurate and worthless insights, yielding ineffective or incorrect decisions accordingly. Then, the most delicate step is the one of "Data Cleaning", that involves data scientists to review all the data collected to fix incorrect data, fill empty fields, erase duplicate records to end up with uniform and standardized data sets that favor analytical tools to uncover reliable answers when applied to the data (Indiamagag). Only after, comes the "Data aggregation" phase, when data are blended from multiple sources and platforms and are then extended. This step is performed by applying guidelines aimed at managing the quality of information and integrating it with the internal data of HR systems, although the level of reliability is heterogeneous and not standardized (McKinsey Global Institute, 2016). During this process, sometimes, workforce analytics teams feel constrained by a dearth of data availability, so they get creative through the process, acquiring new sources or combining existing data sources in new ways to attack the problem at hand in a different way (E. Ledet et All, 2020). At this point, process data are integrated and collected centrally and are available to the different types of applications that require their use from time to time. Besides, other than from differently structured sources that support the processes, unstructured data (e.g. text, audio, video, social media, ...), files are also collected and analyzed, and made available in formats consistent and integrated with other HR systems. All these processes are aimed to add value to existing data: the result is that at this level the quality and reliability of the data is generally high (Osservatorio HRIP, 2016). At this ultimate point, this enhanced data is ultimately converted into practical results or insights by applying analytical techniques in the "Data Analytics" stage, in which data can finally be acted upon. Generally, the more data is refined, the more it becomes applicable to specific uses, raising its value and becoming monetizable. However, to ensure a solid data infrastructure, is important that a diverse landscape of infrastructure providers actually offers the hardware and software, associated with company data management, necessary to execute these stages (McKinsey Global Institute, 2016). Therefore, in the first place, organizations that plan to use advanced HR Analytics & Big Data, should identify and collect the type of employees' data, from various dimensions of human resources, that perfectly fits their workforce analytics projects, and, only on a second stage, develop a strategy for acquiring and extracting these data from the systems at their disposal (Cappelli, 2017).

2.3 From Traditional HR techniques to HRA

We have been stressing so far how nowadays the traditional protocols to analyze and read the HR data have radically improved thanks to new computational techniques, elevating the HR function to a quite new fact-driven level. In explaining how to possible scale to this ultimate level of HR analysis, according to common categorizations in management (Davenport, 2010), within the organization's journey for building up an effective HR Analytics function, analytical tools, visual technologies, and mathematical algorithms used during the process of data treatment can be broadly distinguished into three main classifications depending on the maturity level of analytics: descriptive, predictive, and prescriptive analyses.

2.3.1 The 1o level: descriptive analysis

Descriptive analysis is the first level of analysis and implies an understanding of past events, trends, and results (Fitz-enz & Mattox, 2014). It seeks and describes the relationships between data through an exploratory approach, but without attributing meaning to the models studied. With these types of analyses, organizations are able to analyze and understand historical patterns, summarily grasping how what happened in the past influences the present, but cannot end up with a valid theory for the future. For this reason, in the context of algorithm-based HR, although HR analysis tools can facilitate HR decision-making, descriptive analyses, being useful only to reveal current data patterns, still represent very powerful methods as data are getting more and more granular, as well as their integration from different sources is getting stronger. Basically, leveraging this information about the past and current trends, companies enable costs reduction and improvement of organizational processes. Whatsmore, descriptive algorithms, even though are built on relatively simple statistical terms, such as means, frequency counts, standard deviations, correlations, or percentages successfully explain, for instance, the distribution of variables or the association between variables, helping analysts to find the most-visible business insights behind data (U. Leicht-Deobald et All, 2019). Generally, HR managers use tools for passive analysis that offers a view of the existing situation and mainly envisioned reports are produced to reveal the underlying cause of the event in business: dashboards and scorecards, that include traditional HR Metrics (such as the cost of hire or turnover rate) and KPIs, measuring the efficiency and effectiveness of the HR outcomes and processes, are widely utilized by managers and HR representatives, with the possibility of directly processing data in real-time (Osservatorio HRIP, 2016). These data visualization tools contribute to producing an organization's human capital reporting, built along with predetermined objectives in mind to identify the right HR metrics to use for the task at hand, that is nothing more than the constant production of analytics, data, and narrative information which describes the quality and quantity (from workforce demographics to culture and engagement) of the human capital population inside a company. That's why descriptive techniques are still the most widespread type of analysis performed today (CIPD, 2017). To be more precise, dashboards and balanced scorecards, whose infrastructure is developed through Excel and online, or in-house, dashboard platforms, constitute an interactive system that visually highlights the current snapshot of key HR metrics that are considered important drivers of the organizational performance. Dashboards are made of graphical presentations, charts (such as bar charts, scatter plots, histograms, pie charts, line graphs), and tables, that represent visually and intuitively a large scale of data, to keep track of strategically relevant indices to create standard reports or respond to ad hoc requests (H. Opatha, 2020). This method enables managers for in-depth exploratory analysis every time emerge undesirable scores and trends that push a need for further investigation (P. Leonardi & N. Contractor, 2018). In addition, visualization skills, differing at varying degrees of sophistication, turn out to be really relevant to reach communication effectiveness when explaining the outcomes of an HR analysis to an audience of decision-makers to whom HR analysts depict the situation. On the same descriptive reporting level, moreover, another practice, emerged with the increasing gathering of comparable data across organizations during the last three decades, concerns "benchmarking", that play a key role as companies can evaluate the status of their own HR practices and results in comparison to other organizations in the industry (Fitz-enz, 2010). However, external benchmarks data don't provide competitive advantages, as, not these data being analyzed, they don't create actionable business value. Overall, through all these techniques mentioned so far throughout this section, descriptive algorithms are greatly useful to HR managers for keeping employees motivation, engagement, and performance under control, but cannot foresee these indices in the future (U. Leicht-Deobald et All, 2019). Among the statistical techniques that are specifically utilized by the HR department to process data for descriptive tasks, Cluster analysis, also called group analysis, is by far the most famous and frequent (X. Guo, 2020). Like discriminant analysis and factor analysis, clustering techniques are multivariate analysis methods that sort a plurality of observations into multiple relatively homogeneous aggregates, or centers. For this reason, cluster analysis is known to be a powerful technique that leverages the information contained in a multidimensional observation. This could turn out to be a helpful data-mining system for any organization that wants to discriminate their customer base but, more specifically, translating this into the HR operational framework, with this method employees could be successfully grouped together based on similar characteristics that they share across several variables, or better, computationally speaking, based on how closely associated they are and how distant they are. In fact, the degree of similarity between two employees is given by a determinate index or measure computed over the whole set of dimensions, that generally correspond to the variance, by which, at each stage of the classification process, there must always be, for each employee sample, little variance between members of the same class and large difference among workers belonging to different classes, otherwise, after a certain distance, a new aggregation class is formed (X. Guo, 2020). By the way, differently than for other statistical techniques, being cluster analysis an unsupervised learning method, which I'll deeper discuss about over the machine-learning final part of this chapter, it provides information about where associations and patterns in data exist, but don't suggest favorable information on what those might be, what they mean or by what the latter ones are caused, and for this reason, it remains anchored to the descriptive sphere of analysis. Lastly, is important also to mention the two common methods used in clustering analysis. The first is K-means clustering, also called fast clustering, and the other is the Hierarchical Clustering, whose process of clustering is different from the previous one because, in this case, the number of clusters is not pre-determined in advance (G. James et All, 2017).

2.3.2 The 2o level: predictive analysis

Having provided a wide panoramic of the instruments and main techniques of the traditional way HR department has been treated its data on employees, it's essential now to explore, first in general terms, what predictive analytics, as part of the second level of HR analysis, is about. This type of analysis, rather tha focalizing on past employees record, deals with data-derived insights that give meaning to historical patterns and real-time observations, to predict, to some extent, the probability of future outcomes and their key related factors (Fitz-enz, 2010). Basically, predictive algorithms determine the likelihood of an event to occur (U. Leicht-Deobald et All, 2019); HR managers are then enabled to plan for "what-if" scenarios, making sure they can deliver to the business and understand the impact of interventions (CIPD, 2017). Therefore, analysts need to expand their computational capabilities and mathematical complexity, using more sophisticated statistical techniques, such as regression techniques, machine learning methods, and data mining models. As an instance, computing some more advanced inferential statistics techniques, like using an analysis of variance (ANOVA) than the simplest mean, median and standard deviation, it's possible to uncover meaningful differences between groups, while classification and regression analyses can uncover relationships among variables, so as for HR analytics to finally understand the hidden factors that drive the metrics displayed on the periodic dashboards. With predictive analyses, HR managers can finally set aside, for example, their involuntary biases in recruiting new employees and leave room for the objective nature of the algorithms. Also, companies, at this stage, need to scale up their data-science competencies and operations, as programming languages like R and Python are used to join disparate sources of data, develop models to interpret complex phenomena and provide actionable and strategic recommendations (E. Ledet et All, 2020). In fact, these systems extract possible future evolutionary scenarios by analyzing, classifying, and simulating large amounts of historical data of different characteristics, that must be of high quality and robust nature. Generally, evaluations of simple types of machine Learning applications or structural equation modeling (SEM) are the basics required, as well as is strongly suggested a meaningful and consistent organization of massive amounts of highly accurate data (Fitz-enz & Mattox, 2014). Ultimately, moving beyond this HR reporting limit, anchored on summaries of the current state, the HR Analytics promise, based on Big Data and predictive modeling, demonstrates how various elements of the employee can impact the business and contribute to making HR more strategic, like other business functions, delivering future people outcomes. However, no organization is identical in terms of workforce, talent, environment, strategies, and type of market. Hence, there is no single fixed and successful predictive models that should be applied to all HR functions across different enterprises. Therefore, predictive analytics will become essential for industries that wish to introduce unique decision-making policies and who want to optimize the performance of the organization and improve HR practices.

2.3.3 The 3o level: prescriptive analysis

Scaling up toward the third, most complex level of analysis, once a prediction model is built, there comes the optimization phase, where HR managers can implement programs that improve all HR practices, from enhancing the hiring, training, and promotion of talents to reducing the employee turnover levels, so as to develop the best framework for the highest outcome to happen. This third stage of analysis, known as prescriptive analysis, in this framework, adds to the predictive component the ability to delineate and compare different possible scenarios, and suggested courses of action, by both identifying and demonstrating interrelated cause-effect links of each possible decision (Osservatorio HRIP, 2016). Prescriptive algorithms then reach a more advanced level in analysis. Consequently, keeping into consideration the HR Analytics maturity path, we assume that, as the value of the analysis increases, the computational sophistication raises as well, going far beyond statistical techniques and simple machine-learning algorithms. Therefore, prescriptive techniques, being able to analyze heterogeneous data sources of structured and unstructured type, function not only as decision support but also as decision automation techniques, embracing the proper Artificial Intelligence complexity whose structures are almost impossible for a person to handle. For this reason, they are currently inaccessible by even the most analytically savvy organizations due to the technological and computational competency barriers. In fact, with respect to the predictive algorithms, prescriptive ones are built combining a far greater number of variables and including more sophisticated types of analyses such as simulations, optimization algorithms, scenario-based techniques, and deeper machine learning techniques to foresee outcomes and provide decision options to drive actions (CIPD, 2017).

Eventually, having explained in detail the three-level of analytics present inside the HR Analytics field, it's important to mark that HR analytics is like a "continuum", meaning that a company should posit itself along the HR analytics maturity path depending on the sophistication of its HR processes, data quality, and computational capabilities available, but always looking at the step further and experimenting with new technologies to analyze and disseminate the data.

2.4 Human Capital Metrics for analytics

In this section I will explore the realm of HR Metrics, which represents the widest used instrument for analysis at every level of analysis inside the HR function and, thus, it deserves a deeper examination.

2.4.1 Scope of Metrics in HRA

Based on this basin of HR data inside a firm, metrics represent a set of accountability instruments used for evaluating a function's results, that includes a set of “key performance indicators” (KPI) which can be adapted to each different area of personnel functions.

Mastering some HR metrics suitable for analytics that have been around over the past three decades, by collecting and archiving data on the company population, transforming them into metrics to be analyzed and measuring their value, symbolizes a key factor for the HR department and an initial step on the HR Analytics maturity path: predictive models are in fact built around these metrics to produce actionable insights from historical data. In other words, concrete metrics are essential to show executives how strategic HR actions, adjusted according to these insights, can enhance overall organizational performance. For this reason, the issue around metrics has been profoundly discussed in the literature, which has led to a strategic improvement of its fundamentals, which will be further extensively explained.

However, it’s important to make clear that predictive HR analyst practitioners don’t rely on a strict taxonomy of standard metrics or measures as a reference model valid for all companies, as each HR initiative would entail indices and measures to apply that would be specific for devising a correct forecast, and present picture, of the organization's workforce trends (M.R. Edwards, K.Edwards, 2019). For this reason, infinite different sets of measures have been described in the literature to explain employee performance and other measures. For example, the Human Capital Metrics Handbook (2013) supplies a comprehensive taxonomy of over 600 different human capital measures (CIPD, 2017). Besides, in order to avoid problems related to mono-metric targeting, it is suggested that any efforts to model adequate metrics include taking into consideration and computing a multiple variety of data indicators and measures at the same time of analysis, so as to be provided with exhaustive information and gain a full range of judgments for any metrics-related aspects that contribute to the overall organizational performance (M.R. Edwards, K.Edwards, 2019).

2.4.2 The three levels of Metrics in HR

Despite this specificity issue, generally, HR metrics utilized by organizations are mainly divided into three major levels of metrics (Boudreau & Ramstad, 2002). Only by collecting and using all these three levels, could companies acquire both an accurate strategic and operational picture of their workforce-related activities. In particular, some of those measures are purely descriptive in nature, meaning that these are used to address HR matters as they emerge in a generally unstrategic way. Examples are extrapolated from the answers to questions that aim at informing HR operational policies such as: 'How many people are currently employed? What is their salary level? How many more candidates shall we need to onboard to open a new function?'.

More precisely, the first metrics level is the so-called efficiency metrics and is readily connected to the existing accounting system of each organization. Basically, efficiency measures tell analysts how efficiently HR performs its basic administrative processes and activities, focusing on productivity and cost (J. H. Dulebohn a, R. D. Johnson, 2013). They get managers to reflect on how resources should be allocated within HR to optimize the overall operation of the HR function as they mainly include measures such as turnover, profit, and labor costs. At least, those are relatively easy to compute and then are compared to benchmarks provided by multi-company databases. In fact, they are measures of output over a span of time (hour, day, week, year) that is associated with particular units (single, total, or average employees or teams). For example, 'cost-per-hire returns the efficiency of the recruitment process shows how much it costs the company to hire a new employee, dividing the recruitment budget in a given period by the number of employees hired in a given period (CIPD, 2017). Other metrics of this kind correspond to "profit per employee", "labor cost factor", "Human capital value-added revenue". Having said that, it's important to stress that efficiency metrics do not consider the quality and effects of HR policies (Lawler et All., 2004). For this reason, companies who rely solely on them are most likely to gain only short-term cost savings or productivity increases and not all the other long-term benefits resulting from other approaches, and that is why this type of metrics are still considered descriptive, as they are not based on linking concepts (Boudreau & Ramstad, 2002).

In addition to measuring costs and the value of human capital, metrics also focus on the value of HR practices or programs (CIPD, 2017). The second level of metrics is the effectiveness metrics or HR cost-benefit metrics. This level verifies if HR policies result in the intended outcome on the workforce that they are directed toward, to ultimately define which practices have been effective and which haven't. They want to measure, for instance, whether a training, provided to workers, has resulted in the development of desired skills or not, that is the metrics called training effectiveness (Lawler et All., 2004). Another instance is measuring the worker’s career advancement concerning development plans. Basically, common metrics of this nature measures strategic skills, core competencies, and other employees’ attributes that judge as effective or ineffective the HR policies applied. However, understanding the return-on-investment of HR programs could be highly useful but may tell little about the synergies among HR policies and the overall value of measures in enhancing decisions about human capital, assessing only how HR teams are performing at the present moment (E. Lawler et All., 2004). However, these premises are different from HR Analtics’s utmost goal that is to build unbiased patterns and forecast scenarios to inform strategic decisions on a firm’s talent pool (Levenson, 2011). Taking an example about the area of diversity&inclusion within a firm, efficient KPIs would make an HR manager notice only if the number of workers from different ethnic backgrounds has increased over a period, without suggesting the root causal factors behind this increment or without grasping the future impacts that the latter would have on other measures, like on the turnover rate (M. Nocker and V. Sena, 2019).

Ultimately, the final level of metrics is impact, or strategic HR metrics, and specifically measures HR's impact on financial, customer, process, and people outcomes. Impact metrics are designed to answer human capital-related questions to prove how HR activities affect the organization’s ability to obtain and maintain sustainable competitive advantage (Lawler et All., 2004). Those measures are related to worker attitudes, behaviors, competencies, and culture that are critical to organizational performance: they are concerned with building “causal chains” between HR interventions or employee characteristics and business processes or results (Boudreau & Ramstad, 2002). Thanks to this characteristic, those measures are more strictly adjusted to a firm’s strategy, presenting, therefore, less commonly shared indices among companies. An example of such questions for improving the organization focus on the benefits learning and development initiatives bring, is: “What are the impacts of our training programs on productivity?”. In this case, what this index aims to is not only quantify the direct and indirect cost of the specific training program, but also identify what measures its impact. These could be measures of revenue, improved customer feedback, or any other quantification of additional value to the business, compared before and after the training was delivered (CIPD, 2018). In addition, as we have just noticed with this instance, impact metrics typically go beyond simple ratios of the HR system, into business outcomes, involving an integration process of HR data with other organizational data, giving outcomes that are more customer-focused, process-related, and financial. In this way, companies are greatly enabled to answer questions that involve demonstrating, in measurable terms, a causal relationship between a particular HR metric and other metrics built by other functions of the organization, ending up possibly developing predictive models, based on the reference metrics studied, that will contribute to sustained competitive advantage in managing and deploying their workforce. On top of that, they are mostly implemented by complex statistical techniques and computational models. For example, an impact metric could be based on a multiple-variable model linking employee engagement by age or job tenure, to customer satisfaction, which should provide management with leading indicators of organizational performance.

2.4.3 Examples of Human Capital Metrics

Having explained differences among the three levels of HR metrics, importantly, HR analysts are used to concentrate their efforts to study statistical relationships among a set of metrics that are categorized based on the area of HR interest. Inside the recruitment sphere, as a matter of fact, other than "cost per hire", already mentioned above, there are many more indices, like "time per hire" (which measures the number of days between a position opens and a candidate signs the job contract), "candidate-job satisfaction" and "candidate experience" (that investigate around whether the expectations set during the recruiting procedure have matched reality or how the whole recruitment process has been perceived), "time to productivity" (measuring how long it takes to get people up to speed and productive) and so on. On the other hand, elaborating some turnover metrics is crucial for HR analytics team, like "monthly turnover rate", "separation rate", "employee absent rate", "stability"/"instability rate", and "survival rate". In particular, organizations should be careful to distinguish the type of lever included in this statistic (retirees or voluntary leavers, for example) and generate a formula that fits the company settings and truly indicate the actual cost of replacing a member, including the cost of induction and training of the new hire. Within the diversity&inclusion area, examples of diversity scores frequently used are those aimed at discovering the percentage of a specific human gender or of ethnic minorities that fill certain senior jobs or project teams, that have been promoted in a given period, or the percentage pay gap in comparison to the rest or to the total of the company population (M.R. Edwards, K.Edwards, 2019). Besides, concerning talent retention in critical units, individuals' performance ratings could be mapped against a measure of "potential" for each employee, based on different categories of them both, discriminating which workers are underperformers, who seems to be a valued specialist, and who instead could be an emerging potential talent or a recognizable top talent. This greatly helps in making strategic decisions whose goal is to strengthen the company's organizational development capabilities, devise succession planning, and adjust career paths accordingly (BCG, 2021). Ultimately, strictly related to performance, engagement metrics might be among the most important 'soft' HR index to control as they allow assessments of key employee attitudes that are widely considered in the literature to be predictive of behavior that contributes to greater organizational performance. Even if engagement levels are mainly collected at the individual level through survey questionnaires, because of confidentiality issues engagement scores are then mainly analyzed at the aggregate level, calculating, for example, the percentage of a work unit that has agreed or strongly agreed to certain engagement measures. (M.R. Edwards, K.Edwards, 2019)

To conclude, as the research run by Lawler & Boudreau (2015) indicates, the use of impact metrics is still low in comparison with effectiveness metrics, that are gaining popularity, and to efficiency metrics that are largely the most applied ones. What is advisable for organizations, though, is to develop an alternative approach that, rather than just showcasing a standardized set of traditional HR metrics, integrates the use of metrics and relative KPIs as part of Workforce Analytics activities for achieving a bigger-picture data-driven understanding of the relationships between qualitative and quantitative HR indicators and organizational business performance (W. Cascio & J. Boudreau, 2017). For this to happen, C-level is called to move from the first two levels of metric to the impact metrics to finally reveal the connection between HR metrics and business outcomes, relying upon the use of more complex statistical techniques and predictive algorithms.

2.5 Advanced HR Analytics techniques: Machine Learning

It has been stressed throughout this chapter that, in addressing the challenge to handle and analyze massive and heterogeneous amounts of data, companies are increasingly relying on sophisticated techniques; leveraging solid predictive and prescriptive means of analyses they are enabled to keep gaining new valuable insight, improve their level of computational efficiency and assist their consolidated HR processes in a more result-oriented way. Among the ways to apply AI in decision-making and in the field of Big Data analytics, in this section I will first focus my attention on explaining the concept of machine learning. Subsequently, over the second part of the section, I will introduce the basics of the most widespread supervised learning modeling techniques utilized for predictive scopes, which, in the further chapter, I will practically use in my analyses on the field.

2.5.1 Machine Learning concept

Machine learning is a modeling method, enabled by computational systems, that derives from interdisciplinary research disciplines (such as statistics, information theory, engineering, AI, data science, etc.), that have grown largely over the past two decades; its main function is to devise fast and efficient algorithms and models to predict future outcomes through a process of learning from historical patterns identified in Big Data sets (S. Patel, 2017). Making inferences on the future based on previous patterns, Machine Learning is categorized as part of the predictive analytics world and among the data mining techniques: its main feature consists of discovering relevant information and of making intelligent decisions automatically (McKinsey Global Institute, 2016). In addition, machine learning applications present wide-ranging potential and capabilities to solve big data problems in comparison to traditional technologie: for this reason, machine learning frontiers are impacting many fields, including fraud detection, social media analysis, supply chain, medical diagnosis, recognition systems, finance, informatics, and more (S. Feng at All, 2016). Unlike conventional software where humans specify the instructions on the tasks they need to execute, the underlying concept of machine learning is to build a learning algorithm that is trained by analyzing a portion of dataset observations, - the “train dataset '', without using explicit instructions. The algorithm then continuously processes that data until it learns, recognizing patterns between variables in data, to develop an approximated model that forecast desired variables from new data given as input (R. Hamilton, W. Sodeman, 2020). For these characteristics, these systems become relevant in complex situations where there's not enough information and it's difficult and time-wasting to repeatedly recreate the model manually (J. Berral-García, 2016). Ultimately, machine learning tasks are typically classified into three subdomains, depending on the purposes for the modeling and the level of human interaction they require to operate: supervised, unsupervised, and reinforcement learning.

2.5.2 Supervised Learning

Generally, on supervised learning, the aim is to build a model from observed examples a-posteriori, which should be able to operate classification, estimation, and regression of new examples a-priori (J. Berral-García, 2016). With these methods, a human has trained the machine. In fact, this process implies extracting an inferred function from a labeled pre-built training dataset that is provided to the machine's algorithm. Each set of training samples consists of an associated input value and the desired outcome value, so that the machine knows what is looking for. Eventually, the machine learns to guess and plot new values of answers when given new data (S. Patel, 2017). If the expected output is far different from the one resulting from the machine, the algorithm's parameters can be modified and new calculations can be done with the same data to reach a closer conclusion, improving its accuracy (M.R. Edwards, K.Edwards, 2019).

Supervised learning can be divided into classification and regression. Classification, in particular, is to design a classifier able to predict the classes of output values, given a training set of input data, whose type of variable is categorical, or discrete (L. Wang & C. Alexander, 2016); for example, given the inputs role, function and performance ratings, the classification process will forecast if the turnover risk of an employee will be high, medium or low. The classification models are said to give a qualitative response. Classification algorithms usually consist of a learning phase, when the model is trained and a predefined number of classes or groups are clustered based on a set of observed attributes, and a classification or testing phase, for the verification of the model and the prediction of which class labels the new data belongs to (X. Zhang, 2020).

Respectively, the regression process predicts a value of dependent output variables from continuous input samples from independent variables (M.R. Edwards, K.Edwards, 2019); regression problems return then, differently to classification, a quantitative response. As an instance, given the input average team payroll level, team leadership index and team life duration, the regression process will forecast the average team performance, which will be a number on a continuous scale.

2.5.3 Unsupervised Learning

Unlike supervised, on unsupervised learning techniques, the input data set is given to an algorithm without training labels and the environment only provides inputs without desired targets (S. Feng at All, 2016). The machine teaches itself without any human help, highlighting the real power of AI. The model is then built to draw inferences from observed input examples but it's the algorithm that freely analyzes them trying to make sense of them, thus returning output labels that describe the assigned data properly, basically finding hidden patterns, relationships, or groups in the data. With this method, the estimated outputs correspond to the new knowledge that the algorithm presents (J. Berral-García, 2016). Like it happens to us humans, as we realize to have made a mistake, the machine improves its algorithms to reduce possible errors to occur once it properly understands the patterns among data (M.R. Edwards, K.Edwards, 2019). Unsupervised learning is divided among clustering and association rules techniques. Clustering is when the machine partitions un high-dimensional unlabeled datasets into inherent clusters or groups based on a high degree of similarity of attributes among intracluster and a high level of diversity among intercluster (M. Mani et All, 2021). K-means and hierarchical clustering correspond to the most popular used unsupervised clustering algorithms. Rather, association rules, which will not be further explained, aim at establishing if-then rules and relationships among variables that describe a large dataset.

2.5.4 Linear regression

The most common type of regression analysis is linear regression which is a very straightforward technique for supervised learning, included in the ultimate level of statistical inference, that enables to predict a quantitative dependent dimension Y, that is necessarily continuous, on the basis of one or more input variables X, that could be either continuous or categorical to suit the model (G. James et All, 2017). While correlation analysis seeks the strength of linear association between two variables, regression analysis instead assumes the existence of a linear dependence between the dependent response variable, meaning that one may be predicted in the model by one or more predictor independent variables (G. Tripepi, K. Jager, F. Dekker, C. Zoccali, 2008). Linear regression algorithms, to describe the relation among the corresponding variables, produce the general following mathematical equation: $Y \approx b0 + b1X + \dots + bnX + \varepsilon$, where the symbol "ε" refers to an an unobserved distrurbative random variable that contributes to model the linear relationship between the dependent variable y and the nd the regressors **x**. The relationship can be portrayed visually by a scatter plot with a line running through it and will predict the expected increasing, decreasing, or direction of change in the dependent variable that can be associated with specific independent features (M.R. Edwards, K.Edwards, 2019). For example, stating that X represents a training course and Y the advancement in career, in this case representing our response variable, it's possible to regress the model and find how much participating at training increases the chance of being promoted inside the firm.

2.5.5 Logistic regression

Logistic regression analysis is another statistical and machine learning algorithm, similar to linear regression models, that models patterns of data predicting the likelihood of an event to occur, based on its relationship to a range of independent variables. Likewise linear regression, independent variables could be either continuous or categorical and are related, not exclusively in a linear way, to a dependent variable y that, conversely, must be categorical, and usually of binary nature, meaning that could have two only possible values, such as death or survival in case of clinical status (G. Tripepi, K. Jager, F. Dekker, C. Zoccali, 2008). The logistic regression model has in fact a linear form for the 'logit' of the success probability, corresponding to the logarithm of the odds. The resulting equation, where the sign of the parameter β indicates whether the curve ascends or descends, will take the data as input, and then will compute the estimated parameters and will give out an output of relevant information (M.R. Edwards, K.Edwards, 2019). An example of logistic regression is the investigation of the HR department that aims at discovering what factors have impacted the fact that an employee has left the organization or has remained, relating this binary variable to some work-related characteristics like his performance criteria, contractual status, or salary level. Ultimately, if both linear and logistic models involve two or more independent variables that account for a variation in a single dependent variable, these are called ‘multiple’ regression models.

2.5.6 Support vector machine

Support vector machine (SVM) is another most widely used, discriminative, supervised ML approach, used for both classification and regression models, that attempts, given labeled data, to get the best possible space to categorize the classes in training samples (R. Gupta, S. Tanwar, S. Tyagi, N. Kumar, 2020). In particular, these machines are sometimes called linear binary SVM classifiers as they find linear classifiers in a bi-dimensional feature space where the data points are mapped and the different classes are separated by a line; otherwise, in case classes are divided by a hyperplane in a higher dimensional space called ‘Kernel’, meaning that the number of features given to fit the model are more than three, we should talk about non-linear SVM, and the feature space is enlarged. The model simply draws margins between the classes, categorizing the data set in the two sides of the hyperplane as they were two classes; when a new data is analyzed, the algorithm then forecasts which class, or which side of the hyperplane, it belongs (D. Sisodia, S. Vishwakarma, A. Pujahari, 2017). In addition, the distance between the margin and the hyperplane is as wide as possible to lower the error and inaccuracy in classification (B. Mahesh). SVMs have very good performance levels for datasets of moderate size (L. Wang & C. Alexander). This instrument, then, allows multidimensional aggregation in great simplicity, as, thanks to the creation of hyperplanes or separation lines, the placement of a specific data within a category arises from the comparison between that data and all the others, not only from the comparison with the average of the data covering characteristics similar to that data with respect to a reference variable.

2.5.7 Decision tree

Decision tree is a useful and popular nonlinear predictive supervised learning technique that classifies data based on their feature values (L. Wang & C. Alexander, 2016). Also, it's possible to build regression trees that, differently, make use of continuous target values. Decision trees consist of two phases: during the training phase, the algorithm generates the classification by partitioning, in a recursive manner, the training dataset using a top-down approach in which attributes are selected in sequence so as the training data is organized into a tree-structure plan; then, in the test phase, a decision tree is used to determine the most commonly occurring class, corresponding to a specific terminal node region, to which a new sample belongs (G. James et All, 2017). Basically, a decision tree is a graph representing classification choices in a tree-like structure (B. Mahesh). A decision tree is made of a root, or decision nodes, representing a test on the attribute that is to be classified, whose edges represent a series of test conditions, that work with splitting criteria to expand the leaf nodes, that represents class labels, into which the data subsets are partitioned (L. Wang & C. Alexander, 2016). The splitting criteria consist of simple equations based on some quality measures such as information gain and entropy, associating weights to the the features used and producing automatically rules to apply for the prediction through the iteratively repeated process of splitting that make the trees grow (L. Liu, P. Story, S.Akikineni, C. Davis, 2020). Decision trees can be easily interpreted and explained even by a non-expert, but their results often present low error and high variance. Moreover, trees can be very non-robust, meaning that a small change in the data can cause a large change in the final predicted tree, other than being unsuitable for handling Big data sets (G. James et All, 2017). In fact, decision tree has the occurring problem of overfitting, the phenomenon by which the tree comes out of the model all alike, that allows this algorithm to perform well only in the training phase with the data on which the tree has been fitted, while when giving new data in the test phase, the same trained algorithm seems often unable to understand how to make the prediction since it doesn't generalizes results. For this reason, these types of algorithms are sometimes defined as 'weak learners'. On the other hand, a model is prefered to have enough degrees of freedom to manage the underlying complexity of the data and to be successfully accurate in its predictions, but not too much freedom to avoid high variance and be more robust

(J. Rocca, 2019). Though, this trade-off balance between bias and variance of the trees can be eluded with random forest techniques, where best parameters values could be found to fit a model which has low bias and high variance (C. Hegde & K.E. Gray, 2017).

2.5.8 Random Forest

Random forest is one of the most common and most powerful supervised machine learning techniques that have the advantage of being able to perform both regression and classification tasks, as the decision tree algorithms. Random forest algorithms are based on grouping a combination of several tree classifiers on subsets of training data, rather than a single decision tree, for classification and regression tasks. Explaining deeper how random forests work, usually, to avoid overfitting to happen and converting weak learners into strong ones, two techniques are used when dealing with random forests: 'Bagging' and 'Boosting' methods. Bagging is an algorithm that is used to improve the robustness of a a tree classifiers with several leaves, as, considering concurrently homogeneous weak learners to train, learns them independently from each other in parallel and combines the total set of result processing the average of the predictions in order to obtain a model with a lower variance (J. Rocca, 2019). However, fitting fully independent models because it would require too much data. Therefore, with the 'bootstrap method', the algorithm creates a large number of samples B from the initial dataset of size N by randomly drawing with replacement B observations (J. Rocca, 2019). This way, each tree, trained on a different random dataset, made of random observations, to simulate the totality of observations, will grow independently from other trees: in the classification model, this has the improving effect of de-correlating the trees and their results won't be dependent on the data I provided to create the model. _Having said that, bootstrap datasets share several instances and are somewhat similar, so, in order to grow trees that are dissimilar to one another, some different 'Boosting' algorithms are performed. These take the average of the responses of all the random trees that learn sequentially, rather than contemporarily, thus reducing the chance of always having the same results on its components, letting different features that might not have a high bearing on the end result of each tree emerge. With trees not looking at the same information, this process does not change the expected answer but reduces the correlation between the different returned outputs (J. Rocca, 2019). In addition, it leads to a substantial reduction in the bias over a single tree, that yet could bring to a model outbound overfitting, corresponding to a raise in variance. On the other side, a reduction in variance is reached by applying the methods explained so far that favor the model to learn efficiently and grow diverse trees (as their results are not dependent on the data provided in the training phase because of random input selection) even though the level of bias increases, but marginally. Generally, the more trees in the forest, the more robust the samples' class prediction and thus a better trade-off resulting in a marginally lower accuracy but considerable lower variance that prevent the model from overfitting, yet with good tolerance for abnormal value and noise (L. Liu, P. Story, S.Akikineni, C. Davis, 2020). For all these reasons, these specific types of random forests' algorithm of bagging and boosting are called the 'ensemble method' of the decision trees, that indicate a machine learning paradigm where multiple weak models are correctly combined to obtain more precise and/or robust models than normal decision trees (J. Rocca, 2019).

2.6 Best and worst real-world HRA examples

In this section, I provide real-world examples of best global organizations that lead the way when it comes to people analytics: these firms have put additional effort to overcome the basic descriptive reporting activities, recognized to provide effectiveness, though without completing the whole workforce picture. For these firms that woke up very early with respect to the Workforce Analytics world, the development and incorporation of more advanced practices in their decision-making processes have resulted in significant company business outcomes, benefits to the workforce, and increasing sophistication of the work. On the other hand, cases of bad examples of application of machine learning algorithms, concerning decisions about workforce, will be also illustrated to understand what went wrong in the process and which path to undertake to outline prevention work in this sense.

2.6.1 IBM case study

In the first place, I would like to present the virtuous case of IBM, which represents one of the frontrunners in the domain of HR Analytics. IBM, during the last years, has put a lot of technological efforts to make progress in Big Data analysis and machine learning that now grant the company, in its models, the ability to isolate variables to find out which are responsible for considerable insights. Intending to enhance motivations among its employees, for instance, they have implemented a text sentiment analysis, combined with cluster analysis, that could extract from unstructured data, taken from social media content and annual surveys, reliable engagement trends about its workforce population, based on employees' feelings over time (N. Guenole & S. Feinzig, 2018). Although trusting in this automated algorithm, the human element was not completely set side but played an important checking role on the emotion-sensing algorithms that come out of the software: a small team of analysts routinely scrutinized the identified patterns to verify if they were sufficiently analytically-reliable before sending them up the chain to the C-suite level. The project was named "Social Pulse" and was mainly meant to just continue listening to employees without being intrusive in their private life, making sure IBM's code of ethics around the use of AI in HR was always applied (N. Guenole & S. Feinzig, 2018). Ultimately, positive or negative emotions could then be tracked through this software, and actions taken accordingly to sort out any problematic situation. Eventually, among the patterns, it was discovered a widespread complaint about the performance system which graded employee performances on a curve. Thanks to that, the process has been promptly rebuilt according to a forum, arranged by the HR department, to gather proposals for a new system to engage more all "IBMers" (K.Waddell,2016).

Another area of HR analytics advances for IBM is training and development: the firm has invested in this field as it thought that, with an accurate assessment of employee knowledge, they could save great amounts of money. IBM's AI innovations applied in the learning context have then contributed to maximize skill development at the individual level, and to optimize the overall acquisition of strategic skills at the organizational level (N. Guenole & S. Feinzig, 2018). One of the innovations delivered was the AI tagging of learning content, which contributed to making learning easily available at any time, and a real-time skills inference; in fact, through these initiatives, IBM began to deliver helpful training materials to employees when they actually needed it. More than that, they successfully envisioned a sort of personalized learning dashboards, a sort of skill profiles created analyzing employee data validated by managers (N. Guenole & S. Feinzig, 2018). For example, instead of wasting employee time by making everyone participate in compliance training each year, IBM, ahead of time, could sort out who already knew the regulatory standards. Moreover, IBM has been enabled to implement specific intervention plans to progress towards closing the identified skill gaps in the business but also to discover skill sets that employees weren't aware of mastering. In this sense, to improve workforce planning, IBM used artificial intelligence to match the skills of individuals with internal job opportunities and development programs, thereby encouraging people to develop skills it needs for future growth (M. Nocker & V. Sena, 2019).

2.6.2 Google case study

Other than IBM, Google, one of the leading high-tech companies across the globe, is another great example of a pioneering company in the HR Analytics field, which has reaped enormous market outcomes leveraging a strategic focus for data and facts over opinions, feelings, and intuitions, in order to complement human decision making. What is striking is the profound difference in how Google, relying on insights from data, deals with all issues related to people. Firstly, all its initiatives are results of an intensive process of change in culture toward a more data-driven decision-making approach, that began establishing the People & Innovation Laboratory- the "Pi-Lab". In this way, psychologists, data scientists, and academic researchers carried out applied research on organizational innovative practices at the HR level, to verify whether some theories were accompanied by the support of science and data (S. Shrivastava, K. Nagdev & A. Rajesh, 2018). The multi-year project "Oxygen", started by Google in 2009, is, by far, the one that has paid off the most, and is about leadership capabilities. This predictive research, containing 10,000 rows of internal employee data and more than 100 variables (including a detailed analysis of complaints and praises mentioned on performance reviews, phrases of top manager award nominations, and qualitative comments from employee feedback surveys), has identified a pattern of good-manager traits and behaviors that lead to more engaged staff and higher productivity (J. Sullivan, 2013). This analysis provided a critical understanding of what employees expected in terms of qualities from their leader and also gave guidelines to better target the roles and responsibilities of managers, essential for the overall company performance. Moreover, Google then leverages the resulting eight characteristics to architect management training&development programs; these behaviors became popular in the day-to-day working style of managers, and were utilized for managerial roles interviews and to search career advancements decisions (J. Kaur & A. Fink, 2017). Ultimately, project Oxygen surprisingly revealed that, more than as technical guidance, managers were considered by their under-level workers as important figures for coaching and mentoring. In fact, STEM (science, technology, engineering math) competencies did strikingly not appear on the final list of the 8 essential manager expertise, leaving room for many soft skills such as having good communication and sharing information, being a good coach (expressing interest in the employee and giving frequent personalized feedback), having critical thinking and problem-solving capabilities, have a clear vision for the team and insights into others, and so on. What's more, in the same way, with its "Project Aristotle", based on a deep teamwork analysis, Google found that the same soft skills mentioned above, plus "generosity, curiosity and emotional intelligence and safety constitutes the best recipe to create very productive, inventive and successful teams across high-tech environments" (J. Sullivan, 2013). To conclude, it's remarkable also the efforts done by this innovative firm toward the spheres of hiring and retention of employees, which have always represented one of the top priorities for organizations of this size. As a result, Google created, under "Project Janus", a mathematical algorithm, meant to replace subjective intuition and assumptions to predict which candidates were likely to perform well or which ones were likely to leave soon and, also, another algorithm for each large job family that reconsidered rejected resumes to reassess any candidates profile they might have missed. Developing "what if" analysis to continually improve their forecasts on upcoming people management problems and opportunities, they managed to take action before it was too late (David Green, 2019).

2.6.3 Nestlé case study

A third example of a company that strongly adopts HR Analytics plans in its agenda is Nestlè. Over the last 3 years, in particular, this firm has enabled all of its markets to start approaching, more seriously, milestones like gender pay equity and attrition risk in a statistically sound way. For what concerns the focus on devising a more inclusive workplace, using the statistical package "R", Nestle's HR analysts built a simple step-wise regression model, with the direct collaboration of their diversity and inclusion group, taking into account several variables such as age, grade, function, and talent rising, and then calculating whether all these correlate to gender with pay as an outcome. On the other hand, Nestlè also analyzed the data and made correlations between turnover rates and employee attributes to create a profile of the" common leaver" and afterward predicted the list of people who were likely to leave the company. After consulting data results, the team associated five main reasons why people were leaving: remuneration, leadership, recruitment and induction, leadership, and culture. One of the patterns their team identified was that women were leaving at a higher rate than men, regardless of their level of performance and that the head-office employees had greatly higher flight rates than employees in different hierarchical positions, and they adjusted their HR practices saving on the cost per hire and cost of turnover impacts (David Green, 2019). Lastly, also they used predictive models to test hypotheses on the optimal team size and to drive the structure of the organization for improving their workforce planning and quantitatively forecasting the number of heads and bodies and what skills they needed.

2.6.4 Machine learning inconvenients

Having said that, some machine learning algorithms also present some chronic problems with respect to their accuracy, as explained in the previous section. At one extreme, an algorithm could overfit when the machine identifies patterns that do not have significant validity because of high variance detached, rather, an algorithm is deemed as biased when it has been poorly trained and so inaccurately includes or excludes data in performing its prediction (R. Hamilton, W. Sodeman, 2020). Apart from that, as the algorithms in machine learning are entirely based on data, the quality data issue becomes really critical: once the models are trained with damaged and inappropriate data, errors and inaccuracies will be insistently present in the subsequent test outputs, as the machine, learning from scratch, won't recognize those errors, thinking they should be there on purpose, without then correcting the data on the final predictions. For this reason, these types of biases in the systems are regarded as implicit biases (M.R. Edwards, K.Edwards, 2019). Other than the quality, another issue with data in machine-learning algorithms is about the deficiency of empirical data: in most cases, it gets very difficult to adequately train a reliable model on the basis of small datasets, which can't effectively reflect the characteristics of the company population, preventing the algorithm from inferring accurate forecasts (D. Pessacha et All, 2020). Yet, most algorithms, that are designed with the "good" intention of enabling HR decisions based on objective data preventing operational bias and discrimination, can do better than humans if the training data are sufficiently noisy. A direct instance could be found in the recruitment process, where AI is essential to remove unconscious prejudices by ignoring information such as names, universities, places, and dates of previous jobs, yielding better candidates than assessments by HR recruiters. Still, a few bad experiments have occurred in recent times (A. Tursunbayeva, C. Pagliari, S. Di Lauro, G. Antonelli, 2021). As the most unfortunate case in point, in October 2018, Amazon was found by medias to rely on a sexist recruitment algorithm, through which it used to screen potential applicants, that was systematically and inappropriately favoring males over females. This biased model pusehd quite a time the HR department of the company to navigate unconsciously through an unintended gender-discriminatory pattern by excluding qualified women applicants when it came to finding new talents to onboard (M.R. Edwards, K.Edwards, 2019). This was made possible simply because rationally data showed that, in the past, a male employee would be considered a more successful hire than a female one, leading to consequences that have somehow seriously impacted the entire business. This happened accidentally, as the machine learning model, analyzing data about race, age, gender, sexual orientation, and disability of the candidates, was built in an incorrect way, with the purpose of searching for key patterns and terms in the infinite amount of CVs received by the company that has seemed to prioritize men. Exploring another unlucky example, also Google's Aristotle project, despite collecting multiple data on employees, initially didn't return consistent characteristics of successful teams, as expected (A. Tursunbayeva, C. Pagliari, S. Di Lauro, G. Antonelli, 2021).

To conclude, it seems like the main preconditions behind a useful and unbiased application of HR Analytics correspond for companies to make sure to build a solid data infrastructure and ecosystem while being able to answer the relevant questions on the workforce. Also, by identifying data complexities and contingencies, they would achieve a more sophisticated model that reduces the likelihood of involuntary discriminatory acts.

CHAPTER THREE

DATA ANALYSIS AND FINDINGS

3.1 Areas of Analysis for HR Analytics

This section contains the description of the major areas of application and functioning of HR Analytics, highlighting the benefits that descriptive analytics and prediction purposes could bring in these fields. They are: talent recruitment and selection, turnover and retention, employee engagement, training & development, performance management, diversity&inclusion.

3.1.1 Talent Recruitment and Selection

As talent is a major factor in determining the value of a firm, most of the HR analytics initiatives, so far, mainly draw upon the recruitment area to improve the hiring process accuracy, saving enormous amount of dollars on a year-basis, while automating repetitive tasks. HR managers need to accurately design streamlined hiring strategies so that the investment made in recruiting plans and campaigns will be fully returned once the right person, for the right job, at the right time is hired (W. Momin, K. Mishra, 2015). In the context of talent acquisition, HR Analytics & Big Data acts as a strategic component as it can be used to predict which specific attributes to be used as selection criteria so as to screen best-fit candidates (H. Tilbeç, B. Koçak, S. Köse, 2017). In fact, the most qualified and suitable professional applicants often do not necessarily have to be the ones with the most developed skills or the highest academic and professional background and achievements. Rather, sometimes they could turn out to be the ones who best reflect the corporate culture, who are armed with a resonating personality, or who are more likely to stay in the company for a long time and not to leave shortly after joining the company (Osservatorio HR Innovation Practice, 2016). In recruitment analytics, as examining a considerable amount of behavioral and attitudinal pre-hire and post-hire employee data with only human effort ends up being laborious and not comprehensive, digital technologies can personalize the experience of recruiters and measure candidates' desirable pre-established parameters through a bias-reduced and more efficient process. Importantly, HR managers should analyze and validate their selection methods, without insisting on the interviews which consistently risk bringing about the same pre-defined involuntary judgments about an individual because of a particular characteristic (for example, whether the candidate comes from a specific country) (M.R. Edwards, K.Edwards, 2019). On the other hand, they should rely more on the computational power of advanced machine learning algorithms that are able to judge everyone fairly across different settings. Moreover, these data analytics tools are essential in filtering the talent pool, by predicting new potential candidate profiles, who may contribute most to overall performance once hired, and in-house employees who may likely become a potential investment for the company, favoring an improved matchmaking between candidates and jobs (S. Mishra, D. Lama, Y. Pal, 2016). The ultimate goal is to forecast the traits that top performers seem to show as well as to better understand which factors influence the hiring process: updating its hiring filters, companies will have more chances to successfully identify future candidates who will perform similarly (N. Guenole, S. Feinzig, 2018). Overall, these activities wholly contribute to helping reduce costs, resources, and time to employ new staff for vacant positions

3.1.2 Turnover and Retention

Turnover roughly indicates the rate of employees leaving the company prematurely or performing below standard, representing a business-critical issue in organizations that has a variety of reasons to be a trend inside organizations (Corporate Research Forum, 2017). Overall, HR literature's evidences link turnover rates to organization-level performance indicators and employee morale, affecting companies' ability to leverage a productive personnel pool. Moreover, every time a worker abandons, a wide range of direct and indirect resources, time, and staffing costs are required to source, hire, and train new replacing employees that would fill that vacant positions. Therefore, this represents for any HR function a key area where sound analysis to investigate around the principal causes can make an incredibly valuable contribution in preventing employees from leaving, improving employee work experience and happiness level and then the retention rates (H. Tilbeç, B. Koçak, S. Köse, 2017).

To reach these goals, first of all, from a strategic perspective, organizations need a clear internal consensus that the metrics used to measure turnover are appropriate. With these, regular reporting activities of turnover statistics showing dashboards and spreadsheets containing percentages and proportions of employees that quit over a determined period is essential to realize the overall picture of the turnover trends present in the company. Nonetheless, these descriptive tools provide little insight on the important exploratory turnover predictors that should be modeled and monitored (M.R. Edwards, K.Edwards, 2019). At a higher level, HR analysts should then make a great endeavor, using inferential statistical techniques and predictive machine learning-based algorithms, to predict turnover risk hidden patterns. difficult for managers to target directly, across specific functions, business units, geographies, and countries and to understand the real employee-related and work-related factors consistently associated with resignations, ignoring unfunded speculation. This way, organizations can subsequently develop an effective retention management plan as well as implement the necessary intervention to anticipate problems and make adjustments on current misdirected HR policies (L. Liu, P.Story, S.Akkineni, C. Davis, 2020).

3.1.3 Employee engagement

Another pretty discussed area of concern and application for HR analytics is employee engagement. A significant amount of academic research, supporting the exploration of this very theme, has demonstrated that personnel exhibiting high engagement rates were found to have a strong and consistent pattern: they tend to be positively correlated with both task and contextual performance and negatively related to attrition risks (M.R. Edwards, K.Edwards, 2019). For this reason, it appears evident that having engaged employees inside a firm corresponds to one of the main targets which HR functions struggle to strive for; for this reason, over the last decades, companies have been increasingly putting great effort in trying to identify the many drivers of employees' engagement (Corporate Research Forum, 2017). However, there is no agreed definition of employee engagement: this construct seems to embrace some positive work-related vibes as enjoying the job characteristics, feeling proud of the role, taking energy from the work atmosphere, having a sense of purpose and mental motivations tied to organizational success (M.R. Edwards, K.Edwards, 2019). Therefore, not being a tangible concept, interpreting and ascentaining rigorously employee engagement trends within the enterprises is a dynamic process that is only possible by looking at many different metrics. Most importantly, in all levels of analysis it becomes essential to drive a continuous accumulation of data and real-time understandings of how people are feeling about their work (Kearney, 2020). As a matter of fact, organizations are now analyzing employee-related data collected through pulse surveys and external sources like social media profiles, in diverse ways, including text mining and sentiment analysis (S. Garg, S. Sinha, A. Kar, M. Mani, 2021). Also, employee engagement rates could be quite useful as an instrument for internal benchmarking, favoring the comparison between similar groups within the organization when carrying out clustering analyses on the workforce at hand (Corporate Research Forum, 2017). Ultimately, running engagement analysis on a workforce pool leads managers to adjust their business strategies to adapt their offer to the needs of employees, as happens with turnover analysis (H. Tilbeç, B. Koçak, S. Köse, 2017). Leveraging predictive machine learning analyses and making a proper use of HR Analytics, gleaning insights on the intangible factors that affect an individual employee's well-being at work, could indicate where resources should be directed to improve the working environment, and, in parallel, bring about many benefits related to achieving the business's success like an improved corporate image and an enhanced customer satisfaction.

3.1.4 Training&Development

Around the HR framework, training&development corresponds to a critical area of interest for HR managers that care about their employees' career development and would like to keep them updated in terms of learning competencies and motivated in boosting their performance, with the final result of generating more revenue. Expectedly, existing skills get outdated after a period of time, generating a need to learn new ones. In most cases, the measurement of the training initiative's effectiveness is assessed in terms of improvements of the trainees' performance and learning behaviors after they took that specific training; interpreting the results, training that seems to significantly enhance employee performance would be prioritized, while others might be left aside. In this context, the use of HR Analytics offers the promise to automate several steps of the training processes (S. Garg, S. Sinha, A. Kar, M. Mani, 2021). Machine learning models are built to recommend to employees, based on their relative needs, proper relevant courses and training methods, that are the most effective for defining customized employee development plans for the workforce, above all the youngest part of it, on which the company could finally make strategic and wise investments (Osservatorio, 2016). The offering of customized career development plans is of utter importance also for what concerns building strategic workforce planning capabilities backed up by rigorous supporting statistical analyses. Relying on data patterns, HR C-suite level may select with a far improved degree of objectivity, among its in-house talent pool, the ones who could successfully address current and future requirements. Moreover, other than making decisions on people's progression, predictive models may help in identifying ideal team compositions and the critical skills required to execute a change in strategic direction (Corporate Research Forum, 2017). Other classification algorithms assist companies to offer workers career guidance based on their forecasted occupational level at different time intervals over their career (S. Garg, S. Sinha, A. Kar, M. Mani, 2021). Whatsamore, novel types of analyses allow predicting the risk of poor participation in courses or non-compliance with deadlines for some compulsory courses. The predictive and prescriptive models, therefore, make it possible to make hypotheses of different scenarios to minimize the risk of non-compliance with compliance and at the same time to make the planning and scheduling of training interventions more efficient (Osservatorio, 2016).

3.1.5 Performance Management

Another essential area in which Big Data Analytics is effectively used in the HR framework is performance management. Performance management consists mainly of comprehensively evaluating employees' performance against their task duties and responsibilities, by considering various different perspectives that could uncover hidden employee patterns. HR analytics thus first assists the HR operators to track performance levels of its employees, monitoring the progress, according to a specifically chosen method; secondly, helps, through the insights inferred, leverage a well-designed individual performance plan that would successfully reward top performers among the overall workforce (W. Momin, K. Mishra, 2015). Performance appraisal, in fact, represents nowadays, rather than a stand-alone construct, an integrated system of measures and processes including consolidated customer feedback, sales figures, team productivity, 360-degree feedback, and many more (M.R. Edwards, K.Edwards, 2019). Though, once HR top managers have agreed on the most suitable metrics as a perfect gauge of employee performance to use in the process of performance evaluation, sophisticated analytics techniques play an important role in saving great expenditures of time with particular methods, leading to unbiased automation of the process. Recording a large amount of data such as the daily workloads as well as the levels of task achievement and fulfillment of organizational goals for each worker, both at the individual and team level, companies gain a better understanding of the workforce performance overview. At a higher level, as with employee engagement analyses, some machine-learning algorithms are able to cluster employees into distinct groups on the basis of their performance rates; comparing these findings to other indicators like the salary levels, then organizations verify if some unexpected situations are happening in the company, deciding whether opting for actionable HR interventions (S. Garg, S. Sinha, A. Kar, M. Mani, 2021). On the other hand, other models are aimed at accurately predicting where the performance rates would be at the peak, within different departments, on the basis of the employee background data and other attitudinal indicators, while some classifiers are greatly used in analytically forecast which are the most driving predictors on various levels of performance and which relationship these interrelated each other (S. Garg, S. Sinha, A. Kar, M. Mani, 2021).

3.1.6 Diversity&Inclusion

HR Analytics activities are essential to determining where problems and opportunities are occurring in the organization at the micro and macro level with respect to the area of DIversity&Inclusion, a theme that has been raising attention over the last years. Diversity embraces, in its general definition, the set of social individual features, accessible to HR personnel, like physical characteristics, behaviors, cultural origins, values, and experiences that underline a high degree of differences existing between the members of a group. Inclusion, on the other hand, corresponds to the realization of a job environment in which everyone is treated respectfully and holds equal access to opportunities and resources. Thanks to precious insights obtained by some analyses over these two social trends, companies are offering unconscious bias training, delivering leadership programs sponsorship for talented female or ethnic minority employees and building ethical large-scale culture change programs (M.R. Edwards, K.Edwards, 2019). In addition, beyond the moral obligation argument for diversity to ensure equality, dignity and fairness at the workplace, without any prejudices and stereotypes, there is a growing body of literature showing well-documented statistical relationships of leveraging a diverse workforce to greater business performance and increased profits (M. Astley, O. Cherkashyna, 2021). Importantly, the diversity business case has been found to enhance the corporate reputation, resource and skills acquisition, detection of diverse innovative and creative ideas, organizational flexibility, and team performance while reducing stress levels in the workplace. Having said that, when analyzing the workforce data through frequent descriptive reports, HR departments get to know their diversity statistics, and can start conducting benchmarking activities against competitors and analyzing the current descriptive snapshot characteristics of their workforce, with the ultimate goal to signal the possible occurrence of biases. Examples of this type of reports may include gender representation in a top management position, the disproportionate number of promotions within ethnic minorities, the percentage of diverse candidates dropping off in the hiring funnel, or the inequitable distribution of salary levels among employees clusters. However, most organizations are still missing what the numbers are truly meaning and how their HR managers should use this information to make sound decisions (M.R. Edwards, K.Edwards, 2019). Rather, only by running predictive models to identify causal patterns of team diversity, could the company effectively deal with that problem and diagnose which factors have led these evident biases to persist in the company, promptly intervening.

3.2 Process of analysis

In this section, each stage of my process of HR Analytics experiments will be precisely described and passed through, with the aim of providing a comprehensive theoretical and technical walkthrough of the research method that will be computationally implemented through some applicative cases of analyses further in the third section of this same chapter. In particular, contributing to building an analytic-oriented approach, all the procedures run over the main pre-analysis, analysis, and post-analysis phase will be also meticulously explained, illustrating some explicative figures and tables.

3.2.1 Target Definition

In the very first place, defining the most critical, yet general, objectives of an HR Analytics analysis or project correspond to the very first step of the process model which helps in conducting an effective analysis strategically tied to the top organizational goals. In this research thesis, specifically, as I couldn't afford any real HR department business-like dataset, my deepest intention has been to draw quite reliable and computationally-driven HR theoretical and practical conclusions, starting from a simulation of an HR dataset that I created on my own. These conclusions will be focused on some specific areas of personnel management, such as employee performance, workforce diversity, employee engagement, and a few more others; on these HR topics, some precious data-relevant insights, based on some advanced machine-learning predictive applications, would supposedly allow an HR executive to notice things that would not have been so apparent with the limited use of the only basic descriptive techniques. This descriptive approach, as stressed over the previous chapters, is used generally to read HR data inside organizations, but provides a partial overall picture of the workforce situation. By this analytics means, at the end of the process, the findings would subsequently lead to better adjusting the corporate HR policies to optimize the data collected on the referencing HR population, providing many benefits to the company, overall. This proposed novel-model work could then serve as a trace for any firm that would feel the urge to develop an HR analytical framework that could be implemented as a decision support tool for HR practitioners in real-world settings to efficiently manage the workforce of an organization.

3.2.2 Data Construction Technique and Sample

Data, as we have investigated all over the first two theoretical chapters, represents the initial fueling motor behind a resonant execution of an HR analytics endeavor, and then the most relevant aspect of any practical research which aims to validate a dissertation. Normally, inside the HR functions, both analytics professionals and HR managers can grasp all the necessary information, under the various structural forms of data, from their corporate databases, to start any sort of analysis they are willing to pursue. Likewise, doing analytics is not simply taking a bunch of data and giving it as inputs to some automated algorithms: the pre-analysis stage requires a noticeably long process of data preparation that could start only once data practitioners have identified what business-related objectives and needs are to be targeted. Then, once accurate data are consequently collected based on the variables of interest for the business issues under investigation, it is possible to run the relative analyses. However, while HR managers used to collect, store, and organize the information on their employees, derived from surveys, questionnaires, observations, interviews, the collection of data in this very project has turned out to be quite different. In fact, since during these months of working for the elaboration of this research thesis, I didn't obtain the opportunity to collect nor utilize for my analyses any proper real datasets on the workforce populations of a real company, I promptly decided to enrich my pre-processing efforts by constructing myself a realistic enterprise-shape dataset simulation from which conducting my descriptive and predictive analyses. So, after having primarily envisioned which were the insightful results I hopefully wanted to reach, my concern moved on figuring out which specific HR data I needed to base my analyses on, and, subsequently, on how should I rationally construct them accordingly; considering that my research aimed at providing an application example of HR Analytics, my goal was essentially to not overstep the boundary of reality. With these assumptions in mind, the experimental dataset was built by using a working file on Excel, coherently on the basis of the information HR departments used to collect on the various research papers and case studies present in the literature, which I studied, as a reference, before beginning this project. The theoretical assumptions I have inferred from the literature allowed me to efficiently choose, among the great multitude of employee attributes, which ones to judge relevant and suitable for my envisioned cases of analysis related to some critical area of Human Resources. To decrease complexity, I expressly opted to not include any aggregate or team-level employees data and consequently conduct my analysis completely at an individual level. What's more, since the engagement surveys from which I was supposed to generate ordinal observation (with values differing 1 to 5) were figments of my imagination, I avoided any confidentiality and anonymity issues as I didn't take any private information from a real company. Overall, from the selection of 54 different attributes, I started creating the large simulation of observations, for a total of 5,000 structured observations, that would have served afterward across the analyses. The final cleaned dataset simulation ultimately replicated an Italian company's workforce, made of exactly 5,000 employees. Additionally, I associated some general categories to divide the employee features identified for my analysis and make sure the spectrum of employee features I was selecting was wide enough. The main categories of employee data are the following:

- Basic social, demographic, and personal data
- Education experience and job-related data
- Payroll data
- Historical performance appraisal data
- Promotion data
- Training and development data
- Engagement survey data
- Satisfaction survey data
- Talent Profiling skills survey data

The result is an HLTM Excel file, whose columns represent the variable, or attributes, of the employees inside a company, and whose lines correspond to each employee's observation for each attribute. I chose Excel on purpose as I already had some experience with this software; besides, it is a well-known storing data system that has the capability to export data into standard file types and has good processing quality, though within some bigger data limits it may have caused problems. The details of each variable of the 54 features present in the dataset, along with the basics of statistics performed by the programming language application 'R' on the respective observations, are illustrated in the two tables below, developed in R. Specifically, Table 1 illustrates the continuous, or numeric, variables, while Table 2 the discrete, or categorical, ones.

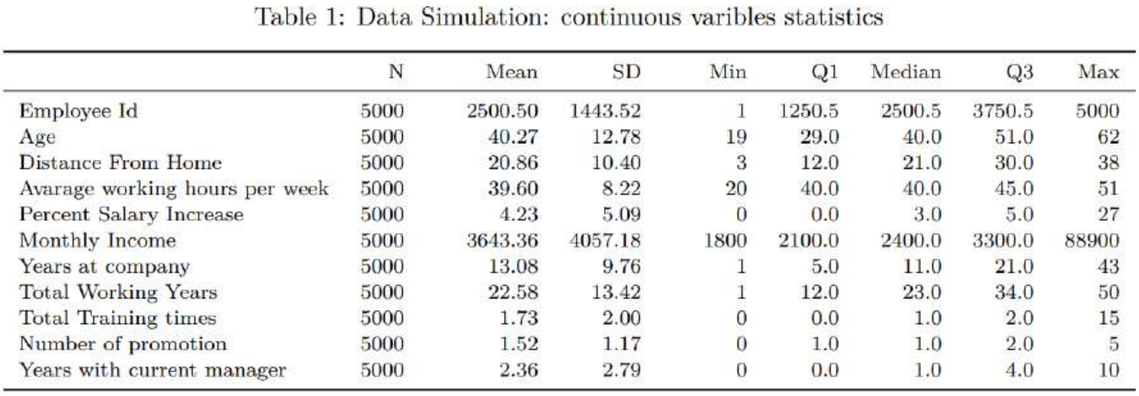


Table 1: Data Simulation: continuous varibles statistics

	N	Mean	SD	Min	Q1	Median	Q3	Max
Employee Id	5000	2500.50	1443.52	1	1250.5	2500.5	3750.5	5000
Age	5000	40.27	12.78	19	29.0	40.0	51.0	62
Distance From Home	5000	20.86	10.40	3	12.0	21.0	30.0	38
Avarage working hours per week	5000	39.60	8.22	20	40.0	40.0	45.0	51
Percent Salary Increase	5000	4.23	5.09	0	0.0	3.0	5.0	27
Monthly Income	5000	3643.36	4057.18	1800	2100.0	2400.0	3300.0	88900
Years at company	5000	13.08	9.76	1	5.0	11.0	21.0	43
Total Working Years	5000	22.58	13.42	1	12.0	23.0	34.0	50
Total Training times	5000	1.73	2.00	0	0.0	1.0	2.0	15
Number of promotion	5000	1.52	1.17	0	1.0	1.0	2.0	5
Years with current manager	5000	2.36	2.79	0	0.0	1.0	4.0	10

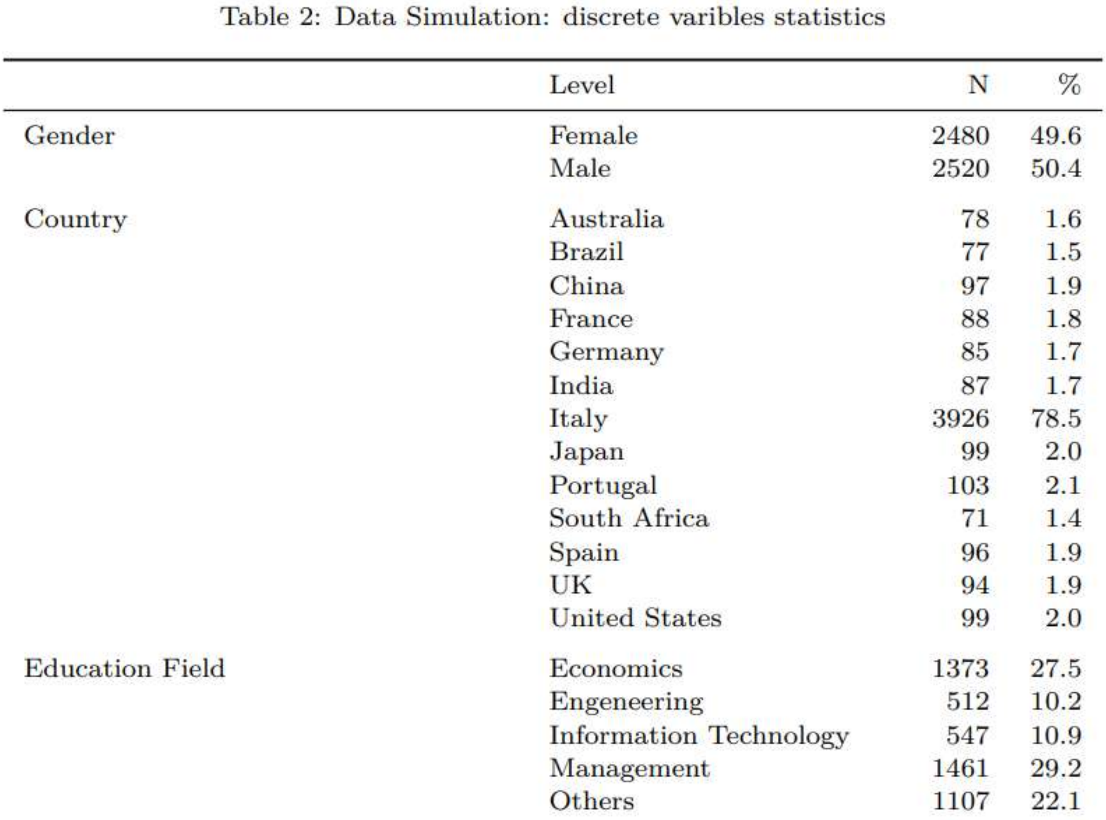


Table 2: Data Simulation: discrete varibles statistics

	Level	N	%
Gender	Female	2480	49.6
	Male	2520	50.4
Country	Australia	78	1.6
	Brazil	77	1.5
	China	97	1.9
	France	88	1.8
	Germany	85	1.7
	India	87	1.7
	Italy	3926	78.5
	Japan	99	2.0
	Portugal	103	2.1
	South Africa	71	1.4
	Spain	96	1.9
	UK	94	1.9
	United States	99	2.0
Education Field	Economics	1373	27.5
	Engeneering	512	10.2
	Information Technology	547	10.9
	Management	1461	29.2
	Others	1107	22.1

Highest Education Level	Diploma	2230	44.6
	Bachelor	2081	41.6
	Master	432	8.6
	Doctoral	257	5.1
Marital Status	Divorced	834	16.7
	Married	2037	40.7
	Single	2129	42.6
Have Children	No	2737	54.7
	Yes	2263	45.3
Employment Nature	Part-time	794	15.9
	Permanent Worker	3756	75.1
	Temporary Worker	450	9.0
Department	Finance	796	15.9
	HR	83	1.7
	Marketing	830	16.6
	Production	1693	33.9
	Purchasing	783	15.7
	Quality	312	6.2
	R&D	213	4.3
	Safety	290	5.8

Table 2: Data Simulation: discrete varibles statistics *(continued)*

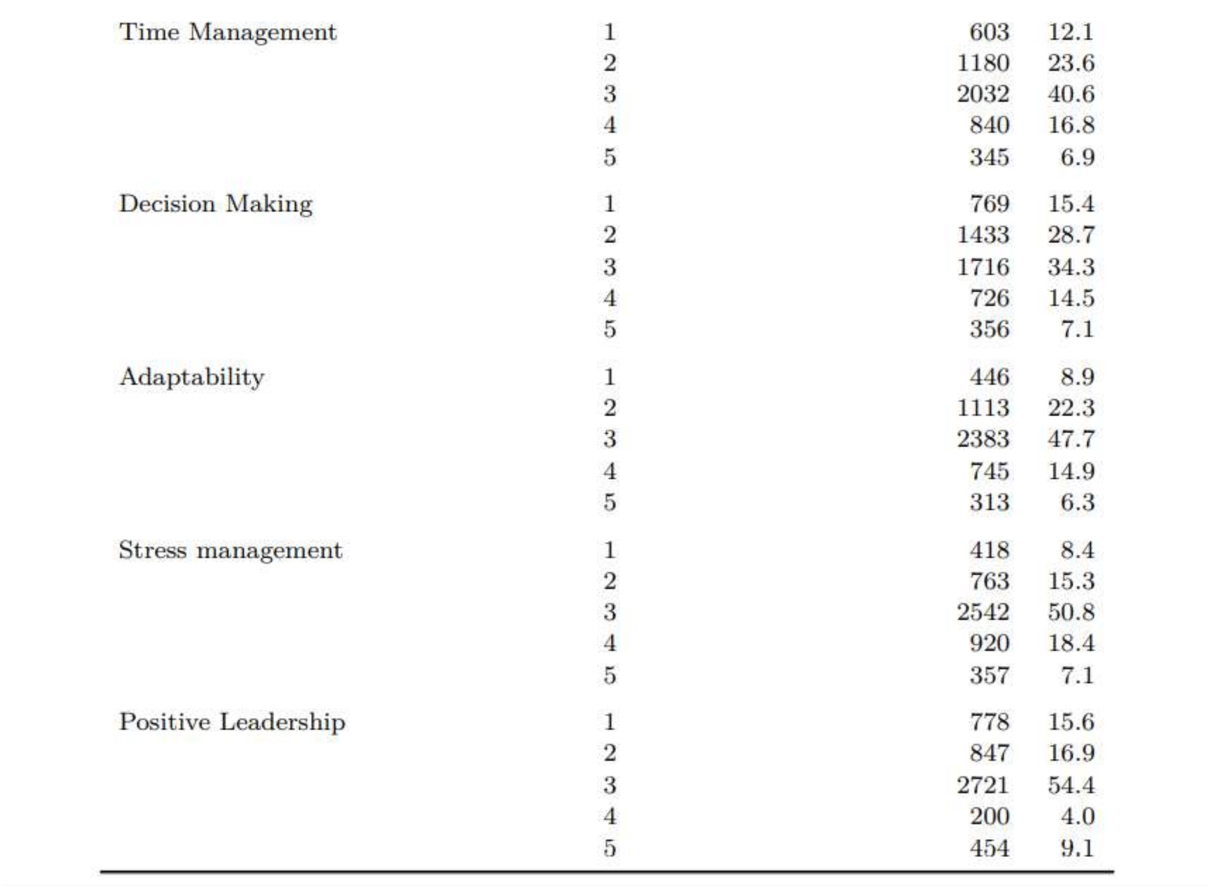
	Level	N	%

Management Level	Staff	3311	66.2
	Junior Manager	832	16.6
	Middle Manager	787	15.7
	Senior Manager	62	1.2
	Executive Manager	8	0.2
Over Time	No	3164	63.3
	Yes	1836	36.7
Historical Performance Rating	Very Low	209	4.2
	Low	1411	28.2
	Medium	1866	37.3
	High	1263	25.3
	Excellent	251	5.0
Collaboration and Cooperation	1	23	0.5
	2	594	11.9
	3	2244	44.9
	4	1283	25.7
	5	856	17.1
Commitment to personal development	1	749	15.0
	2	828	16.6
	3	2162	43.2
	4	496	9.9
	5	765	15.3
Avarage time to answer	1	456	9.1
	2	765	15.3
	3	2562	51.2
	4	906	18.1
	5	311	6.2
Supervisor Evaluation	1	739	14.8
	2	1225	24.5
	3	1492	29.8
	4	1181	23.6
	5	363	7.3
Conflict Management	1	412	8.2
	2	1342	26.8
	3	1100	22.0
	4	1085	21.7
	5	1061	21.2

Number of companies worked	1	564	11.3
	2	730	14.6
	3	1043	20.9
	4	924	18.5
	5	997	19.9
	6	742	14.8
Disciplinary Recalls	No	4278	85.6
	Yes	722	14.4
Job Involvement	1	229	4.6
	2	769	15.4
	3	2489	49.8
	4	1350	27.0
	5	163	3.3
Absenteism Rate	1	1485	29.7
	2	880	17.6
	3	1772	35.4
	4	515	10.3
	5	348	7.0
Job Characteristics	1	414	8.3
	2	782	15.6
	3	2551	51.0
	4	908	18.2
	5	345	6.9
Teamwork Cohesion	1	131	2.6
	2	578	11.6
	3	2220	44.4
	4	1300	26.0
	5	771	15.4
Corporate Culture and Vision	1	397	7.9
	2	773	15.5
	3	2570	51.4
	4	935	18.7
	5	325	6.5

	Level	N	%
Work Environment Satisfaction	1	208	4.2
	2	935	18.7
	3	1383	27.7
	4	1377	27.5
	5	1097	21.9
Job Satisfaction	1	269	5.4
	2	1138	22.8
	3	2478	49.6
	4	1025	20.5
	5	90	1.8
Pay	1	701	14.0
	2	1085	21.7
	3	1815	36.3
	4	925	18.5
	5	474	9.5
Workpace Security	1	873	17.5
	2	1290	25.8
	3	1749	35.0
	4	550	11.0
	5	538	10.8
Peer relationships	1	403	8.1
	2	805	16.1
	3	2564	51.3
	4	885	17.7
	5	343	6.9
Role Clarity	1	442	8.8
	2	968	19.4
	3	1905	38.1
	4	1031	20.6
	5	654	13.1
Supervisor Support	1	878	17.6
	2	1257	25.1
	3	1518	30.4
	4	984	19.7
	5	363	7.3

Career planning	1	953	19.1
	2	1282	25.6
	3	1070	21.4
	4	975	19.5
	5	720	14.4
Turnover Risk	1	575	11.5
	2	1033	20.7
	3	1975	39.5
	4	672	13.4
	5	745	14.9
Work life balance	1	511	10.2
	2	674	13.5
	3	2890	57.8
	4	429	8.6
	5	496	9.9
Learning and Development	1	1380	27.6
	2	1012	20.2
	3	930	18.6
	4	696	13.9
	5	982	19.6
Feedback and Recognition	1	315	6.3
	2	1332	26.6
	3	1809	36.2
	4	580	11.6
	5	964	19.3
Talent Profile	1	253	5.1
	2	1707	34.1
	3	1582	31.6
	4	1261	25.2
	5	197	3.9
Communication	1	22	0.4
	2	594	11.9
	3	2220	44.4
	4	1264	25.3
	5	900	18.0
Problem Solving	1	423	8.5
	2	759	15.2
	3	2512	50.2
	4	937	18.7
	5	369	7.4



Time Management	1	603	12.1
	2	1180	23.6
	3	2032	40.6
	4	840	16.8
	5	345	6.9
Decision Making	1	769	15.4
	2	1433	28.7
	3	1716	34.3
	4	726	14.5
	5	356	7.1
Adaptability	1	446	8.9
	2	1113	22.3
	3	2383	47.7
	4	745	14.9
	5	313	6.3
Stress management	1	418	8.4
	2	763	15.3
	3	2542	50.8
	4	920	18.4
	5	357	7.1
Positive Leadership	1	778	15.6
	2	847	16.9
	3	2721	54.4
	4	200	4.0
	5	454	9.1

3.2.3 Data Description and Methodological Assumptions

Studying the research literature and reviewing some past case studies where HR data were collected and manipulated for implementing some quite of analyses, in building my variables, I came to realize some important actions to consider to proceed with the creation of the applied dataset. In particular, if I expected the ultimate findings to be inferred from enough reliable patterns in the predictive models, which would be later fitted with my data, since my dataset was a simulation, I necessarily made sure that the majority of my variables in the dataset were interrelated by some specific relationships in the generation process of the respective observations on the Excel working spreadsheets. Within this theoretical framework, after exploring the build-in Excel functions, I assumed the best formulas to adopt to give some casual interrelations among my variables was to make use of a set of commands of this same very programming language, through which formulating some conditional logical statements in creatin each cell value. These are the following:

- The ‘RANDOM BETWEEN’ function, that returns a new random number that is between a bottom and a top range each time the spreadsheet recalculates. This was smartly utilized inside my formulas not only with numbers but with string fields, thanks to the use of the function ‘VLOOKUP’, the vertical lookup, with whom Excel could take into consideration in its calculations string values present even in different spreadsheets.
- The Normal distribution function, called “NORM.DIST”, that gives the probability that a random value falls at or below a given value of points on a normal probability density distribution, that, statistically speaking, is a curve that has a bell shape. I used this function, along with the “IF” statements that I’ll explain right below, to generate normal distribution-like curves of value for some specific attributes of the employee dataset with the aim to generate very realistic different data curves, changing the interval parameter points of the standard random curve. This way, each data curve taken into consideration has been shaped according to what I thought could be the best fit between the returning final values and a realistic scenario of those specific data trends inside a firm.
- The ‘IF’ function, that runs a logical test and returns one value for a true result (using “then”), and another for a false result (using ‘ELSE’). I figured that was also possible to run multiple conditions at the same time by “if” combining formulas together, one inside the other. In this way, the first “IF” Statement could appear inside the other one. By “nesting” the if formulas, technically speaking, one formula would then handle the outcome of another formula.
- Direct multiple if statement function, called ‘IFS’, that checks whether one or more conditions are met, and returns, for each statement, the value that corresponds to the first true condition among the total conditions stated, without returning any false value through the calculations. For this reason, IFS can take the place of multiple nested IF statements, and is much easier to read with multiple conditions.
- The ‘AND’ function, that is another logical function to test multiple conditions at the same time “true” or “false” depending on whether they are met or not.
- The ‘OR’ function, that is quite similar to the AND function but returns either true or false results. These two last functions were used to give logical sequence

among the statements written by the means of the other functions to generate observations, in such a way for Excel to always return a value that was not null for each possible concatenated reference variable.

As a next step in describing the generation process of the whole set of my simulative employee observations, I indicated, for each variable, how the latter has been built to give the closest possible similarity with the relative observation of a whatever company made of 500 employees, with a marked international prominence. In creating this formula, apart from a few independent variables that represented an exception to just start the generation of the data simulation, for most of the attributes I purposely connected, in a logical way, using the formulas listed above, many dependent attributes with other dependent ones as well as with many independent ones. This because, as the analyses were running on, I easily figured out that my predictive machine-learning algorithms in the first place, would have not returned reliable results and valuable accuracy scores as they couldn't have recognized any casualty pattern among the data; in fact, being the data generated only by random curve, every single observation would have had the exact same probability to come out as all the other ones. What's more, in this process, to think up all the causality connections, I needed some general methodological assumptions on how to generate some variables. More precisely, I studied in the literature which were the most influencing factors that affect the direct generation of specific grand-indicator variable, like employee satisfaction, engagement and turnover, simulating their final results as they had been gained by carrying out surveys and other things as it used to happen within real companies. Accordingly, each employee attribute was created with Excel formulas using the following methodological assumptions:

'Employee ID': ordinal sequence of numbers from 1 to 5000;
'Age': random distribution between 19 and 62 years;
'Gender': random distribution between the string values "Male" and "Female";
'Country': Of 5000 observations, I have chosen the first 1075 observations, considered the expatriates, to be created through a random distribution including the 13 possible country string values that didn't correspond to the value "Italy".

The rest of the observations, valuated as the locals, equal to 78,5% of the entire population, have been given the value “Italy”, which would represent the country where is located the simulation company’s headquarters;

'Education Field': 5 possible string values logically dependently correlated to the variable ‘Department’;

'Highest Education Level': 4 possible string values derived from a normal distribution but dependently correlated to the variable ‘Age’;

'Marital Status': 3 possible string values derived from a normal distribution but dependently correlated to the variable ‘Age’;

'Distance From Home': random distribution between 3 and 38 kilometers;

'Have Children': binary attribute (Yes/Not) derived from a random distribution but dependently correlated to the variables ‘Age’ and ‘Marital Status’;

'Employment Nature': 3 possible string values derived from a random distribution dependently correlated to the variables ‘Age’, ‘Gender’ and ‘Have Children’;

'Department': 8 possible string values created through a normal distribution;

'Management Level': 8 possible string values derived from a random distribution but dependently correlated to the variables ‘Age’, ‘Gender, ‘Number of Promotions’ and ‘Highest Education Level’;

'Over Time': binary attribute (Yes/Not) derived from a random distribution but dependently correlated to the variables ‘Age’ and ‘Employment Nature’;

‘Average working hours per week': random distribution between 20 and 51 hours derived from a random distribution but dependently correlated to the variables ‘Over Time’ and ‘Employment Nature’;

'Percent Salary Increase': random distribution between 0% and 27% derived from a random distribution but dependently correlated to the variable ‘Historical Performance Rating’;

'Historical Performance Rating': 5 possible string values dependently correlated to the variable ‘Country’ and directly generated from the weighted mean of the observations of the variables ‘Collaboration and Cooperation', 'Commitment to personal development', 'Average time to answer’, 'Supervisor Evaluation', 'Conflict Management', 'Job Involvement' and ‘Job Satisfaction’, each one generated through a different normal distribution;

‘Collaboration and Cooperation': 5 possible integer values (1 to 5) derived from a normal distribution; 'Commitment to personal development ': 5 possible integer values (1 to 5) derived from a normal distribution;

'Average time to answer'. 5 possible integer values (1 to 5) derived from a normal distribution;

'Supervisor Evaluation': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variable ‘Years with Current Manager’;

'Conflict Management': 5 possible integer values (1 to 5) derived from a normal distribution;

'Number of companies worked': integer value created through a normal distribution but dependently correlated to the variable ‘Age’;

'Monthly Income': integer values, in euros, derived from a random distribution between 1800 and 89000 but dependently correlated to the variables ‘Age’, ‘Gender, ‘Country’ and ‘Management Level’;

'Disciplinary Recalls’: binary attribute (Yes/Not) derived from a random distribution but dependently correlated to the variable ‘Absenteeism Rate’;

'Years at Company': integer values derived from a normal distribution and dependently correlated to the variables ‘Age’, ‘Total working Years’ and ‘Number Of Companies Worked’;‘Total Working Years': integer values derived from a random distribution and dependently correlated to the variables ‘Age’ and ‘Highest Education Level’;

'Total Training times': integer values derived from a random distribution between the value 0 and the corresponding value on ‘Years at Company’, and dependently correlated to the variable ‘Number of Promotions’;

'Number of promotion': integer values derived from a random distribution and dependently correlated to the variable ‘Gender’;

'Years with current manager': integer values derived from a normal distribution and dependently correlated to the variables ‘Years at Company’ and ‘Number of Promotions’;

'Job Involvement’: 5 possible integer values (1 to 5) directly generated from the weighted mean of the observations of the variables 'Absenteeism Rate', 'Job Characteristics', 'Teamwork Cohesion', 'Corporate Culture and Vision', and 'Work Environment Satisfaction', each one generated through a different normal distribution;

'Absenteeism Rate': 5 possible integer values (1 to 5) derived from a normal distribution;

'Job Characteristics: 5 possible integer values (1 to 5) derived from a normal distribution;

'Teamwork Cohesion': 5 possible integer values (1 to 5) derived from a normal distribution;

Corporate Culture and Vision':5 possible integer values (1 to 5) derived from a normal distribution;

'Work Environment Satisfaction': 5 possible integer values (1 to 5) derived from a normal distribution;

'Job Satisfaction: 5 possible integer values (1 to 5) directly generated from the weighted mean of the observations of the variables 'Pay', 'Workplace Security', 'Peer relationships', 'Role Clarity', 'Supervisor Support', and 'Career planning', each one generated through a different normal distribution.

Pay': 5 possible integer values (1 to 5) derived from a normal distribution; 'Workplace Security': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variable 'Employment Nature'; 'Peer relationships': 5 possible integer values (1 to 5) derived from a normal distribution;

'Role Clarity': 5 possible integer values (1 to 5) derived from a normal distribution; 'Supervisor Support': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variable 'Years with Current Manager';

'Career planning': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variables 'Number of Promotion' and 'Training Times';

'Turnover Risk': 5 possible integer values (1 to 5) directly generated from the weighted mean of the observations of the variables 'Work-life balance', 'Learning and Development', 'Feedback and Recognition', 'Job Involvement', and 'Job Satisfaction', each one generated through a different normal distribution;

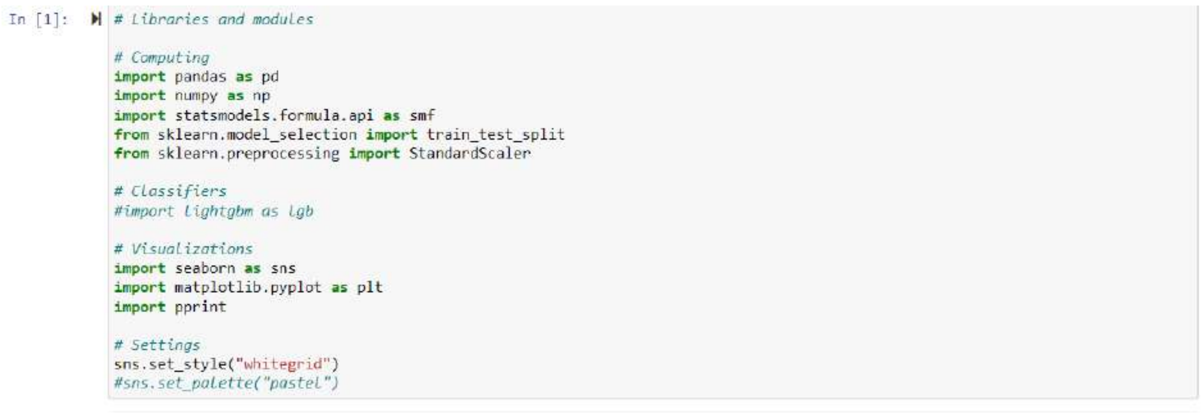
'Work-life balance': 5 possible integer values (1 to 5) derived from a normal distribution; 'Learning and Development': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variables 'Number of Promotion' and 'Training Times';

‘Feedback and Recognition': 5 possible integer values (1 to 5) derived from a normal distribution and dependently correlated to the variable ‘Years with Current Manager’;
'Talent Profile’: 5 possible integer values (1 to 5) directly generated from the weighted mean of the observations of the variables 'Communication', 'Problem Solving', 'Time Management', 'Decision Making', 'Adaptability', 'Stress Management', 'Positive Leadership’ and ‘Historical Performance Rating’, each one generated through a different normal distribution;
'Communication': 5 possible integer values (1 to 5) derived from a normal distribution;
'Problem Solving': 5 possible integer values (1 to 5) derived from a normal distribution;
'Time Management’: 5 possible integer values (1 to 5) derived from a normal distribution;
'Decision Making': 5 possible integer values (1 to 5) derived from a normal distribution;
'Adaptability': 5 possible integer values (1 to 5) derived from a normal distribution;
'Stress Management': 5 possible integer values (1 to 5) derived from a normal distribution;
'Positive Leadership’: 5 possible integer values (1 to 5) derived from a normal distribution.

3.2.4 Procedure of Analysis

This further section will announce the set of methods used for carrying out the empirical research of analysis through some software applications, covering also the data pre-processing methods needed before starting the analyses. First of all, regarding the research study, both qualitative and quantitative methodologies have been used. In fact, qualitative methodology was needed for collecting the employee attributes and, as explained before, selecting which, among all, could have been proposed as suitable predictors for other employee features; for example, which variables could contribute to producing the attribute ‘Turnover Risk’. Whereas, the quantitative methodology was utilized for generating the Excel formulas behind the data simulation, for generating descriptive reports and, ultimately, for applying the predictive models built to deeper analyze some specific variables of interest. Overall, all the computational efforts that I put in this work have lied on the world of Data Analytics, whose processes entail sorting, examining, and testing data at disposal to generate empirical results. Having said that, it's important to underline that the two main research designs and methods undertaken in this thesis are descriptive analysis and predictive analysis. First of all, In the descriptive analysis, known to represent the primary analytics level of techniques for organizations that don't have a sufficient amount of data or are not friendly with more sophisticated techniques, the principal goal, in the HR context, regards understanding the organization's workforce problems; these analyses are applied with the help of applied statistics, HR metrics and intuitive visualizations tools to explore the data and present some insights behind data on which the organization could reflect upon. In running these descriptive analyses, it's been mostly used 'Qlik', one of the top-rated data visualization and business intelligence vendors. In this web-based application, that also supports visual data discovery and self-service BI reporting, after uploading my set of data, I was given the chance to develop intelligent and high-intuitive dashboards through which derive insights on the referred workforce. Also, as I needed a tool that could possibly do both machine learning tasks as well as some data visualization with the same purpose of finding out significant discoveries in the HR picture, some graph and plot codes have been run in the programming language of 'Python' through the web-based application 'Jupyter Notebook', using command-line syntax. This notebook extends the console-based approach providing a thorough computing interface that includes developing, documenting and executing codes. Afterward, along with these first types of analyses, it's been delivered the predictive analysis part processing multiple machine learning models, to emphasize the difference with respect to the descriptive ones, that, as it will be shown over the next sections, can't possibly allow the understanding of the root causes behind some driving patterns on employee data. Using predictive techniques on carrying on my analyses, on the other hand, I tried to effectively find the best predictors of some employee variables, on which HR managers should focus to improve the respective HR policies and initiatives, to verify whether, presumably, the employee data are then effectively improved. This method, which implies rigorously seeking data patterns and trends to support realistic findings from the rough and disparate dataset that I simulated, was applied within the Jupyter Notebook free interface which I accessed by downloading and entering on 'Anaconda navigator'; Anaconda is in fact a "graphical user interface" (GUI), and represents an enormous data-science repositories, as it allows users to launch applications, simplifying Python and R packages and environments management and deployment. I chose to serve myself of this language rather than using R because Python functionalities are known to be more wide-ranging, winding up to be suitable both for descriptive statistical analysis and for machine learning applications. As a first step, I then imported all the packages needed to run the subsequent commands of data manipulation and data visualization. Particularly, ss we can see on Figure 1, these packages have been taken by the following main imported libraries:

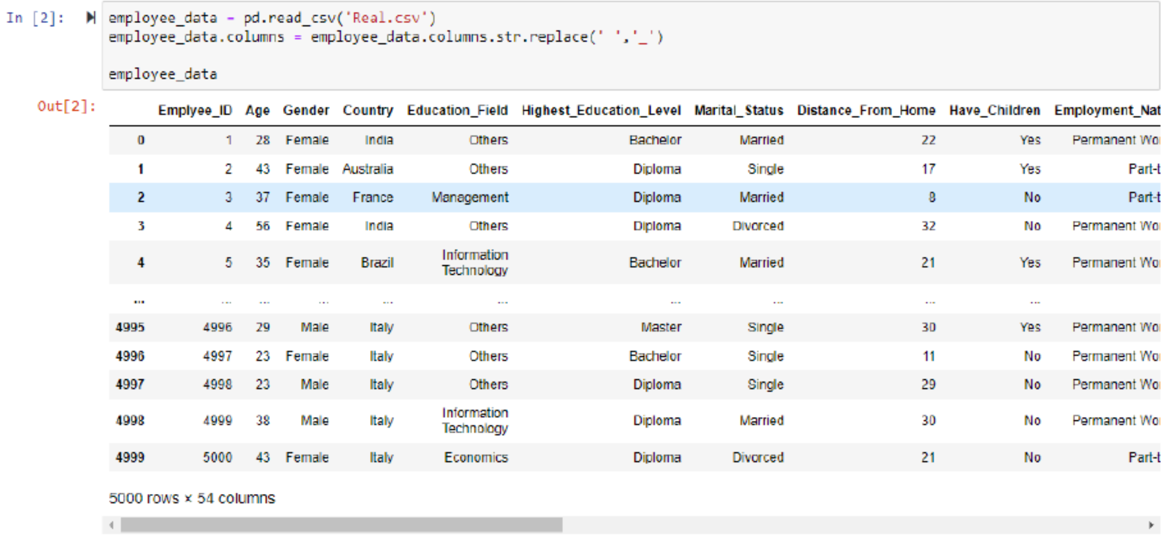
- ‘Pandas’, a powerful and easy-friendly open source data analysis and manipulation tool, built on top of the Python programming language
- 'NumPy', a Python library supporting large, multidimensional arrays and metrics and high-level mathematical functions
- ‘Statsmodels’, a library of Python programming language that supply classes and functions for the estimation of many different statistical models
- ‘Scikit-Learn’, a free software machine learning library for Python programming language that has also an open-source web page specialized in delivering efficient codes, knowledge and tools for predictive data analysis.
- ‘Seaborn’, hat is a Python data visualization library that equips with a high-level interface for drawing attractive and informative statistical graphs
- ‘Matplotlib’, an all-inclusive library for creating static, animated and interactive visualizations in Python



```
In [1]:  # Libraries and modules

         # Computing
         import pandas as pd
         import numpy as np
         import statsmodels.formula.api as smf
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import StandardScaler

         # Classifiers
         #import lightgbm as lgb

         # Visualizations
         import seaborn as sns
         import matplotlib.pyplot as plt
         import pprint

         # Settings
         sns.set_style("whitegrid")
         #sns.set_palette("pastel")
```

FIGURE 1: Essential libraries and modules imported to work my my dataset

Secondly, as it’s shown in Figure 2, I made sure my dataset was loaded in the RAM of my personal computer, using the reading function “pandas.read_csv”, so as for me to take the dataset into a Jupyter Notebook worksheet, making sure the system got the data in the same database form as appeared in Excel. Figure 3 presents instead the characteristics of each employee variable, explicating the name and the type of the variables and counting the total appearances of each observation belonging to each variable in the dataset. This helped me figure out if, once again, the generation of my dataset was functional to my expectations and purposes or if I had to change something to end up with different numbers for my employee observations.



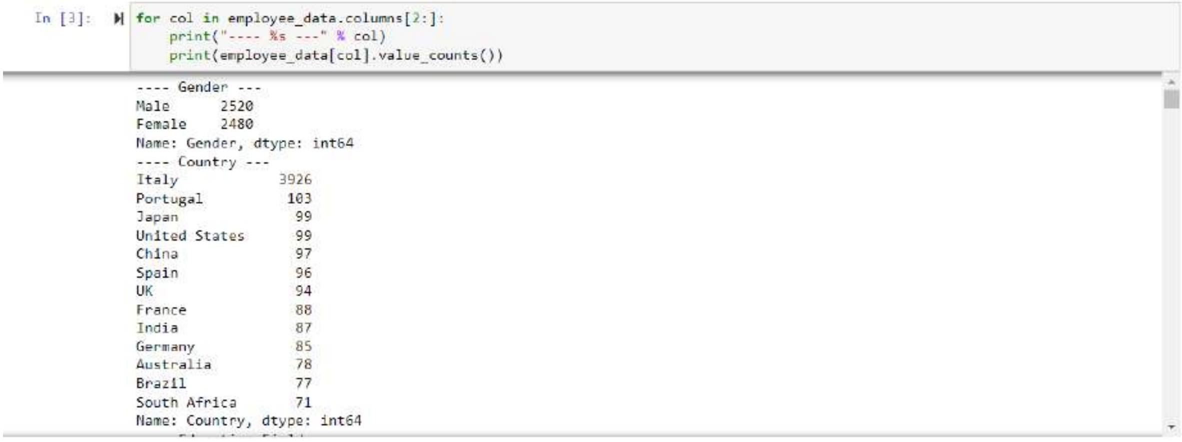
```
In [2]:  employee_data - pd.read_csv('Real.csv')
         employee_data.columns = employee_data.columns.str.replace(' ','_')

         employee_data
```

Out[2]:

	Emplyee_ID	Age	Gender	Country	Education_Field	Highest_Education_Level	Marital_Status	Distance_From_Home	Have_Children	Employment_Nat
0	1	28	Female	India	Others	Bachelor	Married	22	Yes	Permanent Wo
1	2	43	Female	Australia	Others	Diploma	Single	17	Yes	Part-t
2	3	37	Female	France	Management	Diploma	Married	8	No	Part-t
3	4	56	Female	India	Others	Diploma	Divorced	32	No	Permanent Wo
4	5	35	Female	Brazil	Information Technology	Bachelor	Married	21	Yes	Permanent Wo
...	...	...	...	...	...	...	...	...	...	
4995	4996	29	Male	Italy	Others	Master	Single	30	Yes	Permanent Wo
4996	4997	23	Female	Italy	Others	Bachelor	Single	11	No	Permanent Wo
4997	4998	23	Male	Italy	Others	Diploma	Single	29	No	Permanent Wo
4998	4999	38	Male	Italy	Information Technology	Diploma	Married	30	No	Permanent Wo
4999	5000	43	Female	Italy	Economics	Diploma	Divorced	21	No	Part-t

5000 rows × 54 columns

FIGURE 2: my employee simulated dataset read in Jupyter Notebook and imported from Excel. Other values couldn't be contained in this figure for space issues.



```
In [3]:  for col in employee_data.columns[2:]:
             print("---- %s ---" % col)
             print(employee_data[col].value_counts())
```

```
---- Gender ---
Male      2520
Female    2480
Name: Gender, dtype: int64
---- Country ---
Italy            3926
Portugal          103
Japan              99
United States      99
China              97
Spain              96
UK                 94
France             88
India              87
Germany            85
Australia          78
Brazil             77
South Africa       71
Name: Country, dtype: int64
```

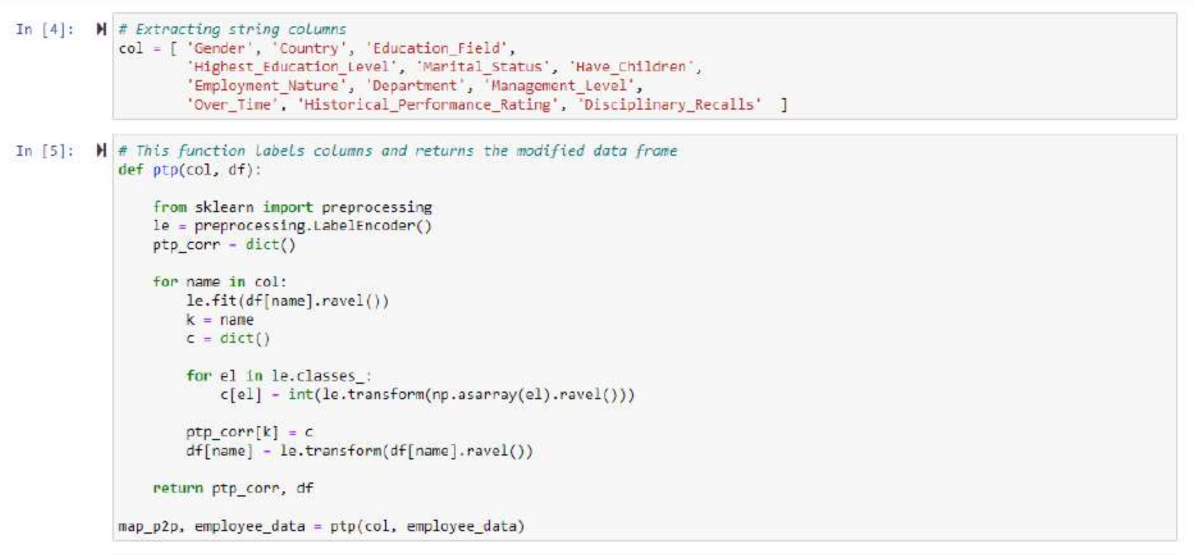
FIGURE 3: Dataset column description counting total number of observations. Other values couldn't be contained in this figure for space issues.

3.2.5 Data Pre-processing Actions

Before approaching the advanced level of predictive analyses, I had to undertake a great deal of endeavor to prepare the data for the analyses; in particular, it's been essential to make sure the data were suitable to be worked by the selected models, to possibly then extract meaningful knowledge and insights. This initial phase of data preparation concerns then the process, performed by data-scientists, of transforming the raw data at disposal into a model-readable format, and resolving other issues with noisy data that could often be made of incomplete, irregular or inconsistent values. This very action, whose goal is to make dataset features able to be smoothly trained by machine learning models, is called 'Feature Engineering' and consists of addressing the following crucial issues with respect to data:

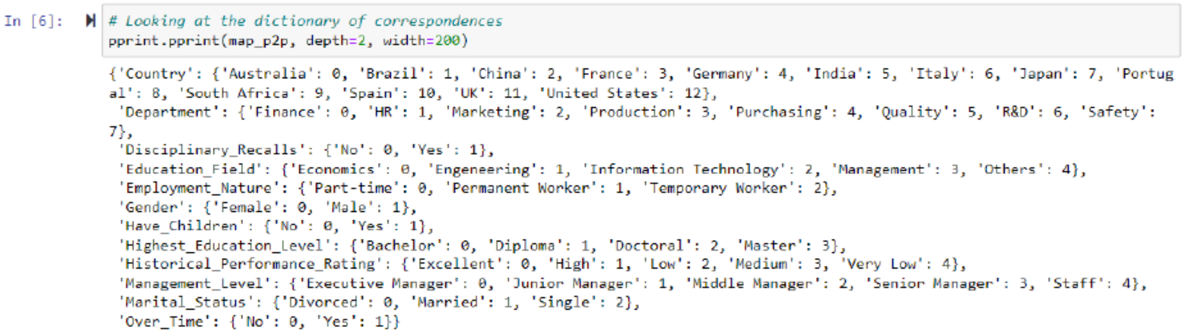
- Variable types, characteristics, and transformation
- Missing value imputation
- Outlier treatment
- Feature scaling process

Without this effort, data would always yield inaccurate and worthless outcomes derived from unreliable algorithms, and consequently to decisions of the same nature (Analytics India Magazine, 2021). Particularly, since in building my dataset simulation, I already performed some intrinsic actions to wind up with a rationale and sense-making set of employee data, then it has been not necessary to clean the data from scratch. For instance, as a matter of fact, in the final dataset, there weren't any null or missing values, as for the totality of observations, to replace with other numbers using specific missing data statistical imputation techniques or to tag with zero; moreover, I found only a couple of outliers, or unique values significantly different from the remaining data, that I purposefully decided to completely ignore instead of treating with special attention. Rather, in the first place, since I knew in advance that the machine-learning algorithms I was going to run worked only with numerical variables, either them being discrete or continuous, I had to first transform all the binary categorical features (whose possible values were ”Yes” or ”No”) into the values 1 and 0 respectively; using the Scikit-Learn package function 'Label_Encoder', used to encode labels replacing the categories with numerical representations, then, I applied the same operations even for the k-nary categorical (either of nominal or ordinal nature) features. This has lead to the conversion of all categorical values to numeric values in order to make the classification tasks more efficient. Label Encoder instrument, basically, allowed me to associate an ordinal number, starting from 0, for each different variable's tuple present within my dataset. Though the system stored my categorical labels as numbers, it's important to remind that these didn't become all of a sudden any meaningful numbers with which I could do arithmetical operations, like adding, subtracting, multiplying, dividing, as the converted numbers are just associated with the class label of reference. Figure 4, along with Figures 5 and 6, show how I performed this operation, eventually returning the dataset variables completely converted into integers values. As an instance, the categorical dimension "Highest Education Level" was manipulated, as, from containing four values such as "Diploma", "Bachelor", "Master" and "Doctoral", these latter have been converted to the value 0, 1, 2 and 3 respectively. Overall, it could be stated that the types of the variable have all been adjusted to integer values, as we could see from Figure 5 and Figure 6. As an alternative, perhaps more efficient but less suitable for my case of analysis, I could have also used the Scikit-Learn module 'One_hot_encoding', which, differently, creates, for each possible categories' value, two Boolean (so, binary) columns, meaning that, as a consequence, for each observation of the categorical converted variables, there will be only one specific column with value "Yes", while all the other created column will have value "No". This operation could prove to be more precise but, as a weak point, it greatly enlarges the final dataset, making the analyses far more complicated to be run by the system.



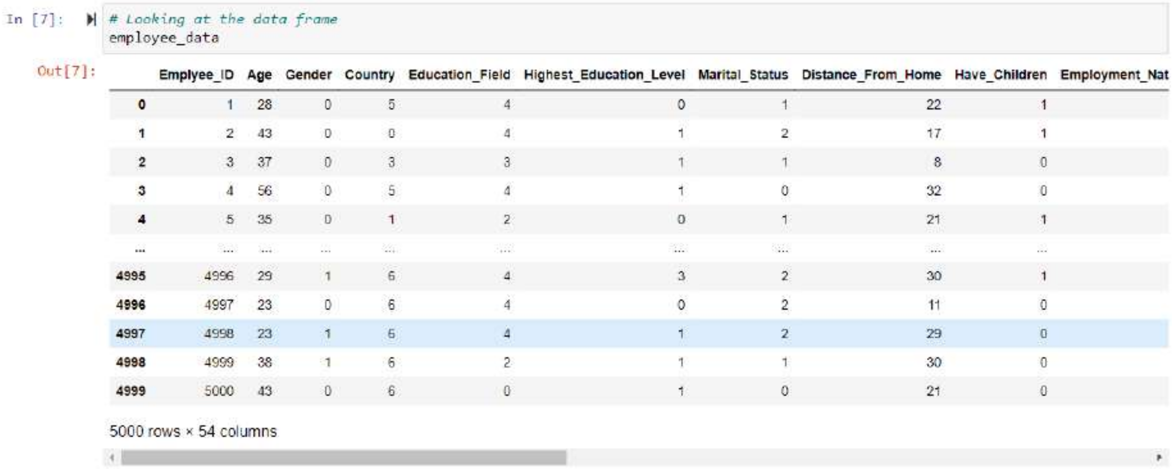
```
In [4]:  # Extracting string columns
         col = [ 'Gender', 'Country', 'Education_Field',
                'Highest_Education_Level', 'Marital_Status', 'Have_Children',
                'Employment_Nature', 'Department', 'Management_Level',
                'Over_Time', 'Historical_Performance_Rating', 'Disciplinary_Recalls'  ]

In [5]:  # This function labels columns and returns the modified data frame
         def ptp(col, df):

             from sklearn import preprocessing
             le = preprocessing.LabelEncoder()
             ptp_corr = dict()

             for name in col:
                 le.fit(df[name].ravel())
                 k = name
                 c = dict()

                 for el in le.classes_:
                     c[el] = int(le.transform(np.asarray(el).ravel()))

                 ptp_corr[k] = c
                 df[name] = le.transform(df[name].ravel())

             return ptp_corr, df

         map_p2p, employee_data = ptp(col, employee_data)
```

FIGURE 4: Label Encoder function to convert all categorical dimensions into numerical



```
In [6]:  # Looking at the dictionary of correspondences
         pprint.pprint(map_p2p, depth=2, width=200)

         {'Country': {'Australia': 0, 'Brazil': 1, 'China': 2, 'France': 3, 'Germany': 4, 'India': 5, 'Italy': 6, 'Japan': 7, 'Portug
         al': 8, 'South Africa': 9, 'Spain': 10, 'UK': 11, 'United States': 12},
          'Department': {'Finance': 0, 'HR': 1, 'Marketing': 2, 'Production': 3, 'Purchasing': 4, 'Quality': 5, 'R&D': 6, 'Safety':
         7},
          'Disciplinary_Recalls': {'No': 0, 'Yes': 1},
          'Education_Field': {'Economics': 0, 'Engeneering': 1, 'Information Technology': 2, 'Management': 3, 'Others': 4},
          'Employment_Nature': {'Part-time': 0, 'Permanent Worker': 1, 'Temporary Worker': 2},
          'Gender': {'Female': 0, 'Male': 1},
          'Have_Children': {'No': 0, 'Yes': 1},
          'Highest_Education_Level': {'Bachelor': 0, 'Diploma': 1, 'Doctoral': 2, 'Master': 3},
          'Historical_Performance_Rating': {'Excellent': 0, 'High': 1, 'Low': 2, 'Medium': 3, 'Very Low': 4},
          'Management_Level': {'Executive Manager': 0, 'Junior Manager': 1, 'Middle Manager': 2, 'Senior Manager': 3, 'Staff': 4},
          'Marital_Status': {'Divorced': 0, 'Married': 1, 'Single': 2},
          'Over_Time': {'No': 0, 'Yes': 1}}
```

FIGURE 5: Numerical values associated to each class label after the conversion of categorical dimensions.



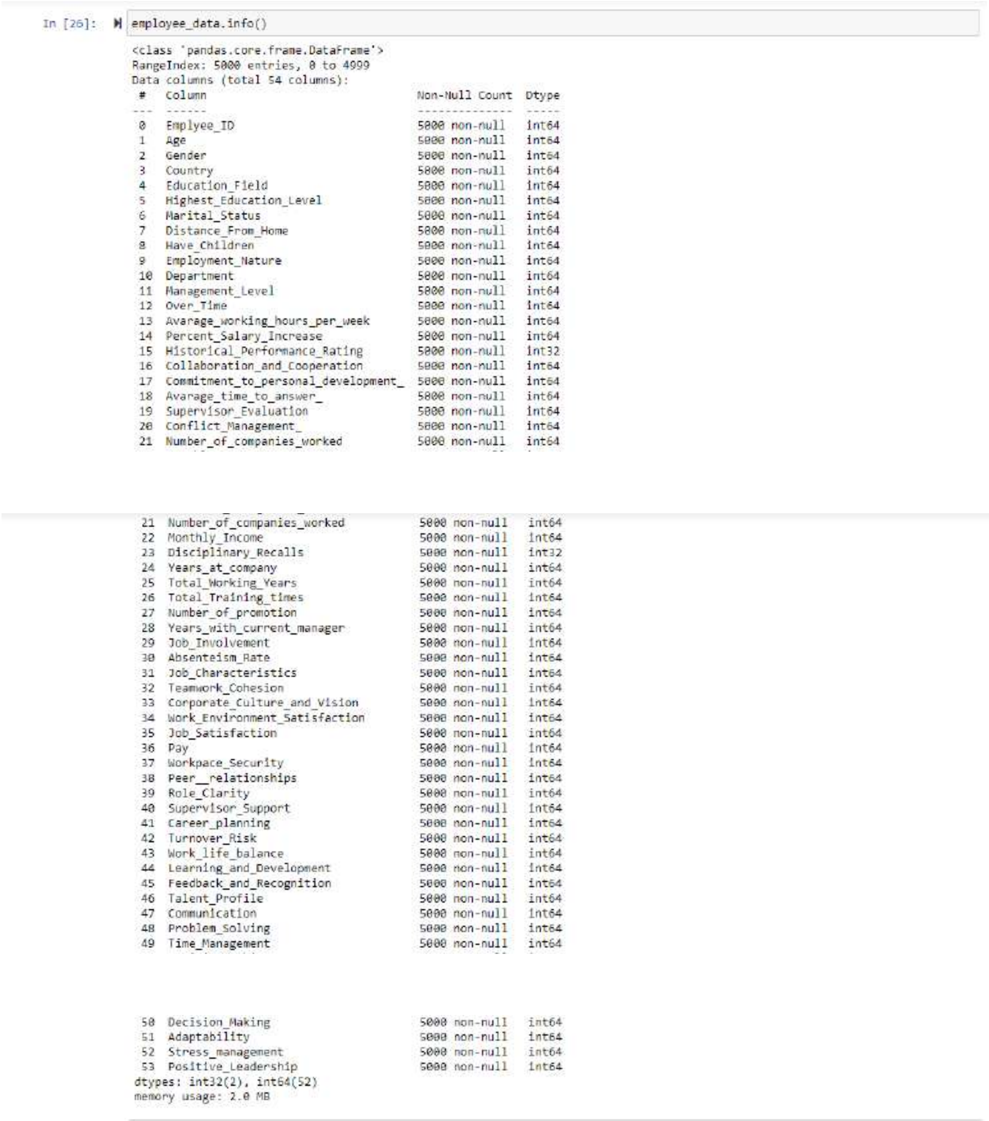
```
In [7]:  # Looking at the data frame
         employee_data
```

Out[7]:

	Emplyee_ID	Age	Gender	Country	Education_Field	Highest_Education_Level	Marital_Status	Distance_From_Home	Have_Children	Employment_Nat
0	1	28	0	5	4	0	1	22	1	
1	2	43	0	0	4	1	2	17	1	
2	3	37	0	3	3	1	1	8	0	
3	4	56	0	5	4	1	0	32	0	
4	5	35	0	1	2	0	1	21	1	
...	...	...	...	...	...	...	...	...	...	
4995	4996	29	1	6	4	3	2	30	1	
4996	4997	23	0	6	4	0	2	11	0	
4997	4998	23	1	6	4	1	2	29	0	
4998	4999	38	1	6	2	1	1	30	0	
4999	5000	43	0	6	0	1	0	21	0	

5000 rows × 54 columns

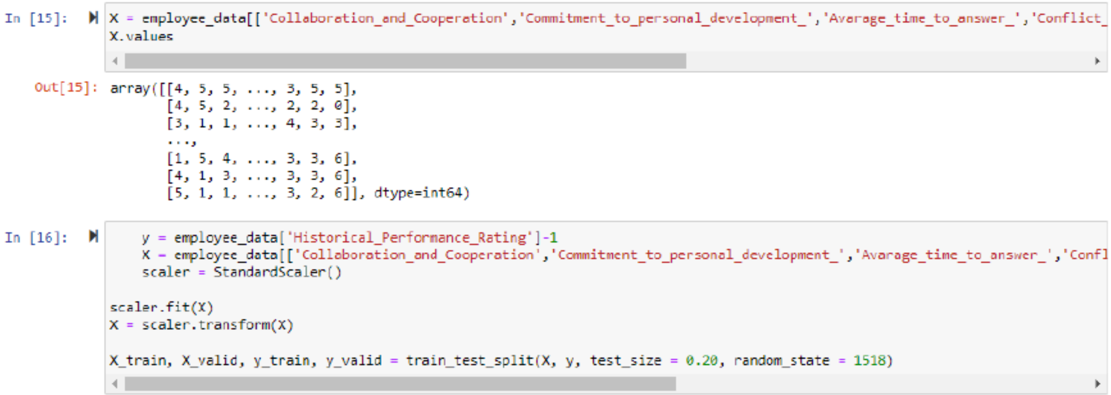
FIGURE 6: Dataset frame picture after the conversion of categorical dimensions. Other values couldn't be contained in this figure for space issues.



```
In [26]:  employee_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 54 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Emplyee_ID                         5000 non-null   int64
 1   Age                                5000 non-null   int64
 2   Gender                             5000 non-null   int64
 3   Country                            5000 non-null   int64
 4   Education_Field                    5000 non-null   int64
 5   Highest_Education_Level            5000 non-null   int64
 6   Marital_Status                     5000 non-null   int64
 7   Distance_From_Home                 5000 non-null   int64
 8   Have_Children                      5000 non-null   int64
 9   Employment_Nature                  5000 non-null   int64
 10  Department                         5000 non-null   int64
 11  Management_Level                   5000 non-null   int64
 12  Over_Time                          5000 non-null   int64
 13  Avarage_working_hours_per_week     5000 non-null   int64
 14  Percent_Salary_Increase            5000 non-null   int64
 15  Historical_Performance_Rating      5000 non-null   int32
 16  Collaboration_and_Cooperation      5000 non-null   int64
 17  Commitment_to_personal_development_ 5000 non-null  int64
 18  Avarage_time_to_answer_            5000 non-null   int64
 19  Supervisor_Evaluation              5000 non-null   int64
 20  Conflict_Management_               5000 non-null   int64
 21  Number_of_companies_worked         5000 non-null   int64

 21  Number_of_companies_worked         5000 non-null   int64
 22  Monthly_Income                     5000 non-null   int64
 23  Disciplinary_Recalls               5000 non-null   int32
 24  Years_at_company                   5000 non-null   int64
 25  Total_Working_Years                5000 non-null   int64
 26  Total_Training_times               5000 non-null   int64
 27  Number_of_promotion                5000 non-null   int64
 28  Years_with_current_manager         5000 non-null   int64
 29  Job_Involvement                    5000 non-null   int64
 30  Absenteism_Rate                    5000 non-null   int64
 31  Job_Characteristics                5000 non-null   int64
 32  Teamwork_Cohesion                  5000 non-null   int64
 33  Corporate_Culture_and_Vision       5000 non-null   int64
 34  Work_Environment_Satisfaction      5000 non-null   int64
 35  Job_Satisfaction                   5000 non-null   int64
 36  Pay                                5000 non-null   int64
 37  Workpace_Security                  5000 non-null   int64
 38  Peer__relationships                5000 non-null   int64
 39  Role_Clarity                       5000 non-null   int64
 40  Supervisor_Support                 5000 non-null   int64
 41  Career_planning                    5000 non-null   int64
 42  Turnover_Risk                      5000 non-null   int64
 43  Work_life_balance                  5000 non-null   int64
 44  Learning_and_Development           5000 non-null   int64
 45  Feedback_and_Recognition           5000 non-null   int64
 46  Talent_Profile                     5000 non-null   int64
 47  Communication                      5000 non-null   int64
 48  Problem_Solving                    5000 non-null   int64
 49  Time_Management                    5000 non-null   int64

 50  Decision_Making                    5000 non-null   int64
 51  Adaptability                       5000 non-null   int64
 52  Stress_management                  5000 non-null   int64
 53  Positive_Leadership                5000 non-null   int64
dtypes: int32(2), int64(52)
memory usage: 2.0 MB
```

FIGURE 7: List of variable descriptions after the conversion of categorical dimensions.

Another step included in the data preparation which I performed at the time of the division between the training and test dataset, which I left a bigger space afterward on this section, has been the 'Feature Scaling' process, which deals with the standardization of the dataset. In machine learning algorithms, some feature values often diverge from others as specific features with higher values, said to have a variance whose order of magnitude is way larger than others, dominating over other variables during the learning process of the algorithm. In this way, algorithms can't, therefore, learn from other features correctly as expected. Since this difference in magnitude does not mean that those features are actually more relevant in driving the final predictions of the model, which erroneously associate different importance weightage to each variable that it receives, dataset standardization to normal distribution represents a common requirement to radically improve the stability and accuracy of the predictive algorithms (Analytics India Magazine, 2021). Thanks to normalization, which eventually brings all the features in the same scale of values to appear similar, a higher quality data is obtained and then, overall, a higher value of insights could be driven. As shown in Figure 8, in my analyses, it was utilized the function 'StandardScaler', taken by Scikit-Learn libraries, that rescaled my variables giving them the properties of a standard normal distribution with a mean of zero and a standard deviation of one (Scikit-Learn: Machine Learning in Python, 2021).



```
In [15]:  X = employee_data[['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_answer_','Conflict_
          X.values

Out[15]:  array([[4, 5, 5, ..., 3, 5, 5],
                 [4, 5, 2, ..., 2, 2, 0],
                 [3, 1, 1, ..., 4, 3, 3],
                 ...,
                 [1, 5, 4, ..., 3, 3, 6],
                 [4, 1, 3, ..., 3, 3, 6],
                 [5, 1, 1, ..., 3, 2, 6]], dtype=int64)

In [16]:      y = employee_data['Historical_Performance_Rating']-1
              X = employee_data[['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_answer_','Confl
              scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

FIGURE 8: Application of Feature Scaling over my dataset utilizing StandardScaler command.

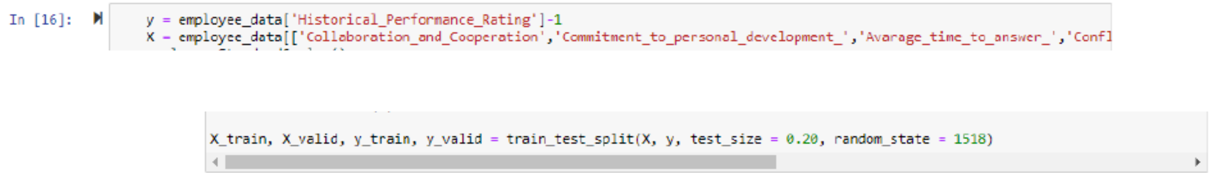
3.2.6 Design of Analytical Models

The second stage of the whole analytical case study at hand, subject to the employee dataset I have created, is the analysis stage, involving the use of developed statistical platform and programming language competencies to carry out both descriptive and predictive analyses, with the ultimate target of making effective human capital decisions. The first step to start thinking on how to structure the proposing descriptive reports, but, above all, to start designing most appropriately the predictive machine-learning models to fit the data, has been establishing which were the main question marks, related to some critical areas of HR (extensively described in section 3.1) that were meaningfully helpful to be answered through this analytical means of investigation. In fact, to find out the impact of certain employee attributes on some critical HR indicators like employee performance, some machine learning algorithms have been evaluated more suitable than others to my envisioned analyses for a number of reasons; this process was necessarily accompanied by a general study on the most efficient machine learning algorithms and on how those could be applied for predicting some important employee dimensions. I, therefore, took into consideration four principal problems that could be detected in an explorative way using descriptive analyses, and most of all could be solved using predictive analysis:

- turnover risk, that allows companies to discover some specific work-related factors for attrition and then predict the most likely future turnover employees in the organization to take action to prevent this to happen
- talent profile, that enables organizations to identify a set of targeting employees who represent the pool of talents on which the company should invest in the future, along with programming some critical retention initiatives
- performance appraisal, that let companies have an overall as well as a more detailed picture of how its diversified workforce is performing and how it will likely perform based on some influencing factors, to verify if motivating incentives for employees are appropriate and if the appraisal processes are totally unbiased

- job satisfaction and job involvement levels, whose predictions, based on the analysis of employee records, sentiments, and behaviors, could influence the ongoing HR policies to improve specific relevant indicators, like training needs or salary increases

Only on a second time, then, I could think of which analysis was best to proceed to run. For what concern the first descriptive part, I decided, in a limited manner, to not utilize all proper statistical tools like SPSS, Minitab, R, or Matlab, neither many statistical techniques like factor analysis,, chi-square analysis, and others; rather, with the idea of implementing a multiple-case study design for this work, I structured the analyses, serving myself of mostly Qlick, and sometimes Python, relying on the inference effectiveness of different types of graphs and plots, human capital KPIs, clustering techniques, and other tools; above all, advanced dashboards would accurately examine the employee variables in question and inform HR managers about the workforce situation pictured by my simulated dataset. On the other hand, regarding the second, more advanced side of predictive analyses, I decided to test a set of five main machine learning-based models, for each area of analysis, that returned various ranges of accuracy: Logistic Regression (LG), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and LGBM (Light Gradient Boosting Method) Classifier. The tasks implemented have necessarily been all of classification nature, simply because all the response variables I wanted to predict in my analyses were categorical class labels. These algorithms, conversely to regression models, are explicitly used to predict discrete, or binary, class label outputs. Importantly, literature also discloses that, while regression, association rules and clustering techniques are less widely applied, classification is the most frequently used predictive objective within HRA functions, as of its effective ability in forecasting classes of employee observations (M. Mani et All, 2021). That said, studying various algorithms, these mentioned classifiers were evaluated as the most suitable to develop predictive models for the employee attributes I build for this project, as well as the ones that were in power to provide the most accurate desired results for the prediction. Ultimately, I decided to report in this thesis only a couple of machine-learning models for each area of examination, the ones with the better resulting scores, even though the algorithms have all been applied, reaching a total of more than 20 predictions made (including the models where some hyperparameter tuning techniques have been processed). Adding complexity, in comparison to the descriptive part, before implementing each predictive model in Python, however, was not only necessary to carry out the time-consuming process of data preparation explained earlier on, but also, in the initial modeling phase, the entire amount of the final dataset observations has been split into a training set and a testing set. In fact, with supervised machine-learning techniques, the algorithms have to first learn how to make predictions for an employee attribute, meaning that have to be trained with a great portion of the total dataset as an input, that is equivalent to the so-called 'train set', that is derived taking a randomly 80% of the entire dataset (included the response variable). Here are then provided to the intelligent algorithm the class label of each training tuple so as for it to learn the patterns present on the data. Only after the algorithm has learned which are the patterns of data to make new predictions, could the algorithms be tested on the so-called 'test set', which correspond to the remaining 20% of left data, held out for a moment; importantly, from this second set of data, is excluded the response variable which, comprehend novel samples of the same dataset, the model is supposed to be able to predict after this dataset division. In fact, evaluating the performance of a predictive model on the same portion of data would be a methodological mistake: the machine would return a 100% score, returning the same labels of the samples that previously used in the learning phase, but it would certainly fail to predict new useful information if provided with other different data. The accuracy of a classifier must be therefore evaluated on the test set, containing novel observation instances not available at training time. This is why the training accuracy rate often is quite different (higher, theoretically) from the test accuracy score, gained by computing the percentage of the test set tuples that have been correctly classified by the algorithm (Scikit-Learn: Machine Learning in Python, 2021). This random splitting process is reported in Figure 9 using the 'train_test_split' helper function from Scikit-learn libraries.



```
In [16]:    y = employee_data['Historical_Performance_Rating']-1
            X = employee_data[['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_answer_','Confl
```

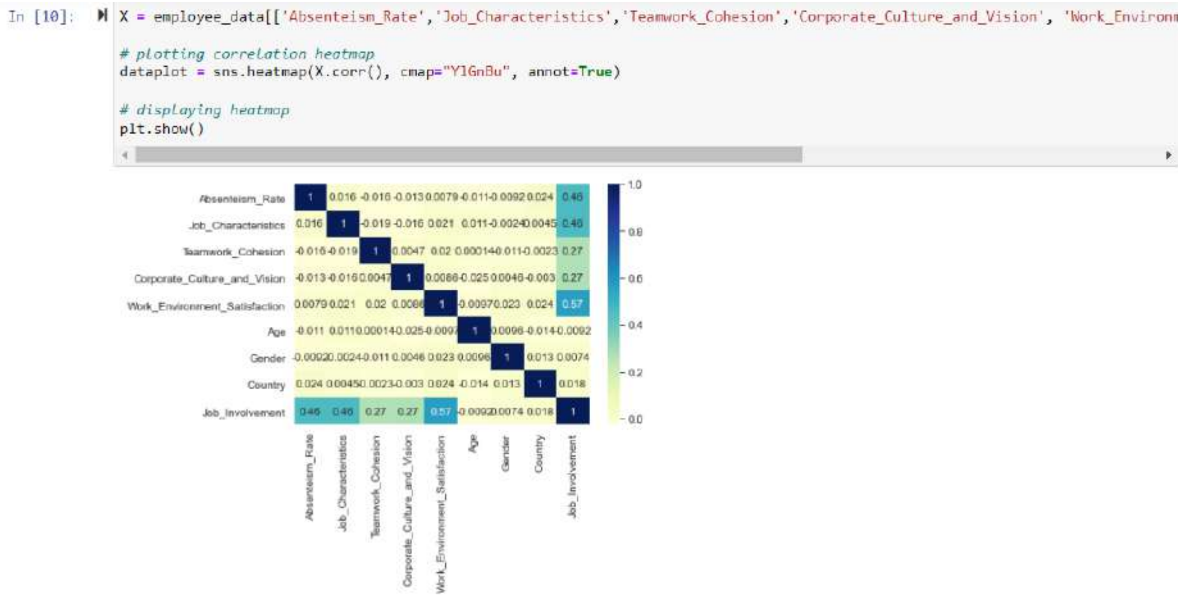
```
X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

FIGURE 9: Dataset partition phase into train and test set

3.2.7 Feature Selection and Model Evaluation

That being said, as a last step, another very helpful action and important procedure to undergo before running each algorithm, was to make sure to include a good set of input employee variables when developing the predictive models to forecast a specific response variable. This process, called 'Feature Selection', has the purposes, when developing a predictive model, of reducing the number of input attributes to both downsize the computational effort of the model, dropping the features least important for model building and selecting a subset of relevant features, as well as possibly improving the estimators' accuracy scores or boost their performance on very high-dimensional datasets. It is often the case that a number of irrelevant variables given as input for a model are in fact not associated with the dependent variable that is to be predicted. With this operation, by removing these variables, we can obtain a model that is trained with less complexity and that results to be more easily interpreted (Scikit-Learn: Machine Learning in Python, 2021). In addition, there are a multitude of methods by which it is possible to apply Feature Selection. In the analyses at hand, I decided to use a simple and little computationally expensive statistical method to rank, regardless of the classifiers utilized, the variables based on how much these are expected to be useful and sufficiently significant to the selected response feature for the classifications to come (Analytics India Magazine, 2021). I then performed a 'Correlation Analysis', or, better, 'Pearson's Correlation Coefficient' Analysis, that identifies the degree to which two continuous variables, one of that being the target variable, are linearly related, without making assumptions about which variables could be "predicted" by the other as happens in linear regression analysis. In addition, during my analyses, I have plotted some heat-maps to better visualize these computed analyses and so more clearly emphasize which features were the most related to the response variable. This process attempts to draw a line of best fit through the data of the two variables, returning values down from 1, meaning the two variables are "perfectly positively correlated", to -1, meaning "perfectly negatively correlated" and passing through 0, indicating between the features there is "no correlation at all". Figure 10 shows an example of correlation analysis applied for understanding the link between the dependent variable "Job

Involvement” and a set of independent variables for which, for the sake of the subsequent predictive analysis, I thought could be helpful to verify how strong and significant their link with the response variable was. In conclusion, these pre-processing findings allow the analyst to decide which variables should be given as input variables for the machine-learning model.

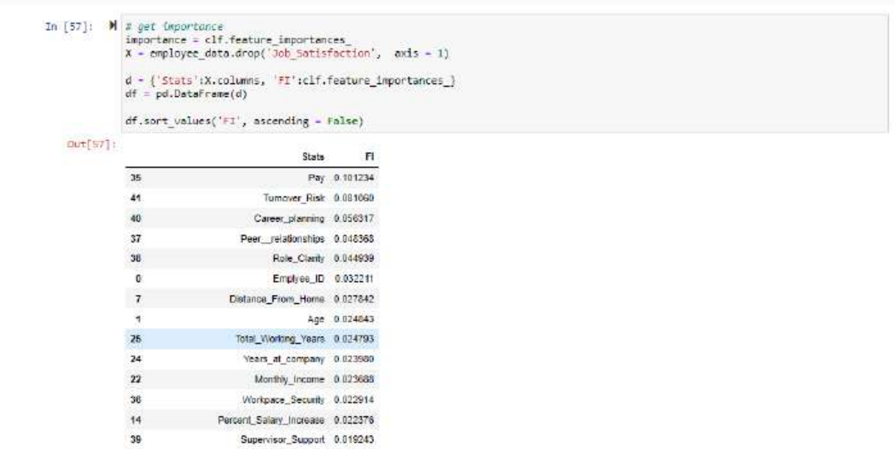


```
In [10]:  X = employee_data[['Absenteism_Rate','Job_Characteristics','Teamwork_Cohesion','Corporate_Culture_and_Vision', 'Work_Environm

          # plotting correlation heatmap
          dataplot = sns.heatmap(X.corr(), cmap="YlGnBu", annot=True)

          # displaying heatmap
          plt.show()
```

	Absenteism_Rate	Job_Characteristics	Teamwork_Cohesion	Corporate_Culture_and_Vision	Work_Environment_Satisfaction	Age	Gender	Country	Job_Involvement
Absenteism_Rate	1	0.016	-0.016	-0.013	0.0079	-0.011	-0.0092	0.024	0.46
Job_Characteristics	0.016	1	-0.019	-0.016	0.021	0.011	-0.0024	0.0045	0.46
Teamwork_Cohesion	-0.016	-0.019	1	0.0047	0.02	0.00014	-0.011	-0.0023	0.27
Corporate_Culture_and_Vision	-0.013	-0.016	0.0047	1	0.0086	-0.025	0.0046	-0.003	0.27
Work_Environment_Satisfaction	0.0079	0.021	0.02	0.0086	1	-0.0097	0.023	0.024	0.57
Age	-0.011	0.011	0.00014	-0.025	-0.0097	1	0.0096	-0.014	-0.0092
Gender	-0.0092	-0.0024	-0.011	0.0046	0.023	0.0096	1	0.013	0.0074
Country	0.024	0.0045	-0.0023	-0.003	0.024	-0.014	0.013	1	0.018
Job_Involvement	0.46	0.46	0.27	0.27	0.57	-0.0092	0.0074	0.018	1

FIGURE 10: Application of a Correlation Analysis, plotted with a heatmap, between the employee dependent dimension “Job Involvement” and a set of independent features, within the process of Feature Selection

Strictly correlated to Feature Selection, the process of ‘Feature Importance’ is pretty similar to the process just described since it is a technique, used with algorithms that support it, as tree-based classifiers, that consists of calculating an “importance” score for all the input features for a given machine learning model, but is performed after fitting the data to the model, in order to to make it easier to be interpreted by the analyst afterward. Overall, a higher score means that a certain feature will have a more significant impact on the predictive power of the model that is being run to forecast a specific attribute (T. Shin, 2021). In fact, when an expert notices a relevant relationship between two variables (one dependent and one independent), it doesn’t necessarily signify that changes in the independent variable drive changes in the dependent variable. Rather, it would just mean that there is a link between the two where changes in one are associated with changes in the other, without these associations being causally defined. As a practical instance, in an analysis with some specific data, the machine could learn that to a high number of ice cream sold, corresponds a high rate of skin cancer diagnoses, leading the analyst to think for a moment that eating ice cream regularly could be an important factor in causing skin cancer. Rather, rationally, these two variables are just casually and not causally interrelated: the algorithm does its job finding a significant pattern in the data, but this has always to be critically analyzed by the researcher. Therefore, this operation then helps to gain a better understanding of the data that goes into the model, giving the intelligent algorithm a human-readable meaning, with respect to the importance of the feature relationships, that is not possible to gain intrinsically for the model itself. Eventually, the prediction performance of the model is improved and the modeling process itself is speeded up, as it is reduced the computational time as well as the dimensionality of the model. Ultimately, Figure 11 highlights how the process of Feature importance, performed running the Scikit helper function "clf.feature_importance" on the prediction of the response variable "Job Satisfaction", has provided insightful information about the role played by specific employee variables in this Random Forest model. Additionally, in my analyses, for the LGB (Light Gradient Boosting) algorithms, a feature importance endeavor will be operated through the so-called 'Shapley Additive explanations': a famous approach from cooperative game theory which calculates the relative single impact of each variables to the prediction performed by the model (X. Wang, 2021).



```
In [57]:  # get importance
          importance = clf.feature_importances_
          X = employee_data.drop('Job_Satisfaction',  axis = 1)

          d = {'Stats':X.columns, 'FI':clf.feature_importances_}
          df = pd.DataFrame(d)

          df.sort_values('FI', ascending = False)
```

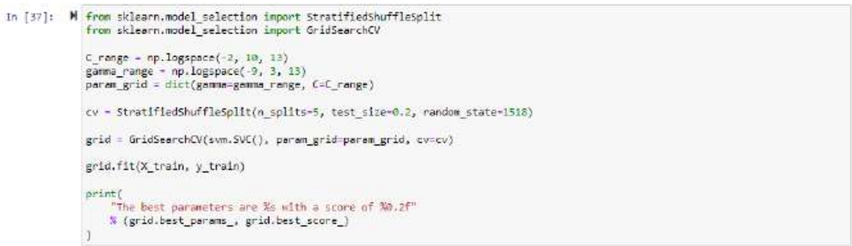
Out[57]:

	Stats	FI
35	Pay	0.101234
41	Turnover_Risk	0.081060
40	Career_planning	0.056317
37	Peer__relationships	0.048368
38	Role_Clarity	0.044939
0	Emplyee_ID	0.032211
7	Distance_From_Home	0.027842
1	Age	0.024843
25	Total_Working_Years	0.024793
24	Years_at_company	0.023980
22	Monthly_Income	0.023688
36	Workpace_Security	0.022914
14	Percent_Salary_Increase	0.022376
39	Supervisor_Support	0.019243

FIGURE 11: Random Forest's feature Importance for the response feature "Job Satisfaction"

Besides, overall, the scope of a learning predictive algorithm is to calculate a function that downsizes errors over a dataset (Analytics India Magazine, 2021). Nonetheless, when running a model, the phenomenon of 'Overfitting' refers to an algorithm that adapts too strongly to the data with which has been trained, memorizing perfectly the underlying patterns as well as the influence of noisy data present in the set of data; for this reason, its resulting analysis, corresponding too closely to the training set of data, most likely would fail to predict effectively new observations across different data inputs. Overfitting is especially frequent in three-like algorithms, as anticipated in the previous chapter. To avoid this to happen, analysts can no longer rely on the simple division between train and test sets, which leads to very divergent accuracy scores, but they need to control the flexibility of the model by adjusting the hyper-parameters - the different settings for the estimators of the model - that are specified outside the training phase (Analytics India Magazine, 2021). Taking an SVM example, through this process that finds the best hyper-parameter, called 'hyper-parameter tuning', users fit the model for each possible combination of typical parameters C, kernel and gamma, and ultimately pull out the model with the highest accuracy.

However, when evaluating different hyperparameters scenarios, some classification problems can still show a high imbalance in the distribution of the response classes (Scikit-Learn: Machine Learning in Python, 2021). To ensure there is no high variability in the relative class frequencies across train and test data sets, the train set can be further split to simulate an unseen test set on which the algorithm's hyperparameters can be tuned, corresponding to a train/validation/test set split in the proportion of 60/20/20. This way, training is still happening on the train set, but a provisional model evaluation is done on the newly formed 'validation set', and, when the analyses appear reliable and satisfying enough, the final evaluation is verified on the last test set (Scikit-Learn: Machine Learning in Python, 2021). Nonetheless, by dividing the total data into three sets, the number of observations that can be used by the algorithm in the learning phase is severely lowered, and the results might be biased by a specific choice based on the train or on the validation set. A solution to this problem is implementing a procedure called 'Cross-Validation' (CV) when selecting the best hyper-parameters. This process falls into the category of the "resampling method", which uses different portions of the data to repeatedly draw different samples from the total dataset, training the algorithm on different iterations in order to get additional information on how the model will generalize its prediction results to independent data inputs (Scikit-Learn: Machine Learning in Python, 2021). In my analyses, I applied cross-validation to evaluate the performance of my Random Forest classifiers for each area of analysis. More precisely, also, during Cross Validation has been possible to employ two basic approaches. In the 'StratifiedKFold CV', useful when the dataset is imbalanced, the training set is shuffled one time and then split into k smaller sets, and the performance score of the model corresponds to the average of the values computed in the iteration. In contrast, using 'StratifiedShuffleSplit', as I chose to do for my case of analysis, as shown in Figure 10, the data are shuffled each time before splitting n_splits times. Both methods, though, preserve the same percentage of samples for each target class, thus preserving the frequencies of the different classes, but in StratifiedKFold the test sets will always be different, whereas in StratifiedShuffleSplit these can overlap.

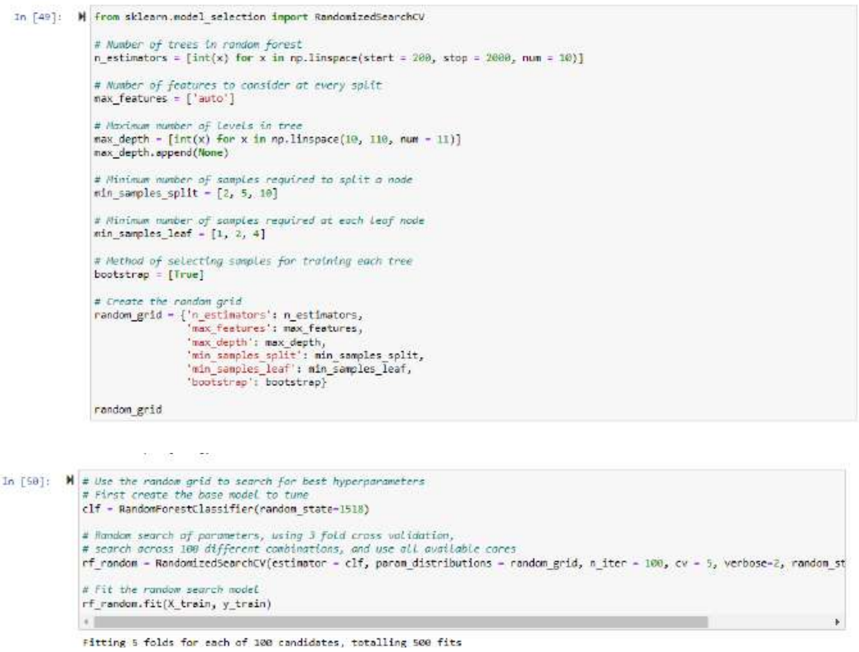


```
In [37]:  from sklearn.model_selection import StratifiedShuffleSplit
          from sklearn.model_selection import GridSearchCV

          C_range = np.logspace(-2, 10, 13)
          gamma_range = np.logspace(-9, 3, 13)
          param_grid = dict(gamma=gamma_range, C=C_range)

          cv = StratifiedShuffleSplit(n_splits=5, test_size=0.2, random_state=1518)

          grid = GridSearchCV(svm.SVC(), param_grid=param_grid, cv=cv)

          grid.fit(X_train, y_train)

          print(
              "The best parameters are %s with a score of %0.2f"
              % (grid.best_params_, grid.best_score_)
          )
```

FIGURE 12: Application of the function StratifiedShuffleSplit to resampling my entire dataset and implementation of a Grid Search for the SVM Classifier in objective

Overall, Cross-validation iterators can be used to directly carry out model selection using 'Grid Search(CV)' or 'RandomizedSearch(CV)', two generic methods to find out the optimal hyperparameter combinations of the model, that I respectively performed picking the respective helper function in Scikit-learn, to see if the accuracy scores of my Random Forest classifiers were indeed improved.

There is a difference between the two approaches: GridSearchCV, whose application is presented in Figure 12 for a SVM model, for given values, is more computationally expensive as it exhaustively considers all parameter combinations of a learning algorithm from a grid of parameter values specified with the "param_grid" parameter. Conversely, RandomizedSearchCV, reported in Figure 13 and utilized for a Random Forest model, implements a randomized search over just a fixed number of possible parameter values. In this second case, it’s the user choosing the number of combinations to examine specifying the “n_iter” parameter (Scikit-Learn: Machine Learning in Python, 2021).

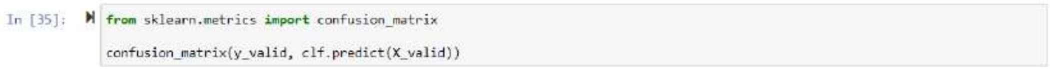


```
In [49]:  from sklearn.model_selection import RandomizedSearchCV

          # Number of trees in random forest
          n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]

          # Number of features to consider at every split
          max_features = ['auto']

          # Maximum number of levels in tree
          max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
          max_depth.append(None)

          # Minimum number of samples required to split a node
          min_samples_split = [2, 5, 10]

          # Minimum number of samples required at each leaf node
          min_samples_leaf = [1, 2, 4]

          # Method of selecting samples for training each tree
          bootstrap = [True]

          # Create the random grid
          random_grid = {'n_estimators': n_estimators,
                         'max_features': max_features,
                         'max_depth': max_depth,
                         'min_samples_split': min_samples_split,
                         'min_samples_leaf': min_samples_leaf,
                         'bootstrap': bootstrap}

          random_grid
```

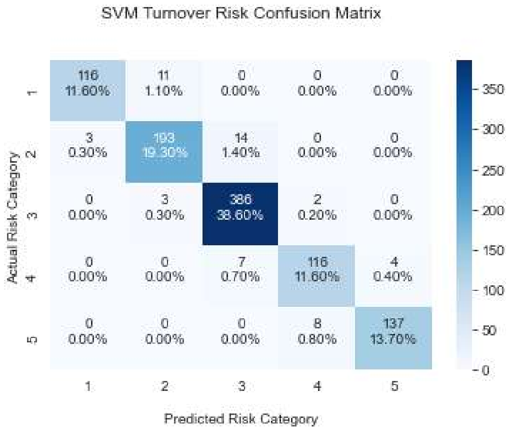
```
In [50]:  # Use the random grid to search for best hyperparameters
          # First create the base model to tune
          clf = RandomForestClassifier(random_state=1518)

          # Random search of parameters, using 3 fold cross validation,
          # search across 100 different combinations, and use all available cores
          rf_random = RandomizedSearchCV(estimator = clf, param_distributions = random_grid, n_iter = 100, cv = 5, verbose=2, random_st

          # Fit the random search model
          rf_random.fit(X_train, y_train)

          Fitting 5 folds for each of 100 candidates, totalling 500 fits
```

FIGURE 13: Application of a Randomized Search for the Random Forest Classifier in objective

Ultimately, to more deeply verify how models were performing in my experimental predictions, other than evaluating the accuracy score, defined as the fraction of correctly classified instances, for some algorithms I used the so-called ‘Confusion Matrix’, imported as a function from the Scikit-Lean libraries. Through this command, the precision was calculated, in particular, depicting the number of input features recorded that resulted as correct or incorrect in the prediction. The matrix returned was a N × N table, considering N as the total number of classes to be possibly predicted for the response feature, where, in one axis appeared the number of tuples predicted by the model for a target variable, and the other axis presented, near, the actual number of labels, putting these two values in comparison (Analytics India Magazine, 2021). Figure 14 illustrates a confusion matrix created after running the prediction, using an SVM classifier, of the employee feature “Turnover Risk”: in the first row of the matrix, for example, the numbers deployed tell the analyst that, of a total of 55 observations that corresponded to the first label of the variable, 116 have been successfully forecasted for that label, while 11 have been associated to the second label in the prediction.



```
In [35]:  from sklearn.metrics import confusion_matrix

          confusion_matrix(y_valid, clf.predict(X_valid))
```



SVM Turnover Risk Confusion Matrix

Actual Risk Category \ Predicted Risk Category	1	2	3	4	5
1	116 11.60%	11 1.10%	0 0.00%	0 0.00%	0 0.00%
2	3 0.30%	193 19.30%	14 1.40%	0 0.00%	0 0.00%
3	0 0.00%	3 0.30%	386 38.60%	2 0.20%	0 0.00%
4	0 0.00%	0 0.00%	7 0.70%	116 11.60%	4 0.40%
5	0 0.00%	0 0.00%	0 0.00%	8 0.80%	137 13.70%

FIGURE 14: Application of a Confusion Matrix to evaluate the prediction precision of a SVM classifier

3.2.8 Decision Making Stage

After developing the prediction models by fitting them with my simulated dataset, I ran the analyses for each of the four cases of analysis identified and mentioned before: Job Satisfaction&Job Involvement, Turnover Risk, Talent Profile and Performance Appraisal. Ultimately, I gathered all the results of my machine-learning analyses, sorted by field of analysis and for each algorithm utilized, creating the following accuracy score Table 3 that shows the various accuracy of the model based on prediction results. Specifically, throughout these analyses, the function utilized for interpreting and comparing the models’ outcomes has been the “accuracy_score” from Scikit-Learn library.

No.	Classifier algorithms	Accuracy on the test dataset (%)				
		Job Satisfaction	Job Involvement	Turnover Risk	Performance Appraisal	Talent Profile
1	Logistic Regression*	87,5	87,7	90,9	40,5	90,1
2	Support Vector Machine	100	100	100	87,1	81
3	Decision Tree	85,6	93,1	91,5	71,9	78,6
4	Random Forest	89,7	93,4	86,9	88,7	83,1
5	Light Gradient Boosting Method	92,5	95,9	88	84,3	87,1

* for this model has not been necessary to apply the splitting division in training and test datasets; moreover, it has been included in this final set of results only the accuracy score of one of the two Logistic Regression models performed, specifically the one that got the highest score

TABLE 3: Algorithms' Accuracy score

Once completed the process of analysis, post-analysis represents the next and last stage in the overall analysis framework, where decisions are taken based on the analysis results. Here the procedure implies trying to interpret the meaningful data insights to draw necessary conclusions and then take actions to change, or better improve, the existing status quo of the directly involved HR policies and procedures which could sort the particular problem that has been critically observed, with the ultimate goal of providing organizational benefits. Overall, insightful results of the analyses become only at this point actionable by the top managerial level of decision-making.

3.3 Experimental analyses

In this section, I show the results of my analyses stage, and also discuss the insights regarding the findings. As mentioned in the last section, each case of analysis will be divided into two parts. The first is the descriptive part mostly containing the HR dashboards I created using “Qlik” to explain some specific trends on my simulated dataset, along with a few insightful clustering techniques; the second part, of predictive nature, includes instead some machine learning models performed in “Python” that provide some accurate employee feature predictions on which based critical managerial reflections on how possibly intervening to appropriately change the future direction of some employees’ trends.

3.3.1 Job Satisfaction and Job Involvement Analysis

Job Satisfaction and Job Involvement, intentionally analyzed together to gain a more comprehensive understanding of the wider employee engagement trend, represent the first critical case of analysis in this experimental work. The following dashboard, in figure 15, highlights two overview indicators on my employee simulated dataset, discriminated by gender. For the hypothetical company, it shouldn’t be a surprise to discover that the “Average Involvement” score, measured on a scale of 1 to 5 in the engagement surveys collected, posits itself at level 3 for both males and females. Going into deeper details, the bar plot stresses out some similar figures regarding the distribution of the level of engagement as for both the gender types within the company. Also, completing the panoramic view and likely weighing on the Job Involvement results, the “Average Training Times” guaranteed for each employee inside the company, and the “Average Number of Promotions” seem to reach reasonable scores. Similarly, figure 16 depicts a bar chart, with a different appearance style, since it has been elaborated in Python rather than in Qlik, explicating the level of Job Satisfaction inside the organizations: the scoring percentages are what we were expecting after commenting the first Job Involvement visualization findings, meaning that they represent a regular distribution of employees job satisfaction ratings, underlying a slightly better trend for the male part of a workforce which turned out to have a slightly higher percentage of workers with a top score of 5 and, at the same time, a slightly lower percentage with the minimum score of 1.

Overall, this partiality given by these graphs still provides little insight on major problems regarding the workforce, which haven’t been grasped yet.

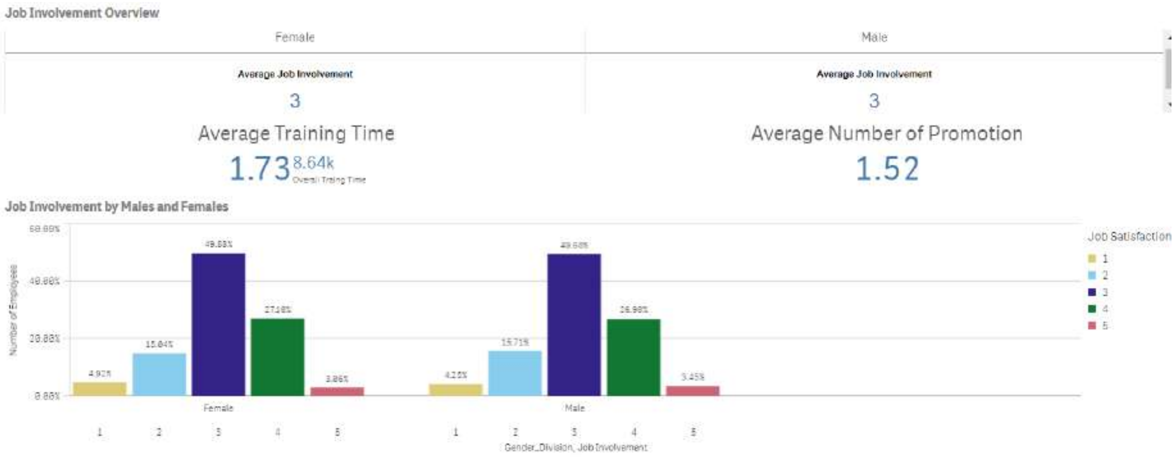


FIGURE 15: Job Involvement Overview KPI and Gender division bar chart

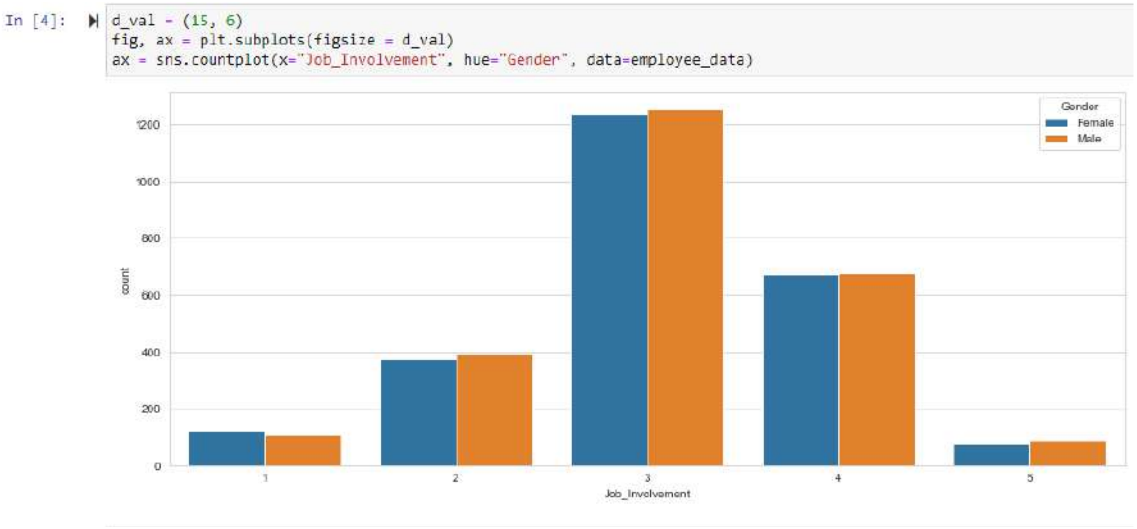

```
In [4]:  d_val = (15, 6)
         fig, ax = plt.subplots(figsize = d_val)
         ax = sns.countplot(x="Job_Involvement", hue="Gender", data=employee_data)
```

FIGURE 16: Job Satisfaction Gender division bar chart

On the graph deployed in figure 17, the interpretation of the dashboarding activity gets a bit more complicated as it's added a second feature, corresponding to the Country of origin of each employee, under which the level of Job Involvement across my HR population is examined performing a type of diversity analysis. What is easily noticeable, first of all, is that for the State corresponding to the company's headquarters, Italy, whose points in the graph were thus elaborated according to a number of observations of a completely different magnitude in comparison with other foreign countries, logically having far less people employed in this pretended Italian firm, there is almost no difference among the job involvement figures with respect to gender. However, involvement average difference, as for gender juxtaposition, gets relatively larger, still remaining within some decimal points of the range for all the other countries of provenience. Australian employees appear the least involved in their job: the company should then dig into what has brought them to register such a low average level of this important construct.

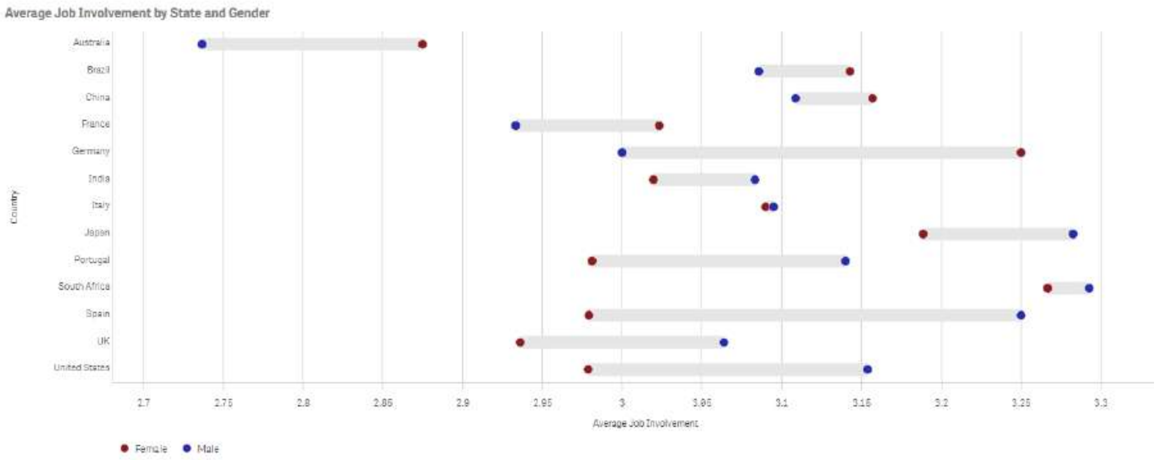


FIGURE 17: Average Job Involvement divided by Country and Gender distribution plot

Insisting on finding some insightful points of reflection concerning the Job Involvement analysis, the following dashboard in figure 18 presents interesting outcomes. Representing a metrics called ‘Average Time to Stay’, that should give an idea of how long long employees stay in a company based on some factors, two different trends are highlighted according to the gender type of the general workforce. In fact, the female staff makes the impression of staying an average longer time, expressed in years in the graph, in comparison to the male part, for each level of job involvement recorded. Moreover, the two lines, describing the behavior of the two variables “Years at Company'' and “Job Involvement”, initially seemed to be directly proportional in all their values but for the observation “4” and “5” of Job involvement, where the two lines don’t grow higher. It can be argued then that for the first three smaller levels of job involvement, both females and males tend to leave the company earlier. This gained awareness may force the firm to adjust something to keep job involvement as high as possible. On the other hand, on the scatter plot illustrated in the same dashboard, the seven different departments are plotted according to the average monthly income level in the y axis and the level of job involvement in the x-axis; also here, at a first glimpse, apparently subsist a direct positive correlation between the two variables on the axes, except for the "Purchasing" department, whose workers are found to be more involved in their jobs even if averagely have a lower salary. This analytical illustration should send a specific message to the top HR managers: that has emerged a problem with respect to the R&D departments, whose personnel, even if holding heavy responsibilities in their delicate tasks, are, like purchasing workers, averagely underpaid than other departments workers, and, not only for this reason, probably, are feeling little involved in what they do in their daily tasks. Additionally, these trends, if investigated further, could most likely be affecting other relevant KPIs of members of these two departments, like Performance and Turnover Risk.

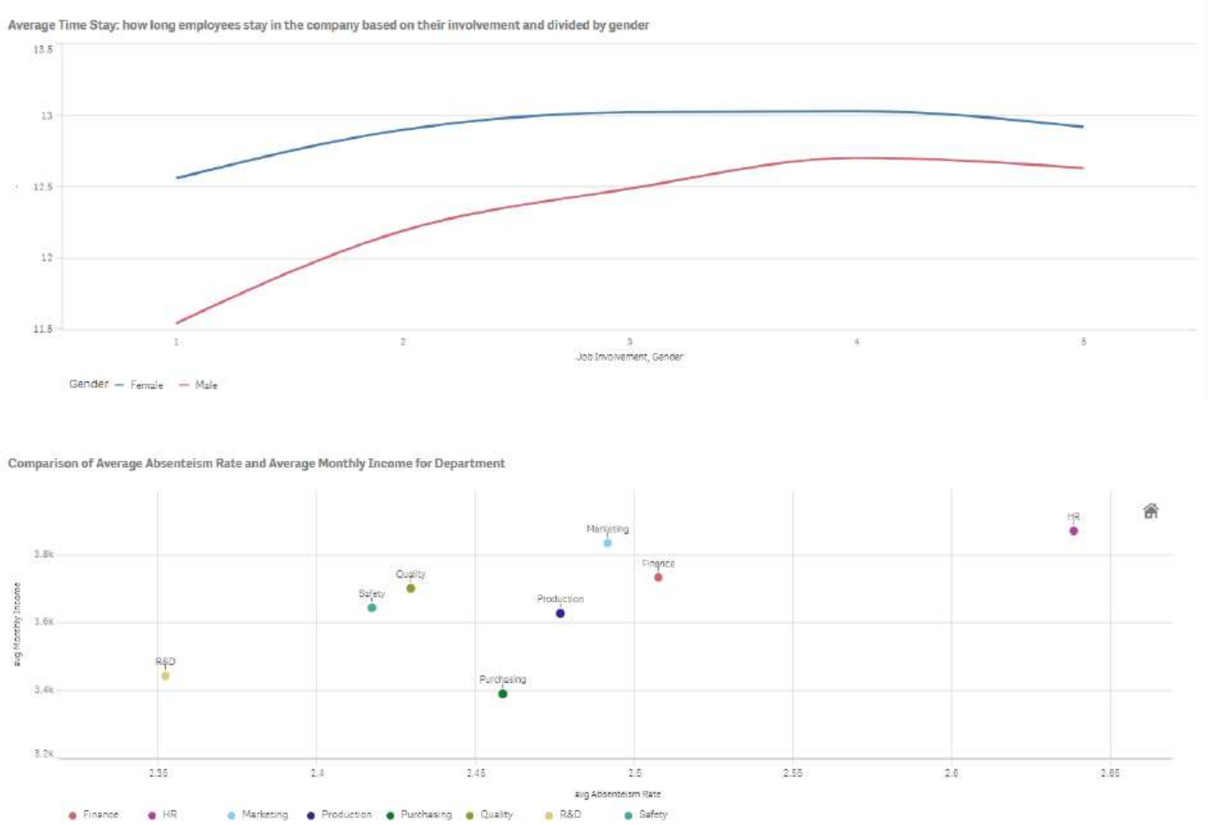
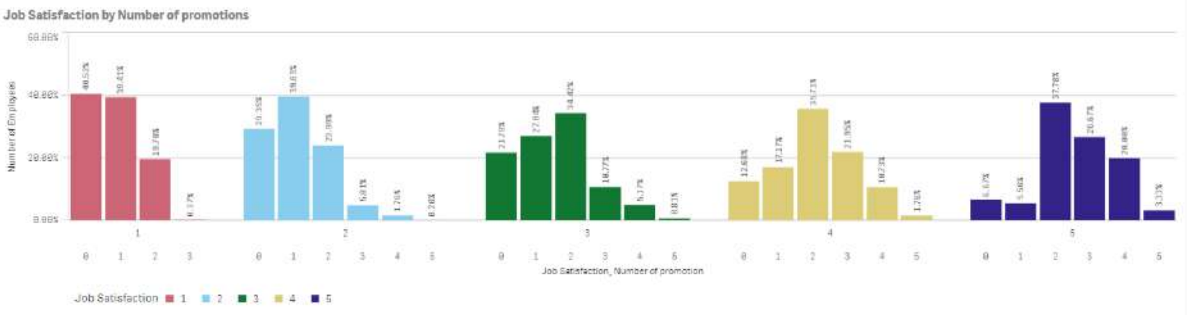
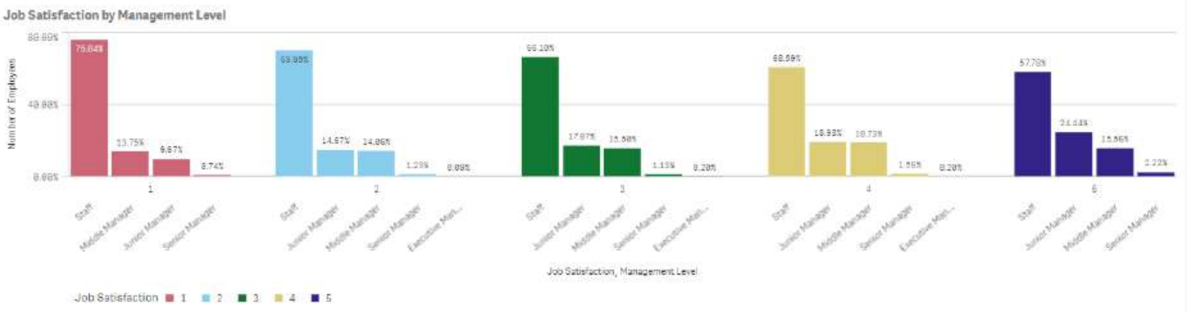


FIGURE 18: Average Time Stay line chart and Average Absenteeism Rate scatter plot

Carrying on the descriptive part, on the next dashboards in figure 19, are illustrated three bar charts depicting the percentages of promotions made and training times delivered to employees based on the respective score of "Job Satisfaction". The graph in the middle, instead, does the same analysis taking into consideration the managerial position of the distributed workforce. As expected, from the first bar chart stands out how, as the level of satisfaction raise, the number of promotions accomplished by the employees is more likely to be higher, passing from most of the observed employees that had 0 or 1 promotion for the minimum level of satisfaction, to a majority of promoted employees two or three times with a satisfaction level attested to 4 or 5. On the second graph, interestingly, the number of workers in the "staff" role diminishes progressively as the level of satisfaction gets higher, meaning that in those advanced levels, there is left room for other employees in higher positions to gain in numbers. Ultimately, on the third graph of the dashboard, a similar pattern is observed analyzing how vary the employee proportions, across the satisfaction scale, that appeared to had no training at all since they have arrived at the company: not surprisingly, over the highest level of satisfaction, employees less fortunate which hadn't any training courses or other development activities, are not even included on the first four training figures that emerge on the graph.





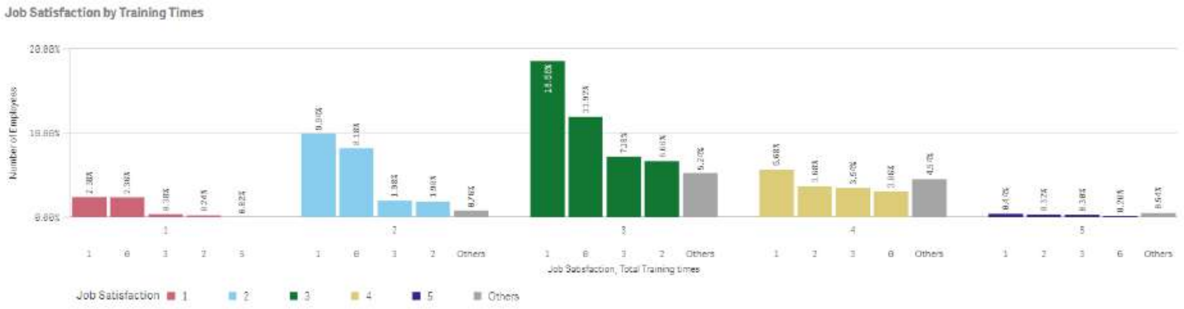


FIGURE 19: Job satisfaction by Promotion, Management Level and Training bar chart

In this last dashboard, represented in figure 20, through a mekko chart, managerial positions of employees and satisfaction levels are plotted to form some frames, as large as the sum of the monthly income associated with the employees of reference. We can notice how, as for each role managerial category, apart from the middle managers, the groups of workers totally underpaid are always the employees recording a satisfaction level of 5, the highest. Therefore, the managers should reflect that simply increasing employee income could not be the most effective ways to improve the level of involvement of employees. This graph in fact is an intuitive visualization from which grasping information about clusters of employees according to some characteristics.

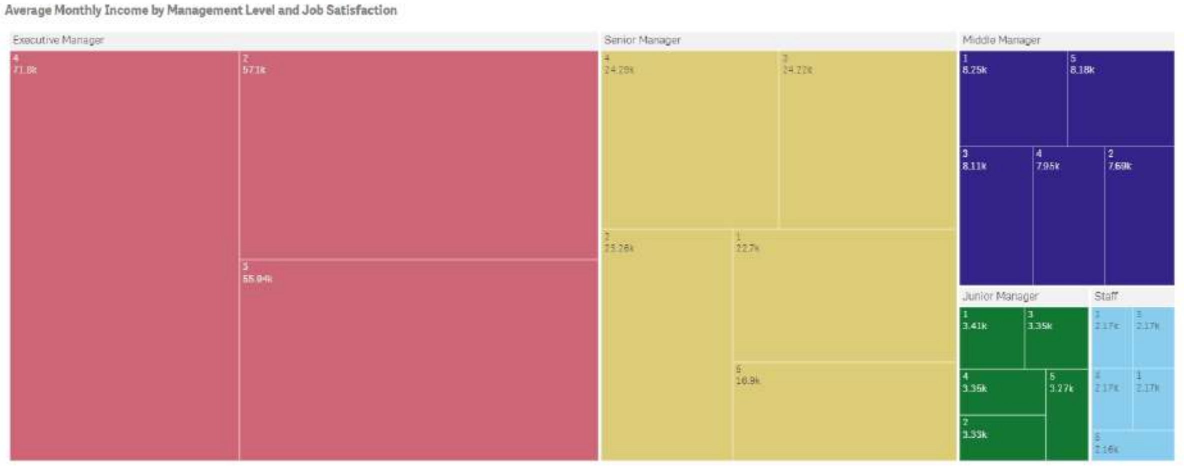
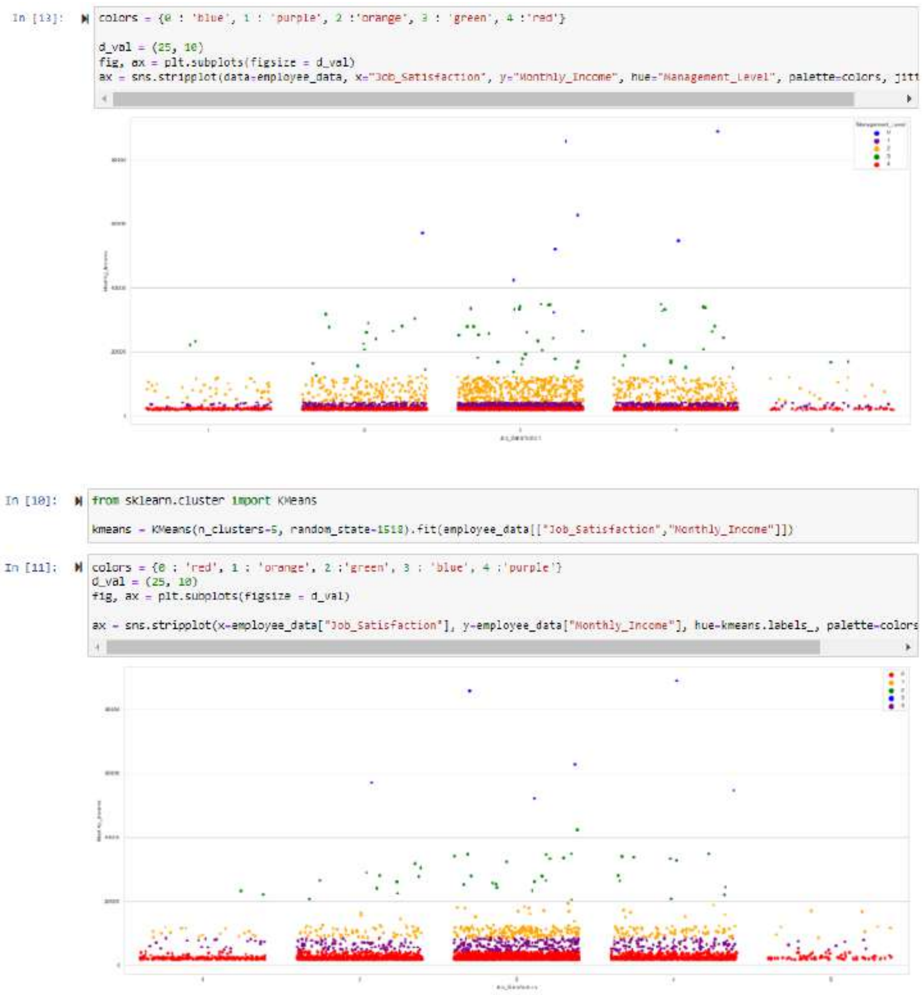


FIGURE 20: Average monthly income by management level and job satisfaction mekko chart

Lastly, two real practical examples of cluster analysis and clustering technique were applied in figure 21, using the Jupyter Notebook operating and graphical interface, which will be extensively utilized for the predictive analyses further on. First, a simple cluster analysis depicts a snap of the actual state-of-thing situation of the role management workforce's levels according to the continuous variable "Monthly Income", plotted on the y axis against "Job involvement" plotted on the x-axis. What is simply telling to the HR analysts is how managerial positions are actually distributed among employees with reference to the involvement score and income of each specific employee. Just below, in comparison, an unsupervised K-means algorithm has been fitted with the same two employee feature; this intelligent algorithm accurately succeeded in separating employees' samples into five groups of similar but not identical variance, as was explicitly requested in the commanding instructions, to see how the five managerial levels should be distributed in the most proper way among the labor force, represented in points in the graph. The algorithm, recognizing the relationship between job involvement rates and the salary levels, given as inputs, returns a division into five clusters of employees embedding a stratified behavior, signaling to HR decision-makers which employees should belong to which cluster. This comparison proves to be pretty useful as some anomalies can be easily detected. Potential biases committed could be in fact verified by observing, say, that an employee that is a purple point in the second clustering chart, reporting the least level of involvement and a moderately low salary, in the other cluster techniques actually covers a "Middle Manager" role. According to the intelligent K-means algorithm, this selected employee should rather be declassed of one managerial position; by the way, mostly for the intermediate two levels corresponding to the purple and yellow categories of "junior" and "middle" roles, managerial position distribution among the workforce should be completely re-organized. This could provide helpful insights for managers to change the overall workforce planning and hierarchical structure according to the job involvement trend present across the workforce.



```
In [13]:  colors = {0 : 'blue', 1 : 'purple', 2 :'orange', 3 : 'green', 4 :'red'}

          d_val = (25, 10)
          fig, ax = plt.subplots(figsize = d_val)
          ax = sns.stripplot(data=employee_data, x="Job_Satisfaction", y="Monthly_Income", hue="Management_Level", palette=colors, jitt
```

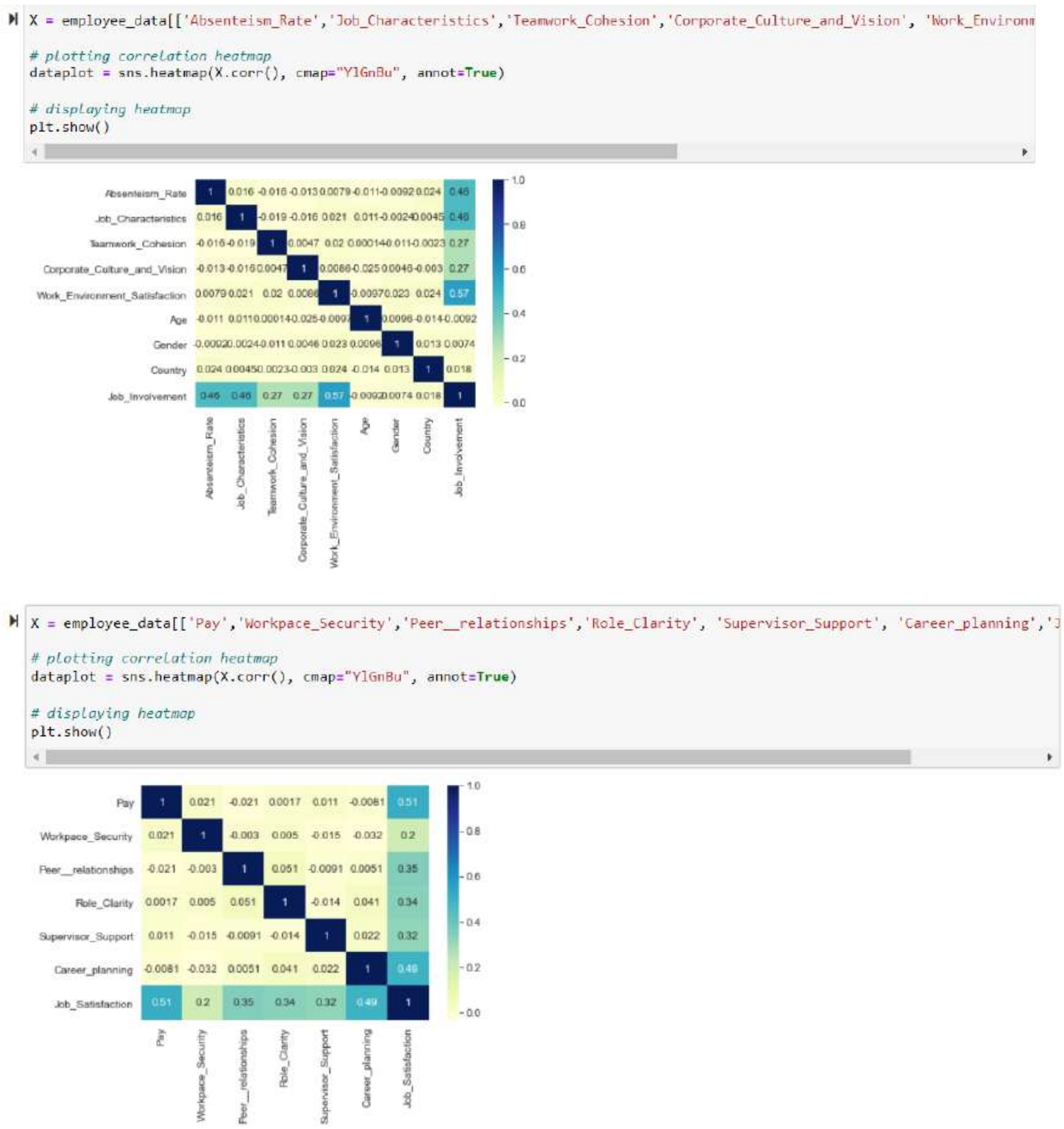
```
In [10]:  from sklearn.cluster import KMeans

          kmeans = KMeans(n_clusters=5, random_state=1518).fit(employee_data[["Job_Satisfaction","Monthly_Income"]])
```

```
In [11]:  colors = {0 : 'red', 1 : 'orange', 2 :'green', 3 : 'blue', 4 :'purple'}
          d_val = (25, 10)
          fig, ax = plt.subplots(figsize = d_val)

          ax = sns.stripplot(x=employee_data["Job_Satisfaction"], y=employee_data["Monthly_Income"], hue=kmeans.labels_, palette=colors
```

FIGURE 21: Real-world situation cluster analysis versus K-Means clustering classification for "Management Level" based on variables "Monthly Income" and "Job Satisfaction"

Moving to the predictive analyses, in this section, the goal is to predict satisfaction and engagement level (1-5) of the employees based on other variables which contribute to delineate it. In general, the main framework of the predictive analysis part is proposed in two steps. Firstly, a correlation analysis is used to run the process of "Feature Selection", reducing dimensions, because, as seen in the last section of the methodology of analysis, it is fundamental to prepare the data to make it readable and usable for machine learning algorithms to use. Secondly, the selected features are used to fit the prediction models. In figure 22, then, I tried to understand which were the most important features that should probably weigh more on the future classification of the two dependent variables, "Job Involvement" and "Job Satisfaction", that will be predicted further on by the models. To do so, it was used the statistical techniques called

“Correlation Analysis”, intuitively represented through a heat map. The results seem to be clear: concerning the first analysis, “Absenteeism rate” (0,47), “Work Environment Satisfaction” (0,47) and “Job Characteristics” (0,57) are the feature strongest linearly correlated, positively, to Job Involvement, with “Teamwork Cohesion” and “Corporate Culture and Vision” still indicating a significant pattern of correlation but less strong (both 0,27). These outcomes are perfectly coherent with how the data simulation has been built. For this reason, just a few other features were selected to be part of the prediction inputs in the subsequent models of machine-learning run, like Age, Country and Gender, even if their evaluation was very poor (around 0,02 or less). Similarly, regarding Job Satisfaction level, I have found out many interesting relationships among attributes that will help in my intention of finding satisfaction estimators. In fact, “Pay” and “Career planning” turned out to be by far the most important feature, having a positive correlation of 0,51 and 0,49 respectively with Job Satisfaction, but also “Peer Relationship”, “Role Clarity”, “Supervisor Support” and “Workplace Security”, in this order, got good grades of positive correlation with Job Satisfaction. I decided to leave aside other features for analysis as these other attributes tried were not meaningfully associated. Moreover, interestingly, for both the response features taken under examination, all the independent variables plotted in this analysis were found to be all positively correlated to them.



```
X = employee_data[['Absenteeism_Rate','Job_Characteristics','Teamwork_Cohesion','Corporate_Culture_and_Vision', 'Work_Environm

# plotting correlation heatmap
dataplot = sns.heatmap(X.corr(), cmap="YlGnBu", annot=True)

# displaying heatmap
plt.show()
```

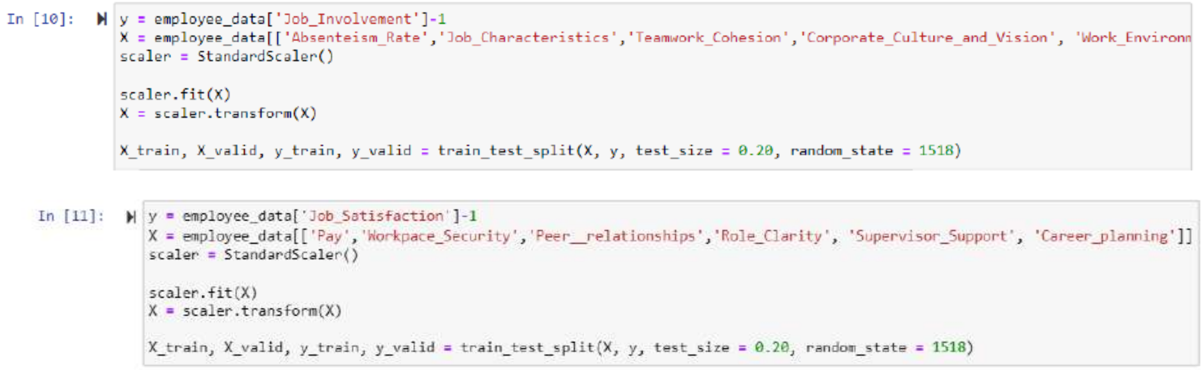
	Absenteeism_Rate	Job_Characteristics	Teamwork_Cohesion	Corporate_Culture_and_Vision	Work_Environment_Satisfaction	Age	Gender	Country	Job_Involvement
Absenteeism_Rate	1	0.016	-0.016	-0.013	0.0079	-0.011	-0.0092	0.024	0.46
Job_Characteristics	0.016	1	-0.019	-0.016	0.021	0.011	-0.0024	0.0045	0.46
Teamwork_Cohesion	-0.016	-0.019	1	0.0047	0.02	0.00014	-0.011	-0.0023	0.27
Corporate_Culture_and_Vision	-0.013	-0.016	0.0047	1	0.0086	-0.025	0.0046	-0.003	0.27
Work_Environment_Satisfaction	0.0079	0.021	0.02	0.0086	1	-0.0097	0.023	0.024	0.57
Age	-0.011	0.011	0.00014	-0.025	-0.0097	1	0.0096	-0.014	-0.0092
Gender	-0.0092	-0.0024	-0.011	0.0046	0.023	0.0096	1	0.013	0.0074
Country	0.024	0.0045	-0.0023	-0.003	0.024	-0.014	0.013	1	0.018
Job_Involvement	0.46	0.46	0.27	0.27	0.57	-0.0092	0.0074	0.018	1

```
X = employee_data[['Pay','Workpace_Security','Peer__relationships','Role_Clarity', 'Supervisor_Support', 'Career_planning','J

# plotting correlation heatmap
dataplot = sns.heatmap(X.corr(), cmap="YlGnBu", annot=True)

# displaying heatmap
plt.show()
```

	Pay	Workpace_Security	Peer__relationships	Role_Clarity	Supervisor_Support	Career_planning	Job_Satisfaction
Pay	1	0.021	-0.021	0.0017	0.011	-0.0081	0.51
Workpace_Security	0.021	1	-0.003	0.005	-0.015	-0.032	0.2
Peer__relationships	-0.021	-0.003	1	0.051	-0.0091	0.0051	0.35
Role_Clarity	0.0017	0.005	0.051	1	-0.014	0.041	0.34
Supervisor_Support	0.011	-0.015	-0.0091	-0.014	1	0.022	0.32
Career_planning	-0.0081	-0.032	0.0051	0.041	0.022	1	0.49
Job_Satisfaction	0.51	0.2	0.35	0.34	0.32	0.49	1

FIGURE 22: Heatmap Correlation Analyses for "Job Involvement" and "Job Satisfaction" and other features

After this Feature Selection Stage, with these same features, as it could be seen in figure 23, the entire datasets for each analysis have been split in a train and test set to subsequently perform the predictive analyses that I am about to present. It's important to remind that, irrespective of test sets, the response variable that was going to be forecast has always been cut off.

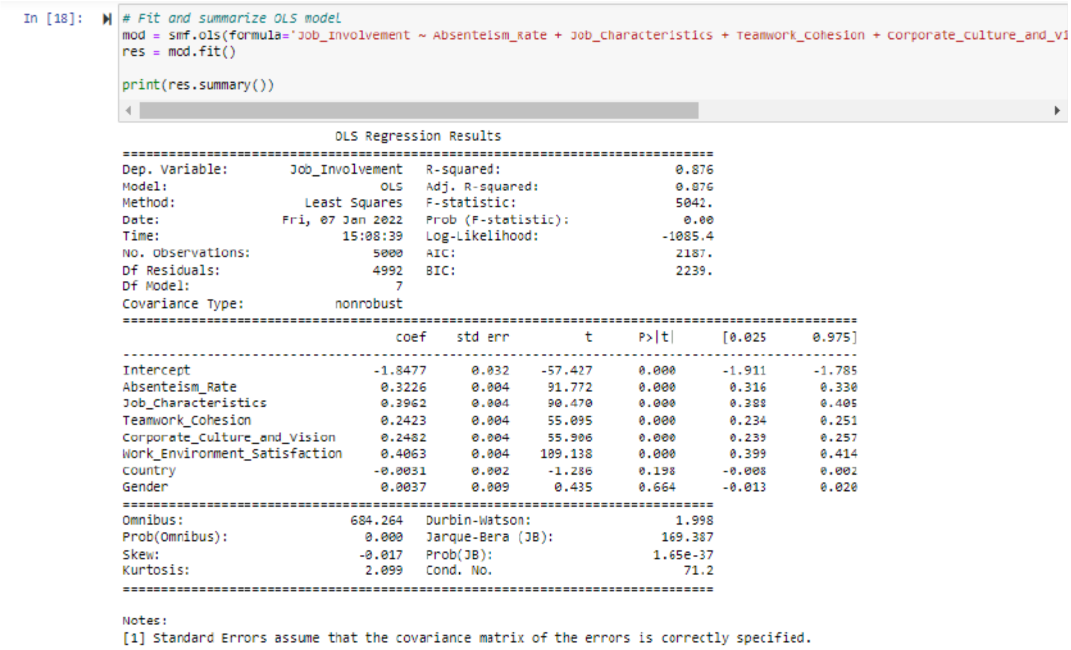


```
In [10]:  y = employee_data['Job_Involvement']-1
          X = employee_data[['Absenteism_Rate','Job_Characteristics','Teamwork_Cohesion','Corporate_Culture_and_Vision', 'Work_Environn
          scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

```
In [11]:  y = employee_data['Job_Satisfaction']-1
          X = employee_data[['Pay','Workpace_Security','Peer__relationships','Role_Clarity', 'Supervisor_Support', 'Career_planning']]
          scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

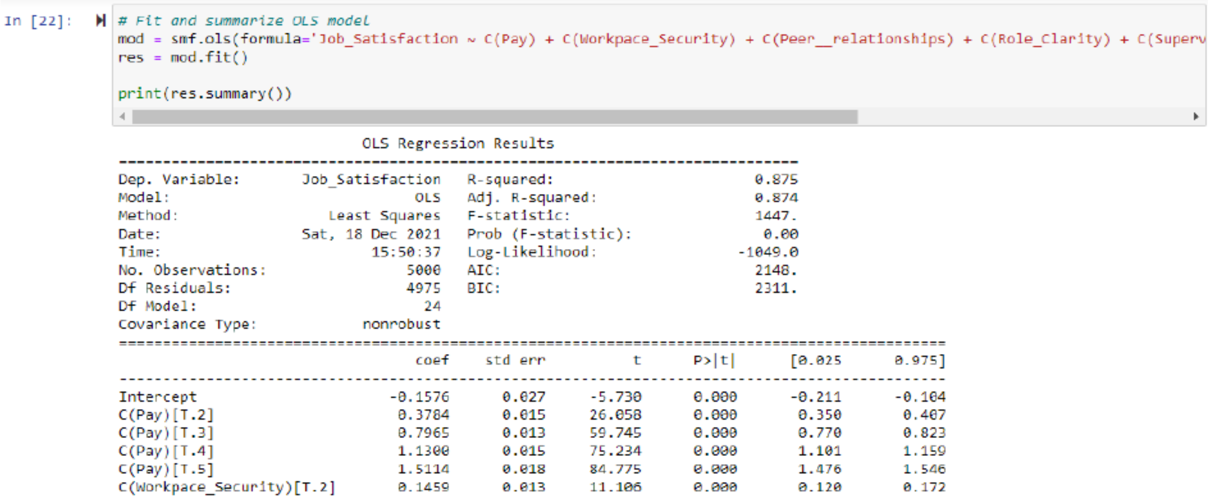
FIGURE 23: Data Preparation with train and test split for predicting “Job Involvement” and “Job Satisfaction”

The two OLS models (“Ordinary Least Square”) shown in figures 24 and 25, are two multinomial logistic regressions with several classes, which are therefore classification algorithms. They get data and try to identify some patterns of linearity (in truth, logistic regression models have a linear form for the ‘logit of success probability’, which refers to the logarithm of ‘Odds Ratio’) that account for variance between two or more features, one of which being a dependent variable; in this case, this one corresponds to "Job involvement" in the first model and "Job satisfaction" in the second. The two models have been constructed by giving, as a set of independent variables considered as predictors, the ones that were found to be significant in the results of the Correlation analysis performed earlier, with the exception of not considering "Age", "Gender" and "Country" in predicting the level of job involvement, as the linearity verification did not yield any good correlation results between these and the response variable. Conceptually, however, there is a difference to underline between the regressor’s characteristics on the two models: in the case of figure 25, since the response categories are ordered, the regressors are ordinal classed, while in the second case of figure 24 the regressors are considered continuous values. In fact, those "(C)" suggest that the system is considering those variables as classes, so, for instance, predicting “Job Satisfaction” feature, I will have the class "Pay" for each labels 2,3,4 and 5 with these numbers specified being each intercept of that independent variable (while the “1” coincides with the class baseline, which is absorbed by the estimator of the intercept), while in the second model we are considering the regressors as integers therefore as continuous variables, as the “(C)” are missing, so the model returns only the coefficients of “Pay” considered as a stand-alone integer. For the sake of analysis, since the two models had identical measures of performances (0.875 versus 0.876) and all the p-values of the variables intercept’ coefficients are significant, therefore it could be a good decision to not opt for keeping the model where the regressors are treated as classes, that results to be less easily explainable and understandable; nonetheless, if on “Pay” I had by chance seen that only label intercept “T3” was statistically meaningful, I would have begun to suspect that it was not even convenient to treat all sub-variables as classes, but better leaving them as single regressors. In this logistic model, on the other hand, I can see how much every single level of the main regressors, that correspond to the independent variables, impacts on the “Job Involvement” response employee attribute. This could result in fewer parameters and more powerful and simpler to interpret models. Overall, turning to the interpretation of the two models, the most important goodness indicator to evaluate corresponds to the "R-Squared'': this measure denotes how successfully, on a convenient 0-100%, the model captures the variance in the dependent variable, that is then collectively explained by the independent variables, measuring the strength of the relationship between the model and the dependent variable. In both outcomes, reading the two results of 0,876 and 0,875 and recognizing the limitations of the analysis, it can be stated that either models are strongly reliable and work well. Reading the values of "F-Statistics' ' together with "Prob (") "on Figure 24 and 25 results for each variable, I can instead interpret the P-value, which to make my model reliable must be under the value of 0.05, the threshold indicating a statistical significance between the two feature taken under analysis. In particular, aside from the regressors “Country” and “Gender” in the first model, the whole set of regressors analyzed and elaborated in the algorithm, turned out to be all absolutely significant for explaining the two variables to be predicted, returning a P-value of 0.000. On the contrary. This means the algorithm is excellent and when applied helps effectively to predict variation in Job Involvement and Job Satisfaction. Ultimately, "Log. Likelihood”, “AIC " and " BIC " are indicators of comparison with other models that present approximately the same figures across the two models below, not providing a real suggestion in deciding which to keep as the top model. Importantly, it must be kept in mind that in any case the "coeff "values in the results of the models, for each independent variable, are not probabilities but “loglik”, then should be converted to be better discussed.

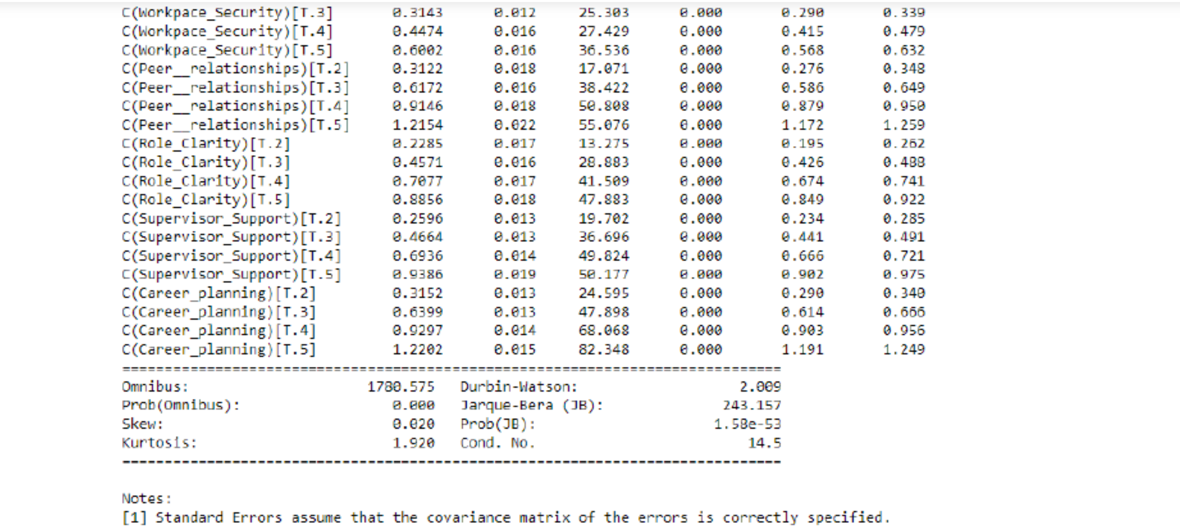


```
In [18]:  # Fit and summarize OLS model
          mod = smf.ols(formula='Job_Involvement ~ Absenteism_Rate + Job_Characteristics + Teamwork_Cohesion + Corporate_Culture_and_Vi
          res = mod.fit()

          print(res.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:        Job_Involvement   R-squared:                       0.876
Model:                            OLS   Adj. R-squared:                  0.876
Method:                 Least Squares   F-statistic:                     5042.
Date:                Fri, 07 Jan 2022   Prob (F-statistic):               0.00
Time:                        15:08:39   Log-Likelihood:                -1085.4
No. Observations:                5000   AIC:                             2187.
Df Residuals:                    4992   BIC:                             2239.
Df Model:                           7
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                        -1.8477      0.032    -57.427      0.000      -1.911      -1.785
Absenteism_Rate                   0.3226      0.004     91.772      0.000       0.316       0.330
Job_Characteristics               0.3962      0.004     90.470      0.000       0.388       0.405
Teamwork_Cohesion                 0.2423      0.004     55.095      0.000       0.234       0.251
Corporate_Culture_and_Vision      0.2482      0.004     55.906      0.000       0.239       0.257
Work_Environment_Satisfaction     0.4063      0.004    109.138      0.000       0.399       0.414
Country                          -0.0031      0.002     -1.286      0.198      -0.008       0.002
Gender                            0.0037      0.009      0.435      0.664      -0.013       0.020
==============================================================================
Omnibus:                      684.264   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              169.387
Skew:                          -0.017   Prob(JB):                     1.65e-37
Kurtosis:                       2.099   Cond. No.                         71.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

FIGURE 24: OLS Linear Regression model to predict “Job Involvement”



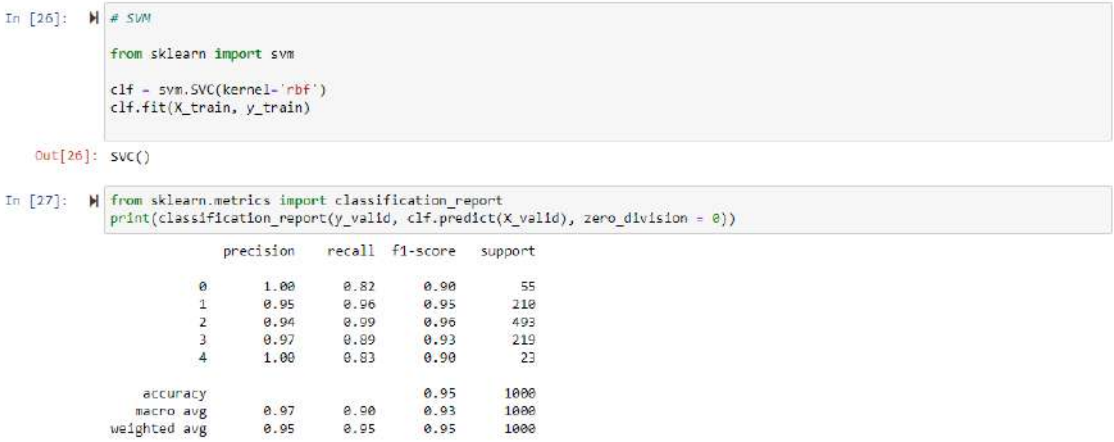
```
In [22]:  # Fit and summarize OLS model
          mod = smf.ols(formula='Job_Satisfaction ~ C(Pay) + C(Workpace_Security) + C(Peer__relationships) + C(Role_Clarity) + C(Superv
          res = mod.fit()

          print(res.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:       Job_Satisfaction   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     1447.
Date:                Sat, 18 Dec 2021   Prob (F-statistic):               0.00
Time:                        15:50:37   Log-Likelihood:                -1049.0
No. Observations:                5000   AIC:                             2148.
Df Residuals:                    4975   BIC:                             2311.
Df Model:                          24
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                    -0.1576      0.027     -5.730      0.000      -0.211      -0.104
C(Pay)[T.2]                   0.3784      0.015     26.058      0.000       0.350       0.407
C(Pay)[T.3]                   0.7965      0.013     59.745      0.000       0.770       0.823
C(Pay)[T.4]                   1.1300      0.015     75.234      0.000       1.101       1.159
C(Pay)[T.5]                   1.5114      0.018     84.775      0.000       1.476       1.546
C(Workpace_Security)[T.2]     0.1459      0.013     11.106      0.000       0.120       0.172
```



```
C(Workpace_Security)[T.3]        0.3143      0.012     25.303      0.000       0.290       0.339
C(Workpace_Security)[T.4]        0.4474      0.016     27.429      0.000       0.415       0.479
C(Workpace_Security)[T.5]        0.6002      0.016     36.536      0.000       0.568       0.632
C(Peer__relationships)[T.2]      0.3122      0.018     17.071      0.000       0.276       0.348
C(Peer__relationships)[T.3]      0.6172      0.016     38.422      0.000       0.586       0.649
C(Peer__relationships)[T.4]      0.9146      0.018     50.808      0.000       0.879       0.950
C(Peer__relationships)[T.5]      1.2154      0.022     55.076      0.000       1.172       1.259
C(Role_Clarity)[T.2]             0.2285      0.017     13.275      0.000       0.195       0.262
C(Role_Clarity)[T.3]             0.4571      0.016     28.883      0.000       0.426       0.488
C(Role_Clarity)[T.4]             0.7077      0.017     41.509      0.000       0.674       0.741
C(Role_Clarity)[T.5]             0.8856      0.018     47.883      0.000       0.849       0.922
C(Supervisor_Support)[T.2]       0.2596      0.013     19.702      0.000       0.234       0.285
C(Supervisor_Support)[T.3]       0.4664      0.013     36.696      0.000       0.441       0.491
C(Supervisor_Support)[T.4]       0.6936      0.014     49.824      0.000       0.666       0.721
C(Supervisor_Support)[T.5]       0.9386      0.019     50.177      0.000       0.902       0.975
C(Career_planning)[T.2]          0.3152      0.013     24.595      0.000       0.290       0.340
C(Career_planning)[T.3]          0.6399      0.013     47.898      0.000       0.614       0.666
C(Career_planning)[T.4]          0.9297      0.014     68.068      0.000       0.903       0.956
C(Career_planning)[T.5]          1.2202      0.015     82.348      0.000       1.191       1.249
==============================================================================
Omnibus:                     1780.575   Durbin-Watson:                   2.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              243.157
Skew:                           0.020   Prob(JB):                     1.58e-53
Kurtosis:                       1.920   Cond. No.                         14.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

FIGURE 25: OLS Logistic Regression model to predict "Job Satisfaction"

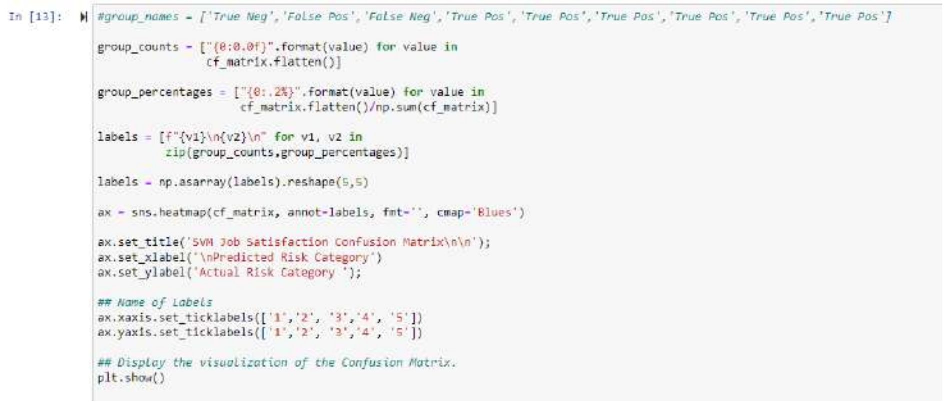
The second machine-learning performed in this analysis regarding the realm of employee engagement refers to the application of a Support Vector Machine classifier (SVM) that would supposedly predict the employee feature "Job Satisfaction", learning first from the data patterns the algorithm identified in the training data set, and the making prediction on the leftover part of the data, which doesn't contain this researched feature on purpose. The two sets, once again, include the six relative correlated variables found through the initial Feature Selection process: "Pay", "Workplace Security", "Peer Relationship", "Role Clarity", "Supervisor Support" and "Career Planning". SVM, as seen in the previous chapters, is a very powerful machine learning tool that, in its process of understanding how to classify the instances that pass through, graphically represents the given data in a high dimensional feature space called Kernel, aiming at maximizing the margin between the classes. Since, in all these experimental cases, I gave the algorithm more than two features to use in the learning phase, the algorithm, categorizing the data set on the sides of many hyperplanes, used then a non-linear separation that could not be graphically reported as well as the hyperplanes themselves. Still, a graphical representation of a partial linear separation could have been obtained applying a PCA, "principal component analysis" to the algorithm, in order to see how much variability just two or three components have in the model, leaving all the other variables aside. In figure 26, once the algorithm has been imported in the system, our classifier has been fitted with the training dataset and given the indication to make a prediction on the test dataset. After ultimating the modeling process on the train set, the process of testing data is performed and the evaluation metrics values can be measured. Generally, in fact, evaluation metrics need to be well designed when analysts ascertain the performance of the model and subsequently compare it with other different models. In this work, four evaluation metrics are employed: precision, recall, f1 score, and support.



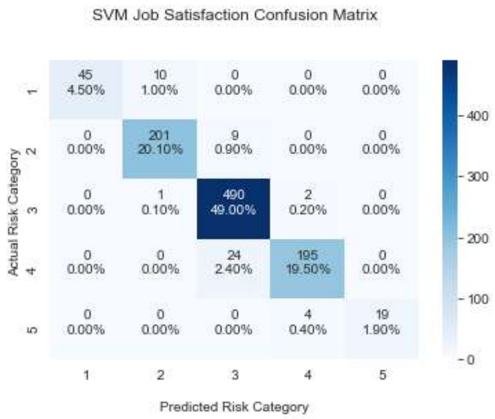
```
In [26]:  # SVM

          from sklearn import svm

          clf = svm.SVC(kernel='rbf')
          clf.fit(X_train, y_train)

Out[26]:  SVC()

In [27]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       1.00      0.82      0.90        55
                     1       0.95      0.96      0.95       210
                     2       0.94      0.99      0.96       493
                     3       0.97      0.89      0.93       219
                     4       1.00      0.83      0.90        23

              accuracy                           0.95      1000
             macro avg       0.97      0.90      0.93      1000
          weighted avg       0.95      0.95      0.95      1000
```

Figure 26: Basic SVM classifier model fit and relative classification report; “Job Satisfaction” prediction

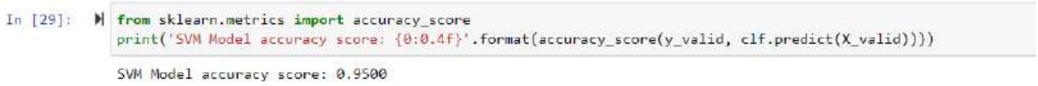
All four measures vary from 0.0 to 1.0; “precision” gives an assessment of the classifier’s exactness in predicting the level of Job Satisfaction associated for each employee of the test dataset, returning the ratio of truly predicted positive values with respect to the total positively true and false values; whereas, “recall” gives a measure of the classifier’s completeness, meaning how the machine correctly identified all positive samples, computing the percentage of all positive instances classified correctly. Precision and recall are inversely related to each other. “f1 score” instead is a kind of weighted mean of precision and recall, that is embedded in its computation, and is used to compare models of each other. Ultimately, the “support” measure corresponds to the number of actual occurrences of the class in the determined dataset, and for this reason, doesn't change across models. Besides the final results, it’s important since it indicates whether there is an evident structural weakness in the reported outcomes of the classifier as a result of imbalance support in the training data, calling for some resampling methods to be applied (Scikit-Learn). Moreover, with the help of the confusion matrix illustrated in figure 27, I find out the number of values that have been misclassified in the “Job Satisfaction” predictions on the test dataset. For instance, along the first row of this SVM’s confusion matrix, “45” indicated the total number of correctly predicted items for the first label, called also “True Positive” values, while “10” refer to the “False Negative” predicted sample that, in this specific example, have been classified as belonging to the second label, but were actually of the first one. Overall, along the diagonal, are found the observations that have been then successfully predicted by the algorithm. Rather, looking at the third column, I could find the number of “False Positive” by looking at the number “9” or “24” which were forecasted inside the label number three but should have been classified in the second and fourth label, respectively. The final accuracy score on the test dataset is of 0.95 which is an extraordinary result that indicates that the model is trusted.



```
In [13]:  #group_names = ['True Neg','False Pos','False Neg','True Pos','True Pos','True Pos','True Pos','True Pos','True Pos']

          group_counts = ["{0:0.0f}".format(value) for value in
                          cf_matrix.flatten()]

          group_percentages = ["{0:.2%}".format(value) for value in
                               cf_matrix.flatten()/np.sum(cf_matrix)]

          labels = [f"{v1}\n{v2}\n" for v1, v2 in
                    zip(group_counts,group_percentages)]

          labels = np.asarray(labels).reshape(5,5)

          ax = sns.heatmap(cf_matrix, annot=labels, fmt='', cmap='Blues')

          ax.set_title('SVM Job Satisfaction Confusion Matrix\n\n');
          ax.set_xlabel('\nPredicted Risk Category')
          ax.set_ylabel('Actual Risk Category ');

          ## Name of Labels
          ax.xaxis.set_ticklabels(['1','2', '3','4', '5'])
          ax.yaxis.set_ticklabels(['1','2', '3','4', '5'])

          ## Display the visualization of the Confusion Matrix.
          plt.show()
```



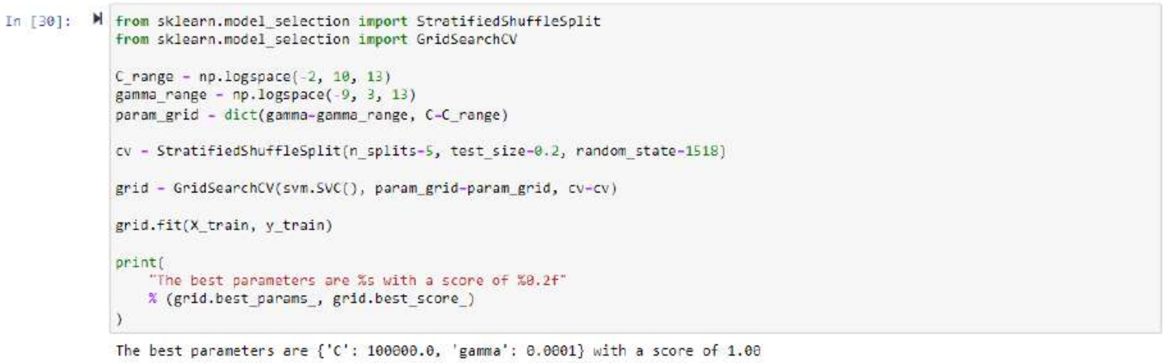
	1	2	3	4	5
1	45 4.50%	10 1.00%	0 0.00%	0 0.00%	0 0.00%
2	0 0.00%	201 20.10%	9 0.90%	0 0.00%	0 0.00%
3	0 0.00%	1 0.10%	490 49.00%	2 0.20%	0 0.00%
4	0 0.00%	0 0.00%	24 2.40%	195 19.50%	0 0.00%
5	0 0.00%	0 0.00%	0 0.00%	4 0.40%	19 1.90%



```
In [29]:  from sklearn.metrics import accuracy_score
          print('SVM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          SVM Model accuracy score: 0.9500
```

Figure 27: SVM confusion matrix and accuracy score; “Job Satisfaction” prediction

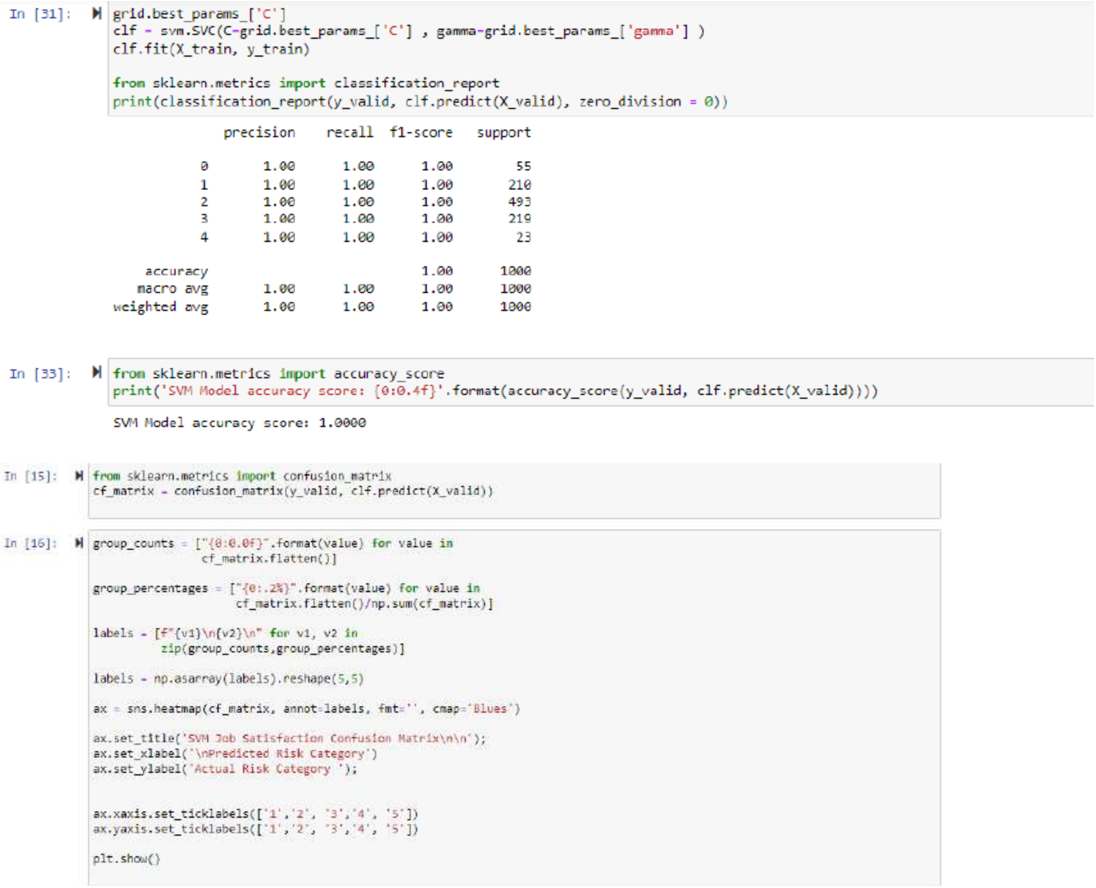
On a second time, in trying to even improve the SVM model, I opted to enter the process of Hyperparameter Tuning, whose goal is to search for the optimal parameter values, fitting the model at hand for each possible combination of, in this case, the two values associated with the parameter “C” and “Gamma”, and ultimately extracting the model with the highest accuracy score after applying a “Grid Search” Cross-validation Method to resampling my train and test sets. This process is reported in figure 28.



```
In [30]:  from sklearn.model_selection import StratifiedShuffleSplit
          from sklearn.model_selection import GridSearchCV

          C_range = np.logspace(-2, 10, 13)
          gamma_range = np.logspace(-9, 3, 13)
          param_grid = dict(gamma=gamma_range, C=C_range)

          cv = StratifiedShuffleSplit(n_splits=5, test_size=0.2, random_state=1518)

          grid = GridSearchCV(svm.SVC(), param_grid=param_grid, cv=cv)

          grid.fit(X_train, y_train)

          print(
              "The best parameters are %s with a score of %0.2f"
              % (grid.best_params_, grid.best_score_)
          )

          The best parameters are {'C': 100000.0, 'gamma': 0.0001} with a score of 1.00
```

Figure 28: Grid Search performed in the SVM model to predict “Job Satisfaction”

With this effort, as it could be seen in figure 29, either the new classification report and the accuracy scores applied to the test dataset returned now have reached 100% of perfection in all the indicators given, meaning that, recognizing the limitations of the model, the SVM built has been able to successfully predict all the labels of Job Satisfaction level f the employees in the test set.



```
In [31]:  grid.best_params_['C']
          clf = svm.SVC(C=grid.best_params_['C'] , gamma=grid.best_params_['gamma'] )
          clf.fit(X_train, y_train)

          from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       1.00      1.00      1.00        55
                     1       1.00      1.00      1.00       210
                     2       1.00      1.00      1.00       493
                     3       1.00      1.00      1.00       219
                     4       1.00      1.00      1.00        23

              accuracy                           1.00      1000
             macro avg       1.00      1.00      1.00      1000
          weighted avg       1.00      1.00      1.00      1000

In [33]:  from sklearn.metrics import accuracy_score
          print('SVM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          SVM Model accuracy score: 1.0000

In [15]:  from sklearn.metrics import confusion_matrix
          cf_matrix = confusion_matrix(y_valid, clf.predict(X_valid))

In [16]:  group_counts = ["{0:0.0f}".format(value) for value in
                          cf_matrix.flatten()]

          group_percentages = ["{0:.2%}".format(value) for value in
                               cf_matrix.flatten()/np.sum(cf_matrix)]

          labels = [f"{v1}\n{v2}\n" for v1, v2 in
                    zip(group_counts,group_percentages)]

          labels = np.asarray(labels).reshape(5,5)

          ax = sns.heatmap(cf_matrix, annot=labels, fmt='', cmap='Blues')

          ax.set_title('SVM Job Satisfaction Confusion Matrix\n\n');
          ax.set_xlabel('\nPredicted Risk Category')
          ax.set_ylabel('Actual Risk Category ');


          ax.xaxis.set_ticklabels(['1','2', '3','4', '5'])
          ax.yaxis.set_ticklabels(['1','2', '3','4', '5'])

          plt.show()
```

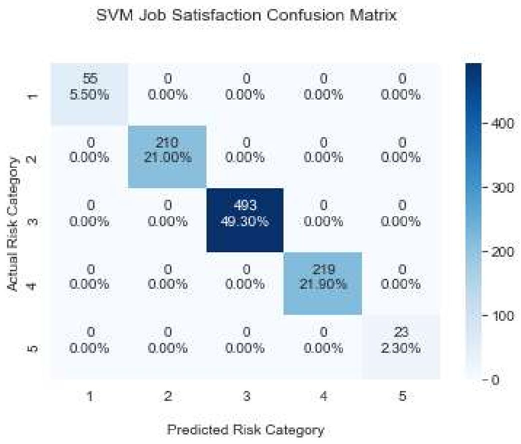
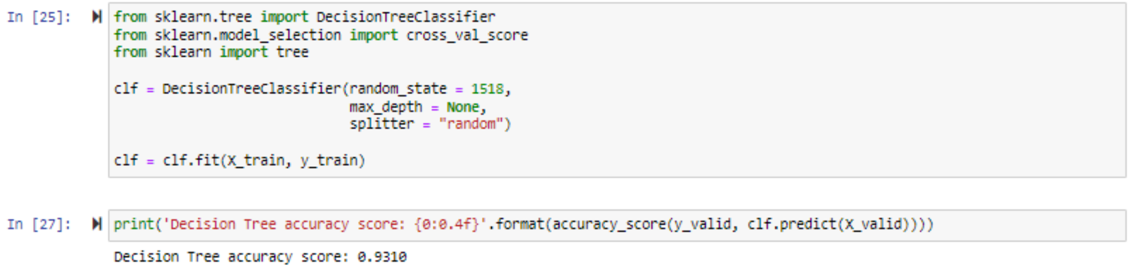
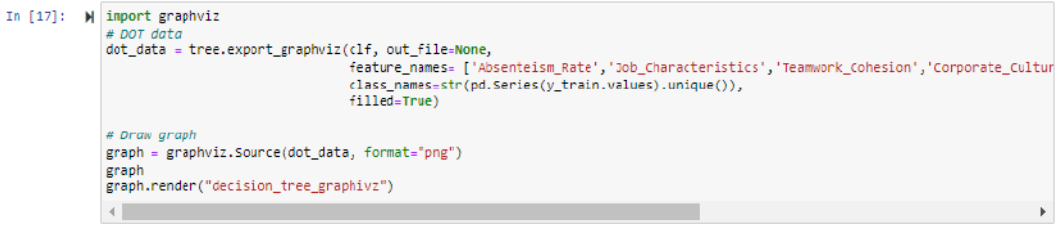


Figure 29: Classification report, confusion matrix and accuracy score of the SVM model after the Grid Search; "Job Satisfaction" prediction

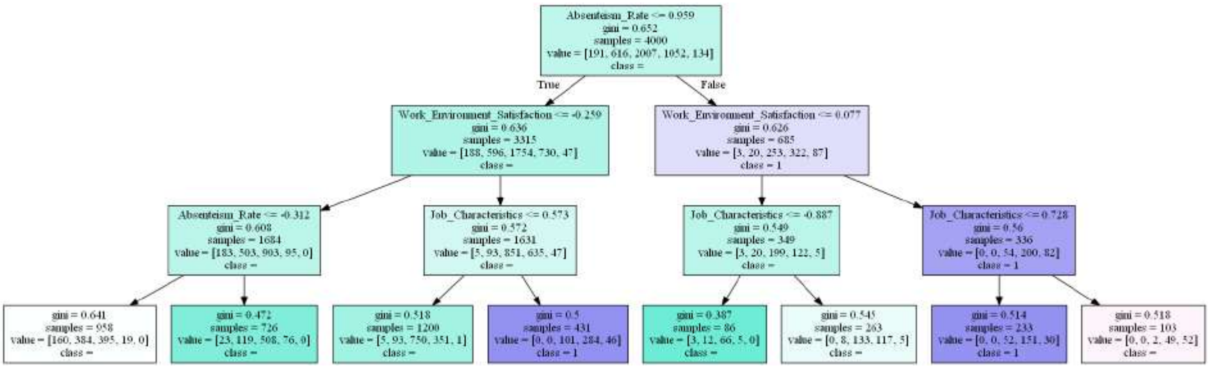
I now approach the last part of employee engagement analysis in which I've tried to predict the level of "Job Involvement" using two tree-based machine-learning algorithms. Oftentimes, these types of algorithms have the "overfitting" problem by which they often create trees that are all the same, meaning they finish with leaves alike each other. In fact, the decision tree performs well only in the training phase with the data on which the tree is fitted, but when I give the model new data on which to test the tree, it is often not able to understand how to make the prediction. To tune a tree and implement a wise tree growing strategy, many constraints could be specified, like "max_depth", " min_samples_split", "min_samples_leaf", "max_leaf_nodes", "min_impurity_decrease", and "min_impurity_split". In figure 30, after importing the algorithm command line, I indicated the maximum depth of the tree giving no precise range and the type of splitter strategy to be used: the result is an accuracy score pretty high of 0.93. In the same figure, has also been attached an image of the first three leaves levels of the tree by importing the Seaborn's package 'graphviz', to give an idea to the readers of how a decision tree classifier looks and works.



```
In [25]:  from sklearn.tree import DecisionTreeClassifier
          from sklearn.model_selection import cross_val_score
          from sklearn import tree

          clf = DecisionTreeClassifier(random_state = 1518,
                                       max_depth = None,
                                       splitter = "random")

          clf = clf.fit(X_train, y_train)

In [27]:  print('Decision Tree accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          Decision Tree accuracy score: 0.9310
```



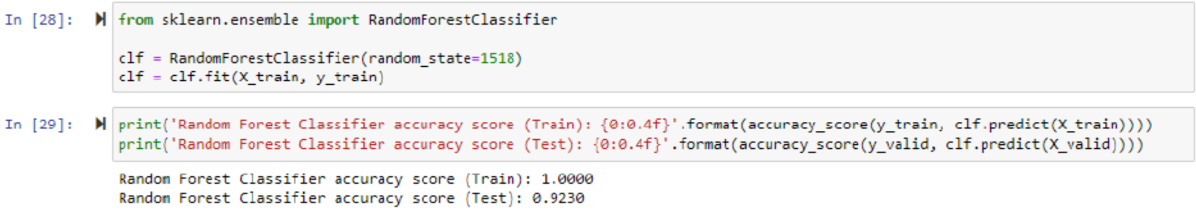
```
In [17]:  import graphviz
          # DOT data
          dot_data = tree.export_graphviz(clf, out_file=None,
                                          feature_names= ['Absenteism_Rate','Job_Characteristics','Teamwork_Cohesion','Corporate_Cultur
                                          class_names=str(pd.Series(y_train.values).unique()),
                                          filled=True)

          # Draw graph
          graph = graphviz.Source(dot_data, format="png")
          graph
          graph.render("decision_tree_graphivz")
```



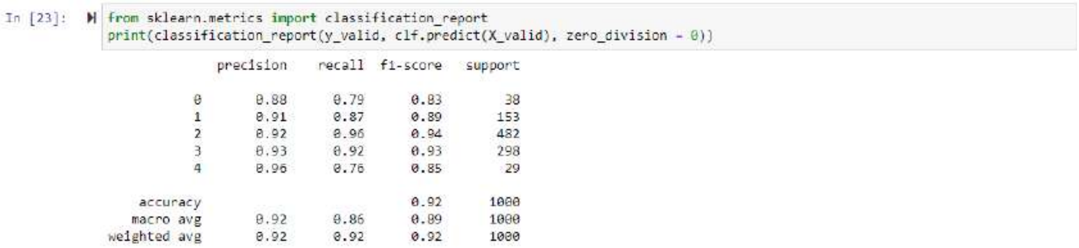
Absenteism_Rate <= 0.959
gini = 0.652
samples = 4000
value = [191, 616, 2007, 1052, 134]
class =
True
False
Work_Environment_Satisfaction <= -0.259
gini = 0.636
samples = 3315
value = [188, 596, 1754, 730, 47]
class =
Work_Environment_Satisfaction <= 0.077
gini = 0.626
samples = 685
value = [3, 20, 253, 322, 87]
class = 1
Absenteism_Rate <= -0.312
gini = 0.608
samples = 1684
value = [183, 503, 903, 95, 0]
class =
Job_Characteristics <= 0.573
gini = 0.572
samples = 1631
value = [5, 93, 851, 635, 47]
class =
Job_Characteristics <= -0.887
gini = 0.549
samples = 349
value = [3, 20, 199, 122, 5]
class =
Job_Characteristics <= 0.728
gini = 0.56
samples = 336
value = [0, 0, 54, 200, 82]
class = 1
gini = 0.641
samples = 958
value = [160, 384, 395, 19, 0]
class =
gini = 0.472
samples = 726
value = [23, 119, 508, 76, 0]
class =
gini = 0.518
samples = 1200
value = [5, 93, 750, 351, 1]
class =
gini = 0.5
samples = 431
value = [0, 0, 101, 284, 46]
class = 1
gini = 0.387
samples = 86
value = [3, 12, 66, 5, 0]
class =
gini = 0.545
samples = 263
value = [0, 8, 133, 117, 5]
class =
gini = 0.514
samples = 233
value = [0, 0, 52, 151, 30]
class = 1
gini = 0.518
samples = 103
value = [0, 0, 2, 49, 52]
class =

Figure 30: Decision Tree Classifier for predicting “Job Involvement”

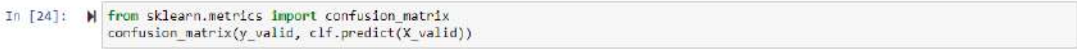
Following on in my experiments, a basic random forest, that is a modification of the standard tree growing algorithm, was applied to my simulated dataset, obtaining an accuracy score of 0.92 as shown in figure 31. In the same figure, it’s also computed and reported the performance evaluation of the random forest classifier using the confusion matrix and the classification report.



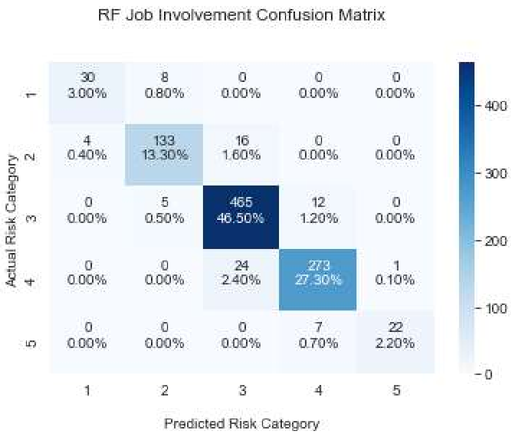
```
In [28]:  from sklearn.ensemble import RandomForestClassifier

          clf = RandomForestClassifier(random_state=1518)
          clf = clf.fit(X_train, y_train)

In [29]:  print('Random Forest Classifier accuracy score (Train): {0:0.4f}'.format(accuracy_score(y_train, clf.predict(X_train))))
          print('Random Forest Classifier accuracy score (Test): {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          Random Forest Classifier accuracy score (Train): 1.0000
          Random Forest Classifier accuracy score (Test): 0.9230
```



```
In [23]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       0.88      0.79      0.83        38
                     1       0.91      0.87      0.89       153
                     2       0.92      0.96      0.94       482
                     3       0.93      0.92      0.93       298
                     4       0.96      0.76      0.85        29

              accuracy                           0.92      1000
             macro avg       0.92      0.86      0.89      1000
          weighted avg       0.92      0.92      0.92      1000
```



```
In [24]:  from sklearn.metrics import confusion_matrix
          confusion_matrix(y_valid, clf.predict(X_valid))
```



	1	2	3	4	5
1	30 3.00%	8 0.80%	0 0.00%	0 0.00%	0 0.00%
2	4 0.40%	133 13.30%	16 1.60%	0 0.00%	0 0.00%
3	0 0.00%	5 0.50%	465 46.50%	12 1.20%	0 0.00%
4	0 0.00%	0 0.00%	24 2.40%	273 27.30%	1 0.10%
5	0 0.00%	0 0.00%	0 0.00%	7 0.70%	22 2.20%

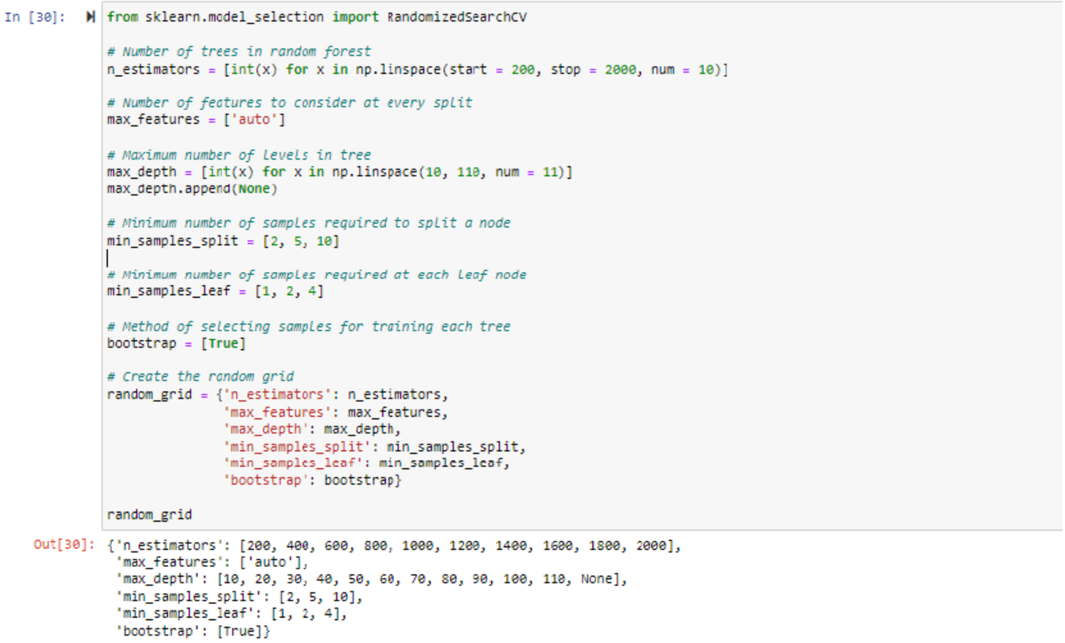
Figure 31: Random Forest Classifier for predicting "Job Involvement" with relative confusion matrix and classification report

In figure 32, proceeding with the RF analysis, it is performed a hyperparameter tuning process, using a "Randomized Search" to select the optimal hyper parameters (that are parameters used by the random forest algorithm to find the parameters to estimate) that would align to my dataset, specifying to the system to search within some parameter ranges for which running many iterations of the model, from which it will be selected the most accurate. Appositely, I chose to use a Randomized Search, that performs a set of iterative attempts, and not a "Grid Search" since it would have required lots of processing power and time for completing the great number of possible combinations for each single parameter. It's important to remember that in tree-like structure it is necessary to deal with a trade-off between the level of (usually very high) accuracy on the prediction based on the train test and the level of accuracy on data external to the learned dataset. Finally, in figure 32 it has been created a random grid of parameters for the RF model, making it clear to the system the range of values from which to start the iterations of the various models. These parameters were the following:

- 'n_estimators', that is the number of trees
- 'max_feature', coinciding to the number of features to consider at every split (the instruction "auto" commands to take the square root of the total number of feature, so

that, differently from standard trees, at each split in the tree, the algorithm considers automatically only a reduced subset of the best available predictors among all features)

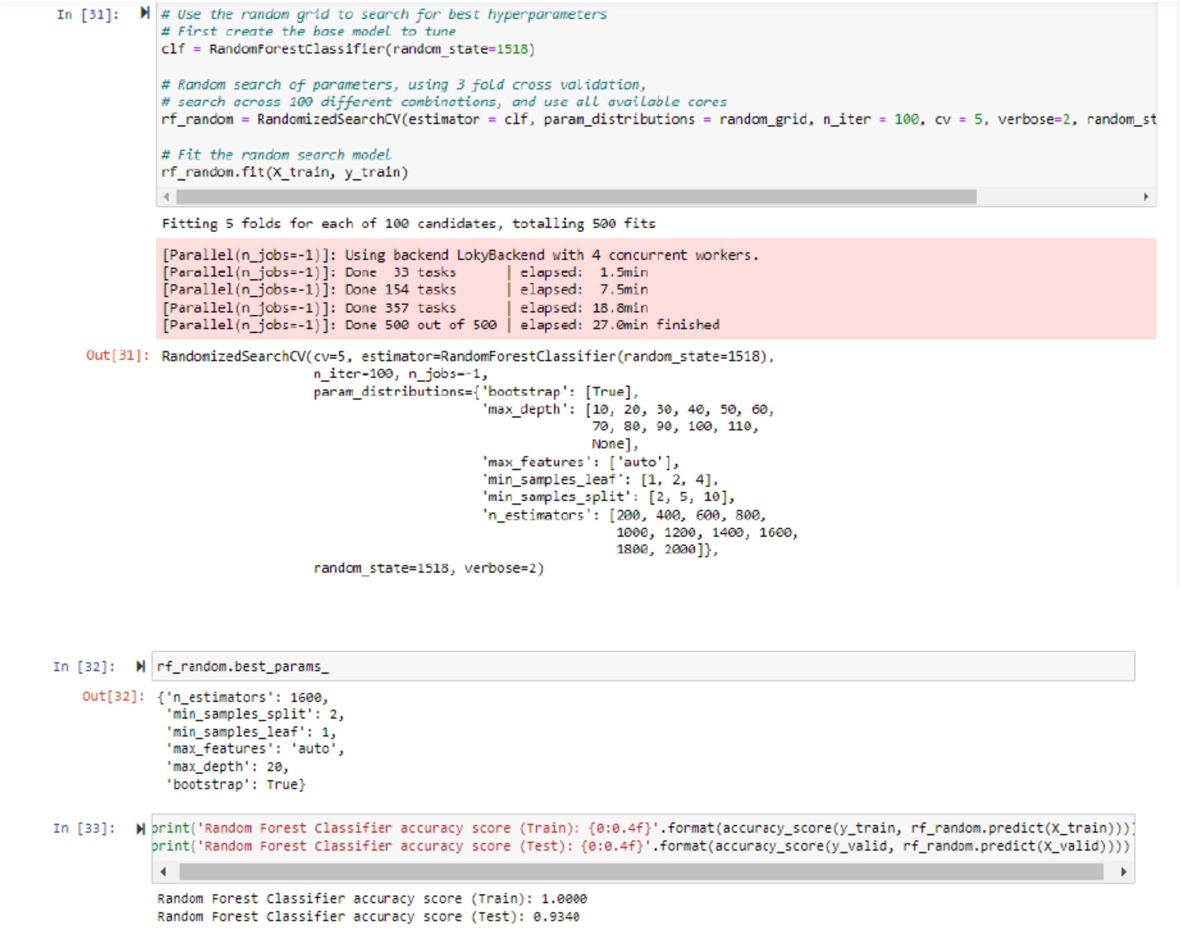
- ‘max_depth’, that represents the maximum number of leaves at which stopping the growth of the trees
- ‘min_sample_split’, corresponding to the minimum number of samples to split a node, meaning that, in a decisional node, a minimum number n of employees were considered
- ‘min_sample_leaf’, the minimum number of samples required at each leaf node
- ‘bootstrap’, that is a boolean value to confirm or not the application of the bootstrapping method as method of selecting samples for training each tree



```
In [30]:  from sklearn.model_selection import RandomizedSearchCV

          # Number of trees in random forest
          n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]

          # Number of features to consider at every split
          max_features = ['auto']

          # Maximum number of levels in tree
          max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
          max_depth.append(None)

          # Minimum number of samples required to split a node
          min_samples_split = [2, 5, 10]

          # Minimum number of samples required at each leaf node
          min_samples_leaf = [1, 2, 4]

          # Method of selecting samples for training each tree
          bootstrap = [True]

          # Create the random grid
          random_grid = {'n_estimators': n_estimators,
                         'max_features': max_features,
                         'max_depth': max_depth,
                         'min_samples_split': min_samples_split,
                         'min_samples_leaf': min_samples_leaf,
                         'bootstrap': bootstrap}

          random_grid

Out[30]:  {'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000],
           'max_features': ['auto'],
           'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
           'min_samples_split': [2, 5, 10],
           'min_samples_leaf': [1, 2, 4],
           'bootstrap': [True]}
```

Figure 32: Randomized Search process to improve the RF model

In figure 33, the process just started below goes further and after having given indications about the forest’s characteristics, I fit the random search model that, completing exactly 500 iterations with different combinations of hyper-parameters, returns me, using the line command “rf_random.best_params_”, the best parameters to improve the accuracy of the model, that will reach an outline accuracy score of 0.934.



```
In [31]:  # Use the random grid to search for best hyperparameters
          # First create the base model to tune
          clf = RandomForestClassifier(random_state=1518)

          # Random search of parameters, using 3 fold cross validation,
          # search across 100 different combinations, and use all available cores
          rf_random = RandomizedSearchCV(estimator = clf, param_distributions = random_grid, n_iter = 100, cv = 5, verbose=2, random_st

          # Fit the random search model
          rf_random.fit(X_train, y_train)

          Fitting 5 folds for each of 100 candidates, totalling 500 fits

          [Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
          [Parallel(n_jobs=-1)]: Done  33 tasks      | elapsed:  1.5min
          [Parallel(n_jobs=-1)]: Done 154 tasks      | elapsed:  7.5min
          [Parallel(n_jobs=-1)]: Done 357 tasks      | elapsed: 18.8min
          [Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed: 27.0min finished

Out[31]:  RandomizedSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1518),
                             n_iter=100, n_jobs=-1,
                             param_distributions={'bootstrap': [True],
                                                  'max_depth': [10, 20, 30, 40, 50, 60,
                                                                70, 80, 90, 100, 110,
                                                                None],
                                                  'max_features': ['auto'],
                                                  'min_samples_leaf': [1, 2, 4],
                                                  'min_samples_split': [2, 5, 10],
                                                  'n_estimators': [200, 400, 600, 800,
                                                                   1000, 1200, 1400, 1600,
                                                                   1800, 2000]},
                             random_state=1518, verbose=2)
```

```
In [32]:  rf_random.best_params_

Out[32]:  {'n_estimators': 1600,
           'min_samples_split': 2,
           'min_samples_leaf': 1,
           'max_features': 'auto',
           'max_depth': 20,
           'bootstrap': True}
```

```
In [33]:  print('Random Forest Classifier accuracy score (Train): {0:0.4f}'.format(accuracy_score(y_train, rf_random.predict(X_train)))
          print('Random Forest Classifier accuracy score (Test): {0:0.4f}'.format(accuracy_score(y_valid, rf_random.predict(X_valid))))

          Random Forest Classifier accuracy score (Train): 1.0000
          Random Forest Classifier accuracy score (Test): 0.9340
```

Figure 33: New Random Forest Classifier, with the application of best parameters, fit with the data and with an improved accuracy score

3.3.2 Turnover Risk Analysis

In this second case of analysis, the level of turnover risk trend within my simulated labor force has been descriptively inspected in the first part, while in the second this important employee feature has been predicted with the help of the array of machine learning algorithms selected for this study, to discover what employee factors are causing employees leaving in the company. The starting dashboard illustrates a first general panoramic of this indicator's trend among employees. Figure 34 shows the "Average Turnover Risk" KPI that corresponds to a score of 3, which, always taking into consideration that the range scale is from 1 to 5, is a result on which to keep a special eye. In fact, a better score of 2 or even 1, averagely, could make the HR top manager way less worried about employees wanting to dismiss; as we have learned over the previous sections, this trend could turn out to be a source of many problems for the company, in terms of costs and reputation. Still, at least this measure is not as problematic as it could have been if the average score had reached a score of 4 or 5 of attrition rate. Examining the values of turnover risk for each particular level and divided by gender, however, HR decision-makers should ask themselves some questions with regard to the low but significant difference between male and female likely intentions to quit: From the same engagement survey, the female population is found to be perceivably more prone to show a higher turnover risk having, in proportion, more employees giving a score of 4 or 5 and at the same time, fewer employees belonging to the first three more desirable scores of Turnover Risk. The highest variation between the two clusters is 8 percentage points corresponding to a larger part of male employees who graded their turnover risk with the lowest score. This dashboard then provides a point of reflection to reconsider the retention plans for the workforce part of the population who is female.

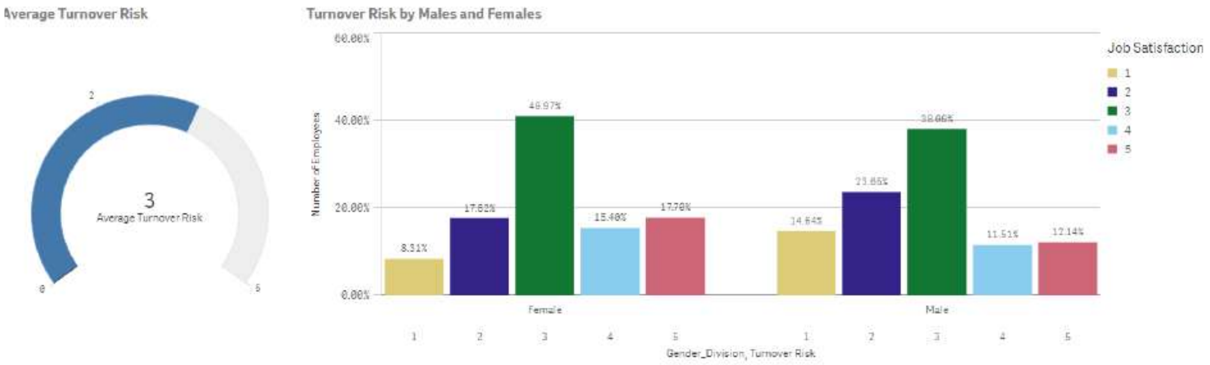
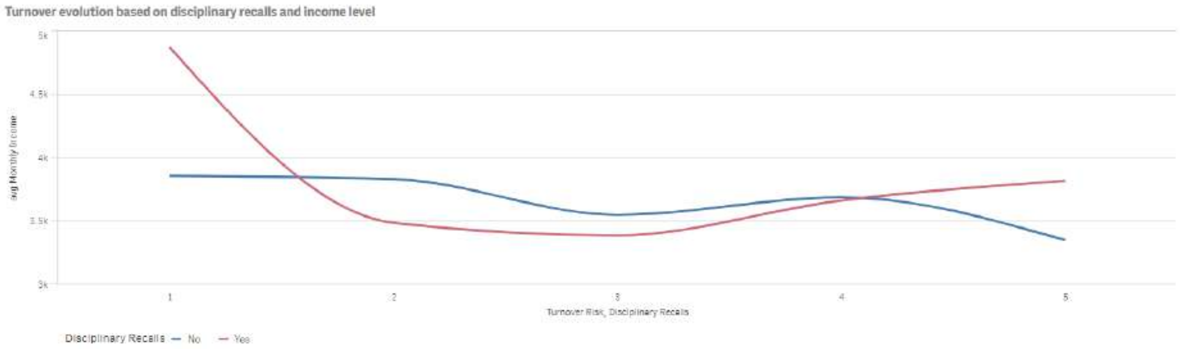


FIGURE 34: Turnover Risk Overview KPI and Gender division bar chart

In attempting to analyze a bit deeper the turnover trend inside this company, I studied how its curve changed by comparing average differences in turnover intentions for employees across a number of other possibly affecting variables. For instance, the dashboard designed in figure 35 is exhaustively explicative of the situation; as a matter of fact, the first line chart put in relation each other, across the five different scoring levels of employee turnover risk, the average monthly salary for two generated clusters of workers: who have received, along their permanence at the company, at least a disciplinary recall and or doesn't, have been always faithful to company rules and managers' procedures. The results orient the point of view of HR analysts clearly pointing out the difference between the two groups behaviors: the curve of the workforce with no disciplinary warning appears to be quite constant, prefiguring a slight descending trend that brought employees to have a greater risk of quitting as their salary level lowers under the average of 3.8 thousand euros per month. On the other hand, from the shape of the disciplinary warned employees curve, a clear pattern is highlighted for what concerns employees who showed the lowest risk of turnover intentions: in comparison to other turnover levels assessed, they belong to a segment of the workforce that averagely have an income significantly higher, distancing other of more than a thousand euro per month of extra earnings. In the same dashboard, instead, a bar chart, in the first place, stresses out another behavioral characteristic of the labor force, that is explained by the HR metrics "Stability Rate": this measure gives an idea of the turnover risk distribution, over time, in years spent in the company, and depending on the employment nature, for which employees are tied to the company with a permanent, temporary or part-time contract. Curiously, for each level of turnover risk, part-time employed workers, though far less in number than permanent-contract employees and a bit more than temporary-contract ones, as highlighted in the Employment nature overview pie chart to the right side, tend to evidently have a longer tenure in the company, while, oppositely and logically, temporary employees turn out to have been in the company for fewer years.



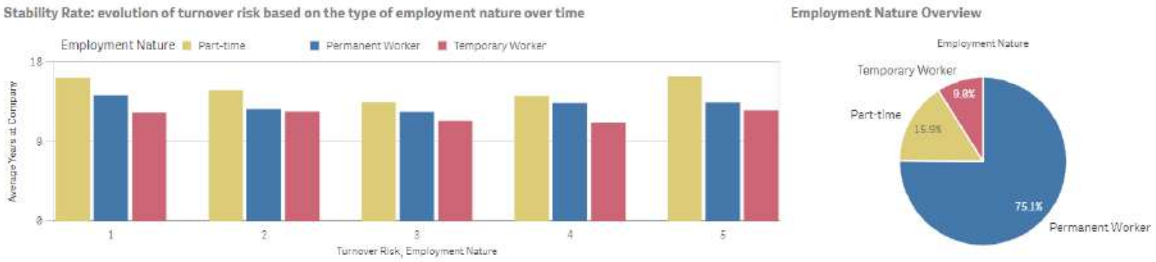


FIGURE 35: Disciplinary Recall and Income incidence line chart and "Stability Rate" bar and pie charts

On the following dashboard depicted in figure 36, a diversity heat-map analysis returns some numerical figures to base a juxtaposition of employees on, dividing the turnover trend by the country of origin of the workforce. For the evident reason of dimension, the critical red boxes that visually stand out the most are those associated with the Italian country, where, as for the other countries, it is highlighted that the majority of employees belong to the third, average level of turnover risk score. What HR managers should place their attention on, by the way, is that for almost every country there are more employees in the fifth level of turnover range, that is more problematic for the company than in the first level, that is rather the most auspicious. This means that even if, from the KPI or the previous dashboard, the mean lies in the score 3, some dangerous cases should be carefully handled by the HR staff to guarantee to not go into a worse overall trend direction. Bad signals are emerging, above all, from German, American and Portuguese employees, which had almost twice as many employees scoring 5 than employees scoring 1. The second mekko chart of this same dashboard shows another interesting relation in how workforces feel about a chance to abandon based on their managerial position and on their average salary rate. Surprisingly, in the first top-three managerial roles in the company, Executive, senior and middle managers exhibit a controversial attitude: personnel receiving higher income seem to have scores, for either three positions taken in considerations, 5 or 4 as turnover risk rate, meaning that there seem to not exist a negative correlation corresponding to a reduction of the perceived employee level for any increment of the pay levels as it happened before analyzing the behaviors of employees with or without at least a disciplinary recall. On the contrary, from this graph, it's noticeably evident that for each of the five managerial positions' division, the smallest area in terms of income is constantly the one associated with the lowest level of turnover risk expressed by the employees; as employees least prepared to voluntary leave the firm are the ones who have the average lowest monthly income level, C-suite level should therefore not being stick to thinking that the only variable affecting the turnover risk level is the salary, but should be enacted by some other more advanced analytical instruments to infer critical consideration that could effectively reduce the overall risk of turnover in the company keeping indispensable workers in order to meet the business targets.

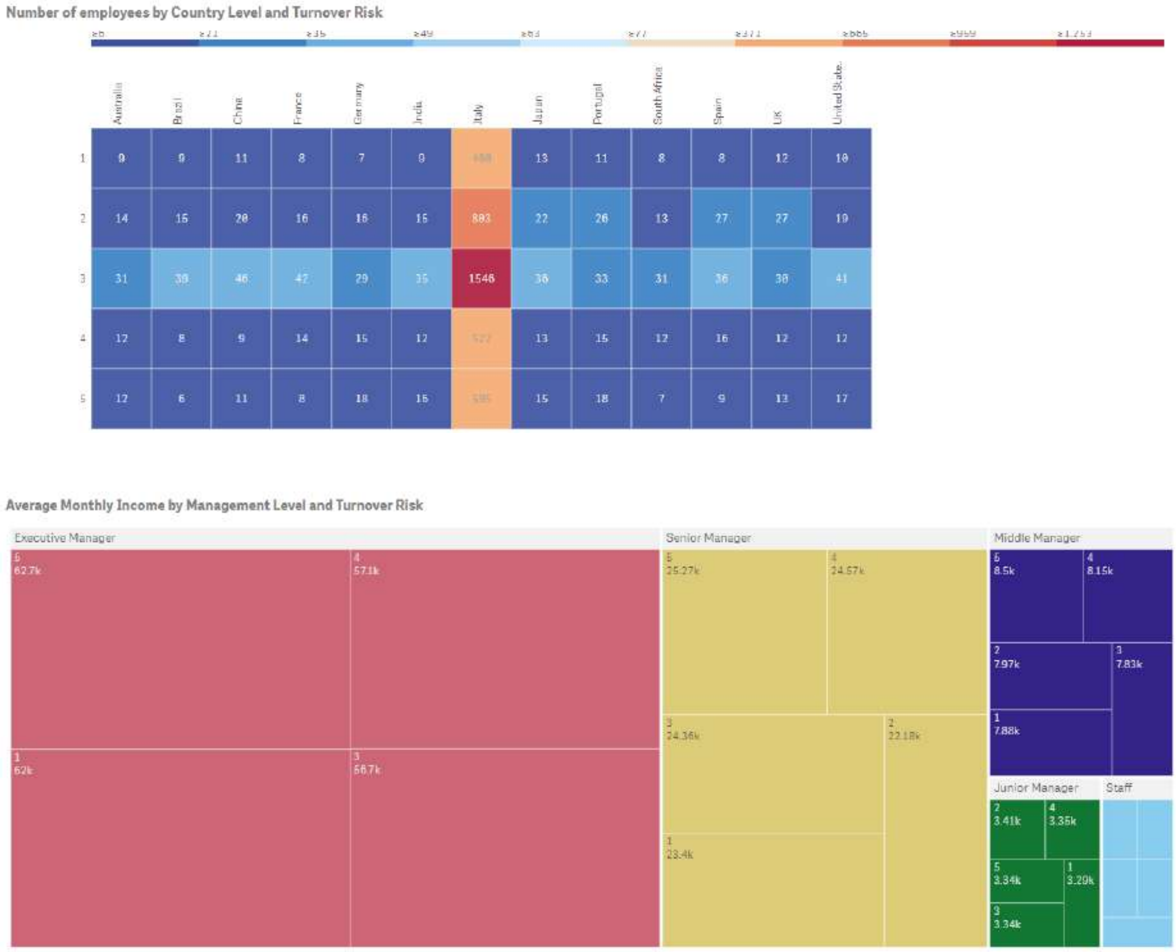
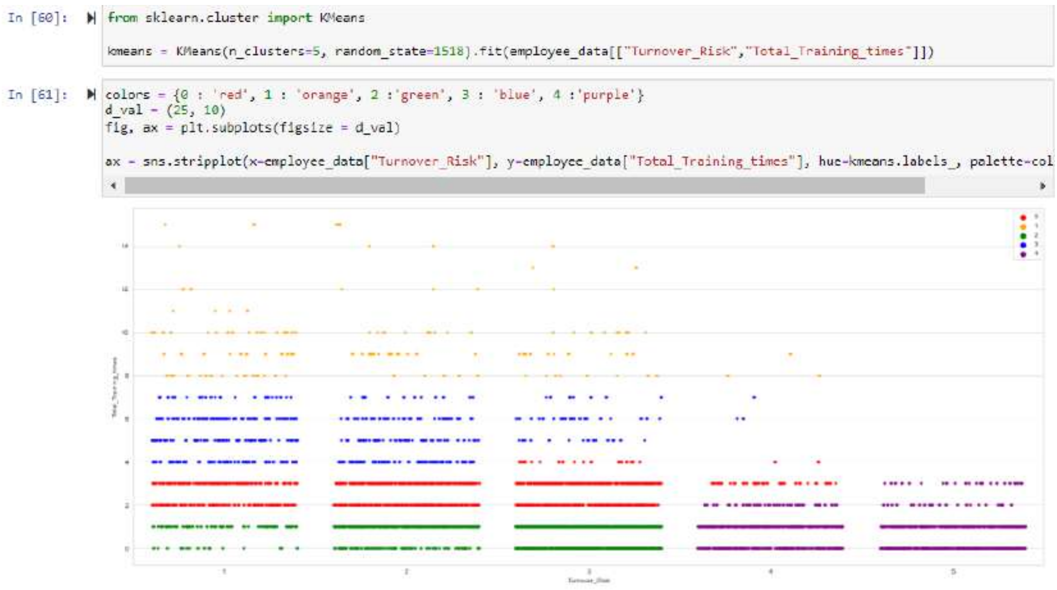


FIGURE 36: Turnover Risk by country heat map and Average Income per role position mekko chart

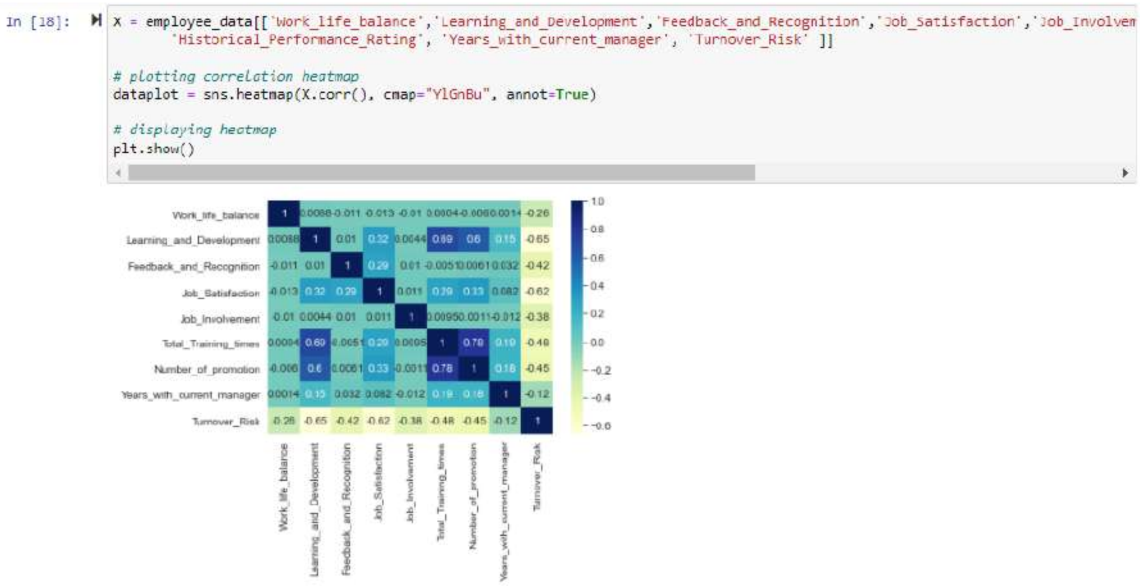
The last techniques applied within this descriptive part of Turnover analysis has been focused in getting more aware of the a relevant employee clusters trend to ascertain if potential biases regarding HR decisions and processes were occuring. Implementing the K-means unsupervised clustering illustrated in figure 37, for instance, HR decision makers are given the chance to give a closer analytical look to the distribution of the turnover risk levels across employees according to the number of training that have been offering them since they were employed by the recruiters. As expected, from this graph it could be straight inferred that employees having a likely higher risk of leaving the company nowadays are the ones which were given significantly less training times to update their competencies or develop new specialized knowledge. The algorithm, classifying the total workforce into five clusters based on the variance found across these two employee features, signals to the C-suite level on which exact group of employees HR departments should promptly intervene to prevent them from leaving. The purple cluster, that doesn't include only workers perceiving the worst rate of turnover risk, should be considered a priority to be taken care of if the company would not incur costly consequences.



```
In [60]:  from sklearn.cluster import KMeans

          kmeans = KMeans(n_clusters=5, random_state=1518).fit(employee_data[["Turnover_Risk","Total_Training_times"]])

In [61]:  colors = {0 : 'red', 1 : 'orange', 2 :'green', 3 : 'blue', 4 :'purple'}
          d_val = (25, 10)
          fig, ax = plt.subplots(figsize = d_val)

          ax = sns.stripplot(x=employee_data["Turnover_Risk"], y=employee_data["Total_Training_times"], hue=kmeans.labels_, palette=col
```

FIGURE 37: K-Means Turnover Risk clustering algorithm based on variables "Turnover Risk" and "Total Training Times"

Only by scaling up to the second level of predictive analytics, successfully forecasting the turnover risk and radically identifying the personal workforce traits, work environment factors, and job attitudes affecting it, HR departments could address beforehand the need to hire replacements quickly, and make other adjustments to retain employees in key positions; basically, with the use of predictions, HR operates efficient turnover-controlling actions neither fragmented nor misguided as these could be with only the information derived by the descriptive reports. This work offers then a new analytics procedure to make companies realize the relevance of finding the key factors influencing employee turnover intention by showing the most accurate, among the many ones built, machine learning models. The first step to building these models has been to enable the process of Feature Selection from which to identify the variables most correlated linearly with "Turnover Risk" to use in the machine-learning models to make the final prediction. In figure 38 I then performed a heat-map correlation analysis indicating the variables that, according to the methodology assumptions made when creating the simulated dataset, were the ones reported in the figure. As expected, all these employee features exhibited a pattern of negative linearity with my response variable of turnover.

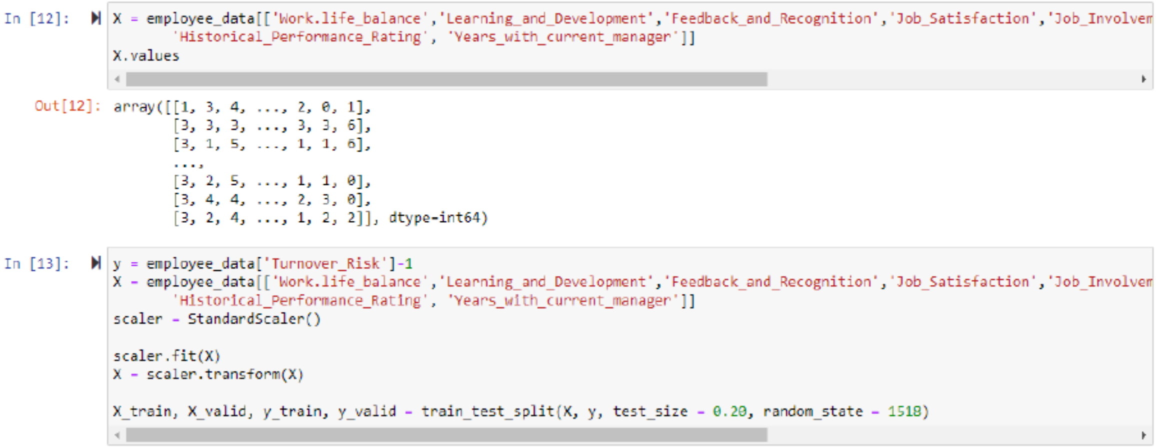


```
In [18]:  X = employee_data[['Work_life_balance','Learning_and_Development','Feedback_and_Recognition','Job_Satisfaction','Job_Involvem
                   'Historical_Performance_Rating', 'Years_with_current_manager', 'Turnover_Risk' ]]

          # plotting correlation heatmap
          dataplot = sns.heatmap(X.corr(), cmap="YlGnBu", annot=True)

          # displaying heatmap
          plt.show()
```

Work_life_balance 1 0.0088 -0.011 -0.013 -0.01 0.0004 -0.006 0.0014 -0.26
Learning_and_Development 0.0088 1 0.01 0.32 0.0044 0.69 0.6 0.15 -0.65
Feedback_and_Recognition -0.011 0.01 1 0.29 0.01 -0.0051 0.0061 0.032 -0.42
Job_Satisfaction -0.013 0.32 0.29 1 0.011 0.29 0.33 0.082 -0.62
Job_Involvement -0.01 0.0044 0.01 0.011 1 0.0095 0.0011 -0.012 -0.38
Total_Training_times 0.0004 0.69 -0.0051 0.29 0.0095 1 0.78 0.19 -0.48
Number_of_promotion -0.006 0.6 0.0061 0.33 -0.0011 0.78 1 0.16 -0.45
Years_with_current_manager 0.0014 0.15 0.032 0.082 -0.012 0.19 0.16 1 -0.12
Turnover_Risk -0.26 -0.65 -0.42 -0.62 -0.38 -0.48 -0.45 -0.12 1
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
Work_life_balance
Learning_and_Development
Feedback_and_Recognition
Job_Satisfaction
Job_Involvement
Total_Training_times
Number_of_promotion
Years_with_current_manager
Turnover_Risk

FIGURE 38: Heatmap Correlation Analyses for "Turnover Risk"

In particular, with a score of -0.65, the "Learning and Development" variable turned out to be the one with the strongest correlation pattern to turnover among the entire set of variables and then could be one of the most critical factors that affect turnover inside the firm. Having done this research, the following effort is to prepare the train and test sets including only the samples of the most correlated variables just selected to fit the machine-learning models. This passage belonging to the phase of data preparation is illustrated in figure 39, where the train test has been created including the following employee attributes: “Work-Life Balance”, “Feedback and Recognition”, “Job Satisfaction”, “Job Involvement”, “Total Training Time”, “Years with Current Managers”, “Number of Promotion” and, most of all, “Learning and Development”. On the other hand, the test set has been shaped in an identical way but the response variable “Turnover Risk”.

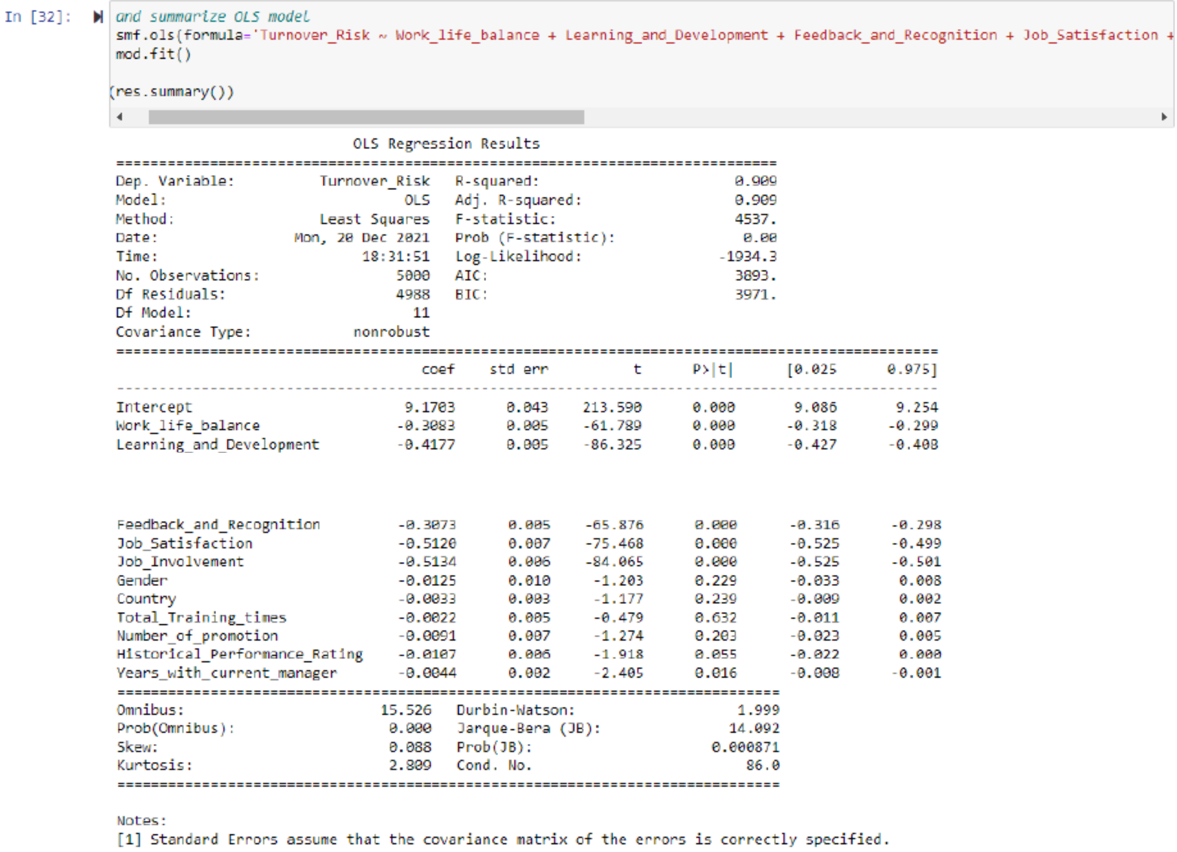


```
In [12]:  X = employee_data[['Work.life_balance','Learning_and_Development','Feedback_and_Recognition','Job_Satisfaction','Job_Involvem
                 'Historical_Performance_Rating', 'Years_with_current_manager']]
          X.values

Out[12]: array([[1, 3, 4, ..., 2, 0, 1],
                [3, 3, 3, ..., 3, 3, 6],
                [3, 1, 5, ..., 1, 1, 6],
                ...,
                [3, 2, 5, ..., 1, 1, 0],
                [3, 4, 4, ..., 2, 3, 0],
                [3, 2, 4, ..., 1, 2, 2]], dtype=int64)

In [13]:  y = employee_data['Turnover_Risk']-1
          X = employee_data[['Work.life_balance','Learning_and_Development','Feedback_and_Recognition','Job_Satisfaction','Job_Involvem
                 'Historical_Performance_Rating', 'Years_with_current_manager']]
          scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

FIGURE 39: Data Preparation with train and test split for predicting “Turnover Risk”

Though this first classification model hasn’t been trained with any data, this logistic regression, thanks to which we could realize which are the most critical factors that cause workers to have a high risk for voluntary departure, has been found to be really reliable, returning an R-Squared grade of 0.909 which indicates the features given well explain the dependent variable. In this OLS model pictured in figure 40, which accounts for variance between the dependent variable “Turnover Risk” and some independent features given as inputs, the ones analyzed earlier in the Feature Selection, the regressors are considered single integer values. The other proposed logistic regression model with every variable analyzed in its corresponding labels’ relationship with turnover, retuned me a lower total measure of model goodness, as the labels had many p-values beyond the statistical threshold of 0.05, so was left aside, like some other machine learning models whose accuracy score was not that high. In this model, in addition, all the features but “Gender”, “Country”, “Total Training Times” and “Number of

Promotions", recognizing the limitations of the analysis, have been deemed statistically significant in terms of p-value and then strongly impacting the generation of the turnover risk level, with most of them having a p-value of 0.00 that prove them to be totally explicable of turnover scores.

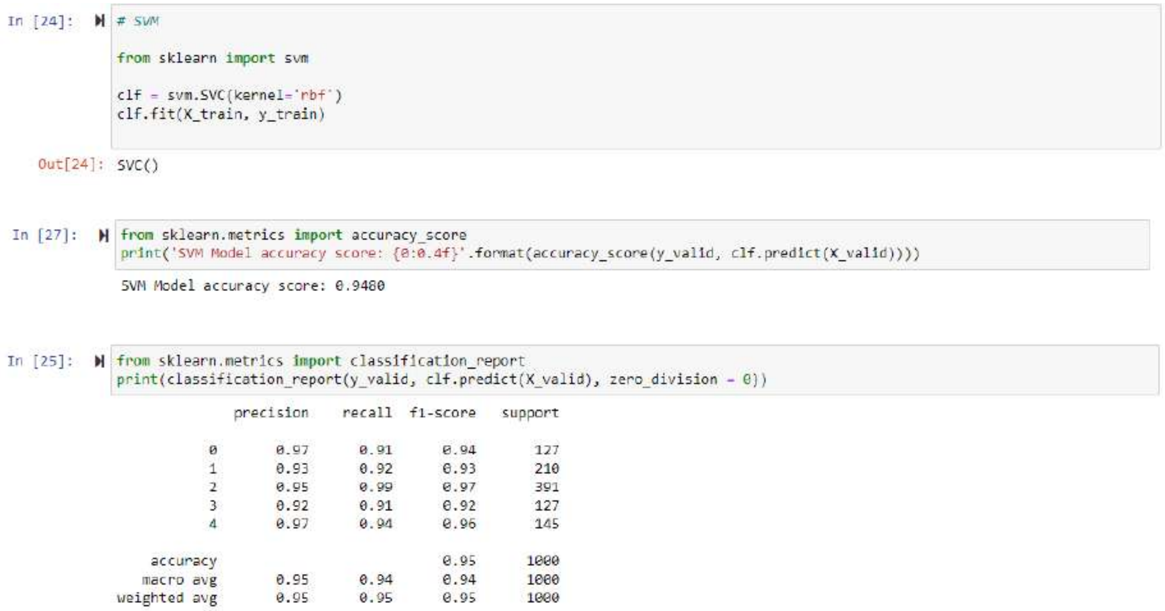


```
In [32]:  and summarize OLS model
          smf.ols(formula='Turnover_Risk ~ Work_life_balance + Learning_and_Development + Feedback_and_Recognition + Job_Satisfaction +
          mod.fit()

         (res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Turnover_Risk   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.909
Method:                 Least Squares   F-statistic:                     4537.
Date:                Mon, 20 Dec 2021   Prob (F-statistic):               0.00
Time:                        18:31:51   Log-Likelihood:                -1934.3
No. Observations:                5000   AIC:                             3893.
Df Residuals:                    4988   BIC:                             3971.
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                        9.1703      0.043    213.590      0.000       9.086       9.254
Work_life_balance               -0.3083      0.005    -61.789      0.000      -0.318      -0.299
Learning_and_Development        -0.4177      0.005    -86.325      0.000      -0.427      -0.408


Feedback_and_Recognition        -0.3073      0.005    -65.876      0.000      -0.316      -0.298
Job_Satisfaction                -0.5120      0.007    -75.468      0.000      -0.525      -0.499
Job_Involvement                 -0.5134      0.006    -84.065      0.000      -0.525      -0.501
Gender                          -0.0125      0.010     -1.203      0.229      -0.033       0.008
Country                         -0.0033      0.003     -1.177      0.239      -0.009       0.002
Total_Training_times            -0.0022      0.005     -0.479      0.632      -0.011       0.007
Number_of_promotion             -0.0091      0.007     -1.274      0.203      -0.023       0.005
Historical_Performance_Rating   -0.0107      0.006     -1.918      0.055      -0.022       0.000
Years_with_current_manager      -0.0044      0.002     -2.405      0.016      -0.008      -0.001
==============================================================================
Omnibus:                       15.526   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               14.092
Skew:                           0.088   Prob(JB):                     0.000871
Kurtosis:                       2.809   Cond. No.                         86.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

FIGURE 40: OLS Logistic Regression model to predict "Turnover Risk"

As already explained, since some other algorithms' accuracy scores were not strong enough, regarding this turnover study I decided to not report all the models performed. Still, to make real companies aware of the spectrum of predictive possibilities to forecast employee turnover, I will introduce right in the figures below, a machine-learning model that ultimately returned 100% level of accuracy. In figure 41, a basic SVM algorithm, following the same procedure didi before for the previous engagement analysis, and taking into account the limitations of the model by which the hyperplanes can not be graphically defined, has been fitted with my just-built train dataset. This algorithm already has proven to be very reliable, giving an accuracy score of 0,948, accompanied by the usual confusion matrix to underline for which "Turnover Risk" classes the few errors have been committed. Plus, the classification report emphasized other outstanding figures referring to the same model's results.



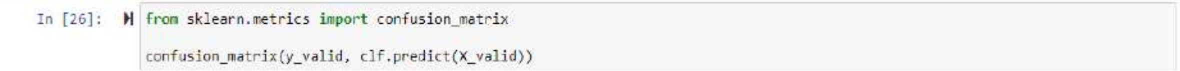
```
In [24]:  # SVM

          from sklearn import svm

          clf = svm.SVC(kernel='rbf')
          clf.fit(X_train, y_train)

Out[24]:  SVC()
```

```
In [27]:  from sklearn.metrics import accuracy_score
          print('SVM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          SVM Model accuracy score: 0.9480
```

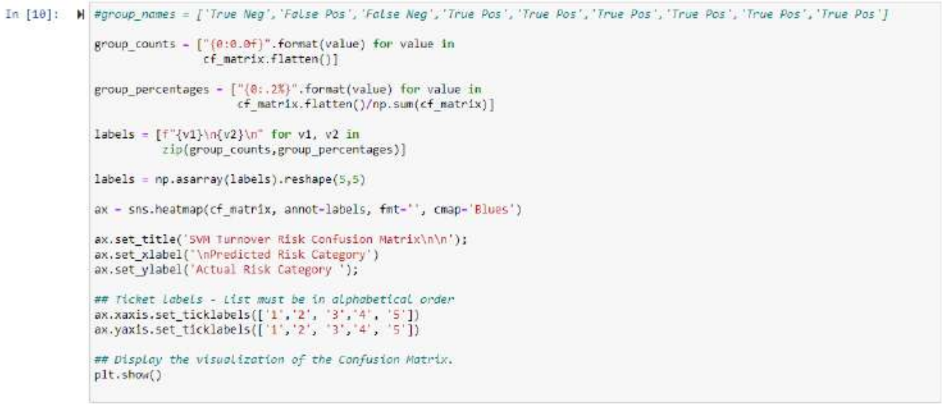
```
In [25]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       0.97      0.91      0.94       127
                     1       0.93      0.92      0.93       210
                     2       0.95      0.99      0.97       391
                     3       0.92      0.91      0.92       127
                     4       0.97      0.94      0.96       145

              accuracy                           0.95      1000
             macro avg       0.95      0.94      0.94      1000
          weighted avg       0.95      0.95      0.95      1000
```

FIGURE 41: Basic SVM classifier model fit to predict "Turnover Risk" and relative classification report and accuracy score

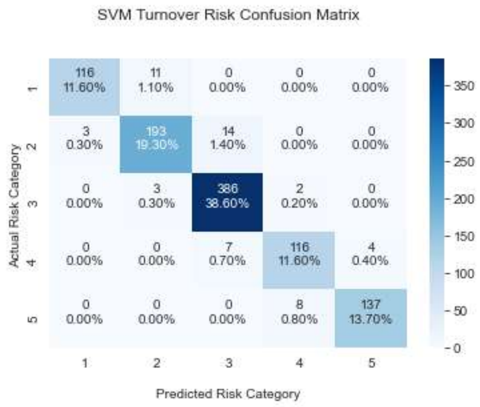
In the following figure 42 has been created a confusion matrix that shows to the readers where the SVM algorithm has failed to guess the turnover risk's label in the prediction.



```
In [26]:  from sklearn.metrics import confusion_matrix

          confusion_matrix(y_valid, clf.predict(X_valid))
```



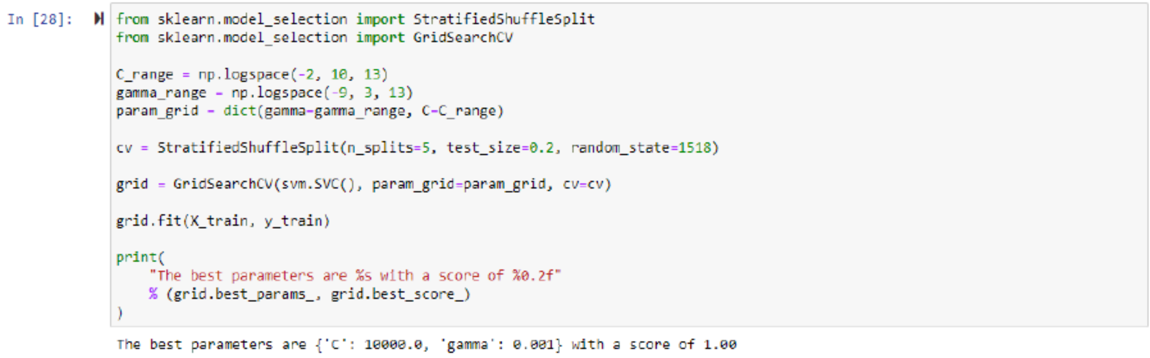
```
In [10]:  #group_names = ['True Neg','False Pos','False Neg','True Pos','True Pos','True Pos','True Pos','True Pos','True Pos']

          group_counts = ["{0:0.0f}".format(value) for value in
                          cf_matrix.flatten()]

          group_percentages = ["{0:.2%}".format(value) for value in
                               cf_matrix.flatten()/np.sum(cf_matrix)]

          labels = [f"{v1}\n{v2}\n" for v1, v2 in
                    zip(group_counts,group_percentages)]

          labels = np.asarray(labels).reshape(5,5)

          ax = sns.heatmap(cf_matrix, annot=labels, fmt='', cmap='Blues')

          ax.set_title('SVM Turnover Risk Confusion Matrix\n\n');
          ax.set_xlabel('\nPredicted Risk Category')
          ax.set_ylabel('Actual Risk Category ');

          ## Ticket labels - List must be in alphabetical order
          ax.xaxis.set_ticklabels(['1','2', '3','4', '5'])
          ax.yaxis.set_ticklabels(['1','2', '3','4', '5'])

          ## Display the visualization of the Confusion Matrix.
          plt.show()
```



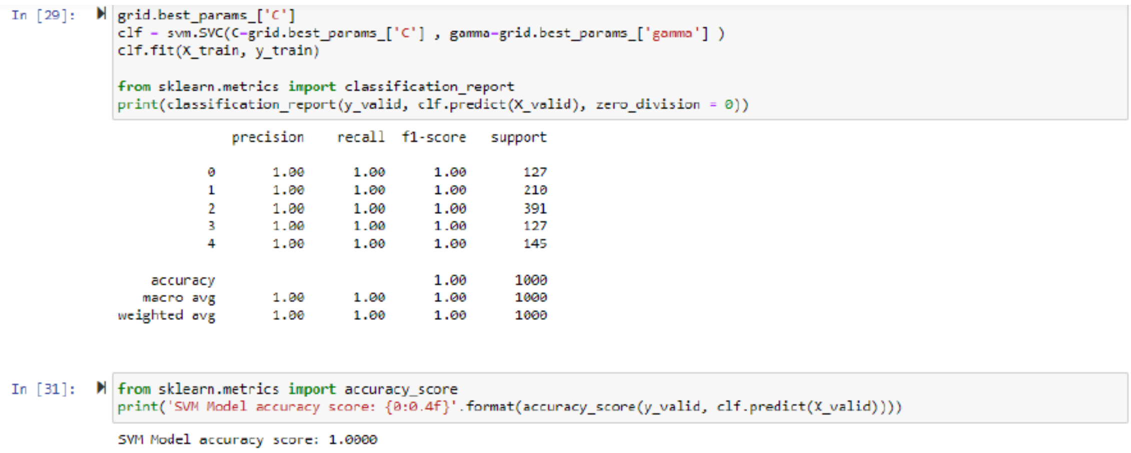
	1	2	3	4	5
1	116 11.60%	11 1.10%	0 0.00%	0 0.00%	0 0.00%
2	3 0.30%	193 19.30%	14 1.40%	0 0.00%	0 0.00%
3	0 0.00%	3 0.30%	386 38.60%	2 0.20%	0 0.00%
4	0 0.00%	0 0.00%	7 0.70%	116 11.60%	4 0.40%
5	0 0.00%	0 0.00%	0 0.00%	8 0.80%	137 13.70%

FIGURE 42: Confusion matrix for the SVM model predicting "Turnover Risk"

However, trying to even improve this already greatly performing model, has been carried out an identical process based on conducting a deeply explorative "Grid Search", with the hope to find the set of best hyper-parameters for this very model at hand. This analytical effort, shown in figure 43, has been proven to be quite efficient in bringing my SVM model to the perfection of ability to predict the turnover level on the test employee datasets, which the "Stratified Shuffle Split" resampling method has been applied on. Finally, modifying the values of the SVM's parameters, "C" and "gamma", with the ones returned by the Grid Search, the entire set of measures inside the new classification report, as well as the accuracy score, has improved to 1, meaning that this Support Vector Machine is now absolutely able, with no errors at all, to predict the Turnover Risk level of my workforce.



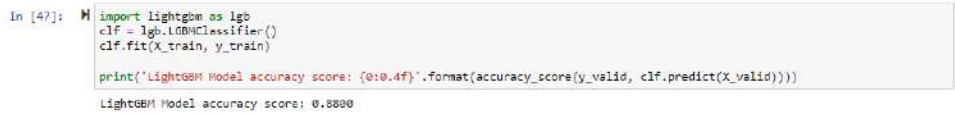
```
In [28]:  from sklearn.model_selection import StratifiedShuffleSplit
          from sklearn.model_selection import GridSearchCV

          C_range = np.logspace(-2, 10, 13)
          gamma_range = np.logspace(-9, 3, 13)
          param_grid = dict(gamma=gamma_range, C=C_range)

          cv = StratifiedShuffleSplit(n_splits=5, test_size=0.2, random_state=1518)

          grid = GridSearchCV(svm.SVC(), param_grid=param_grid, cv=cv)

          grid.fit(X_train, y_train)

          print(
              "The best parameters are %s with a score of %0.2f"
              % (grid.best_params_, grid.best_score_)
          )

          The best parameters are {'C': 10000.0, 'gamma': 0.001} with a score of 1.00
```



```
In [29]:  grid.best_params_['C']
          clf = svm.SVC(C=grid.best_params_['C'] , gamma=grid.best_params_['gamma'] )
          clf.fit(X_train, y_train)

          from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       1.00      1.00      1.00       127
                     1       1.00      1.00      1.00       210
                     2       1.00      1.00      1.00       391
                     3       1.00      1.00      1.00       127
                     4       1.00      1.00      1.00       145

              accuracy                           1.00      1000
             macro avg       1.00      1.00      1.00      1000
          weighted avg       1.00      1.00      1.00      1000


In [31]:  from sklearn.metrics import accuracy_score
          print('SVM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          SVM Model accuracy score: 1.0000
```

FIGURE 43: Grid Search to improve the SVM classifier and relative new classification report and accuracy score

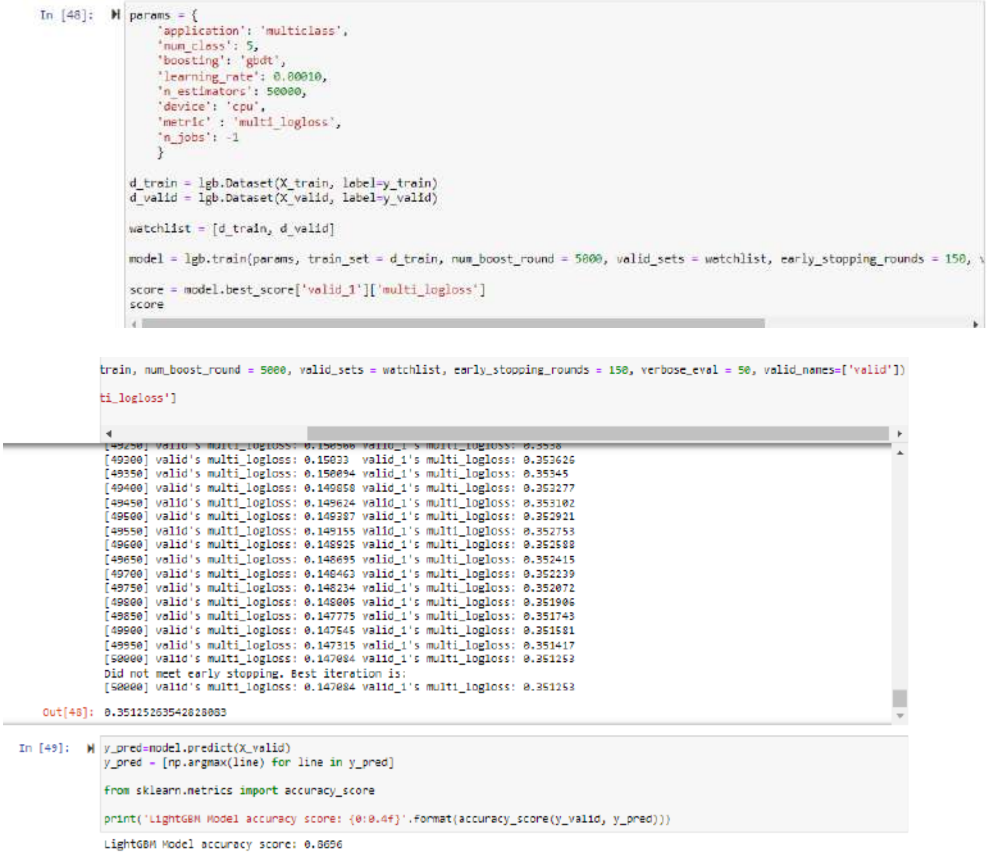
Ultimately, the next model run was a Light Gradient Boosting, called "LGBM", that is a gradient boosting framework that uses tree based learning algorithms that tries to understand, on the basis on some internal parameters, adds a step to the random forest implementation, whether splitting features to right or left. Differently from random forests that utilize pre-sort-based algorithms, this model uses histogram-based algorithms to bucket continuous features values into discrete bins, speeding up learning and lowering memory usage. Once my data have been bucketed, I should get how many samples have been categorized correctly in each bucket and which not. With the gradient boosting component, in this process, for each successful categorization a reward is given while a penalty for the wrong ones, each one very small and based on a learning rate. Applying this penalty and reward scheme at the next iteration of the model, there will be an iteration where no penalties will be given. My model has then been trained on my data as shown in figure 44: the accuracy function returns me a score of 0,88 which is very good, giving the limitations of the model.



```
In [47]:  import lightgbm as lgb
          clf = lgb.LGBMClassifier()
          clf.fit(X_train, y_train)

          print('LightGBM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          LightGBM Model accuracy score: 0.8800
```

Figure 44: LGBM predicting "Turnover Risk"

In attempting to improve the precision of the algorithm, I subsequently match specific values for some parameters as could be suggested in Figure 45. In particular, I ordered the system to associate a 'learning rate', which is the value of each reward and penalty, of 0.0001 and 5000 as 'n_estimators', which indicates the total number of trees. Moreover, with 'metrics', I specify that when a single iteration is completed, the algorithm has to calculate the loss measured on 'multi_logloss' as I was dealing with a multi classification. In addition, setting a 'number_o_boost_rond', I imposed the number of iteration of the algorithm, with 'verbose_eval' of 50, I say to the system to write down a row showing the outcomes every 50 iterations, while with 'early_stipping_rounds' of 150, I let the algorithm stop running if after 150 iterations there was no improvement yet, just taking the best value emerged so far.



```
In [48]:  params = {
              'application': 'multiclass',
              'num_class': 5,
              'boosting': 'gbdt',
              'learning_rate': 0.00010,
              'n_estimators': 50000,
              'device': 'cpu',
              'metric' : 'multi_logloss',
              'n_jobs': -1
              }

          d_train = lgb.Dataset(X_train, label=y_train)
          d_valid = lgb.Dataset(X_valid, label=y_valid)

          watchlist = [d_train, d_valid]

          model = lgb.train(params, train_set = d_train, num_boost_round = 5000, valid_sets = watchlist, early_stopping_rounds = 150, v

          score = model.best_score['valid_1']['multi_logloss']
          score
```

```
train, num_boost_round = 5000, valid_sets = watchlist, early_stopping_rounds = 150, verbose_eval = 50, valid_names=['valid'])

ti_logloss']
```

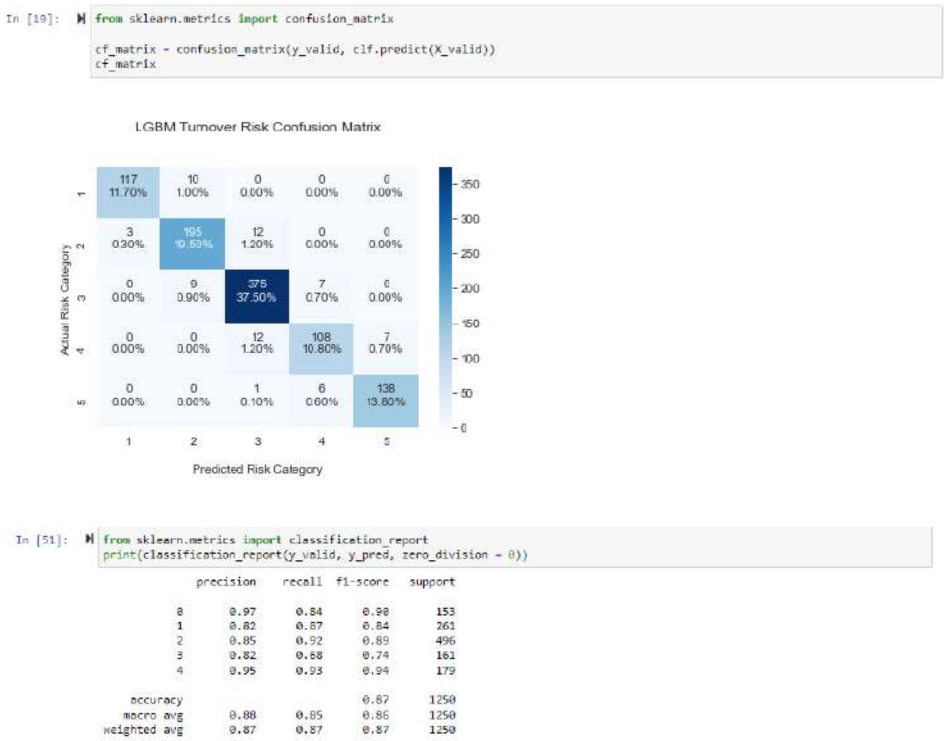
```
[49250] valid's multi_logloss: 0.150566 valid_1's multi_logloss: 0.3538
[49300] valid's multi_logloss: 0.15033  valid_1's multi_logloss: 0.353626
[49350] valid's multi_logloss: 0.150094 valid_1's multi_logloss: 0.35345
[49400] valid's multi_logloss: 0.149858 valid_1's multi_logloss: 0.353277
[49450] valid's multi_logloss: 0.149624 valid_1's multi_logloss: 0.353102
[49500] valid's multi_logloss: 0.149387 valid_1's multi_logloss: 0.352921
[49550] valid's multi_logloss: 0.149155 valid_1's multi_logloss: 0.352753
[49600] valid's multi_logloss: 0.148925 valid_1's multi_logloss: 0.352588
[49650] valid's multi_logloss: 0.148695 valid_1's multi_logloss: 0.352415
[49700] valid's multi_logloss: 0.148463 valid_1's multi_logloss: 0.352239
[49750] valid's multi_logloss: 0.148234 valid_1's multi_logloss: 0.352072
[49800] valid's multi_logloss: 0.148005 valid_1's multi_logloss: 0.351906
[49850] valid's multi_logloss: 0.147775 valid_1's multi_logloss: 0.351743
[49900] valid's multi_logloss: 0.147545 valid_1's multi_logloss: 0.351581
[49950] valid's multi_logloss: 0.147315 valid_1's multi_logloss: 0.351417
[50000] valid's multi_logloss: 0.147084 valid_1's multi_logloss: 0.351253
Did not meet early stopping. Best iteration is:
[50000] valid's multi_logloss: 0.147084 valid_1's multi_logloss: 0.351253

Out[48]: 0.35125263542828083
```

```
In [49]:  y_pred=model.predict(X_valid)
          y_pred = [np.argmax(line) for line in y_pred]

          from sklearn.metrics import accuracy_score

          print('LightGBM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, y_pred)))

          LightGBM Model accuracy score: 0.8696
```

FIGURE 45: LGBM with some specific parameters; "Turnover Risk" prediction. Other values couldn't be contained in this figure for space issues.

After performing the new accuracy score on the validation set,, with relative confusion matrix and classification report illustrated in Figure 46, this time it could be at first sight striking the the model didn't raise its performance, returning a score slightly reduced of

0,869: only way to do that could have been by implementing a more complicated and prepared hyperparameters search which had the system running for quite a long time.

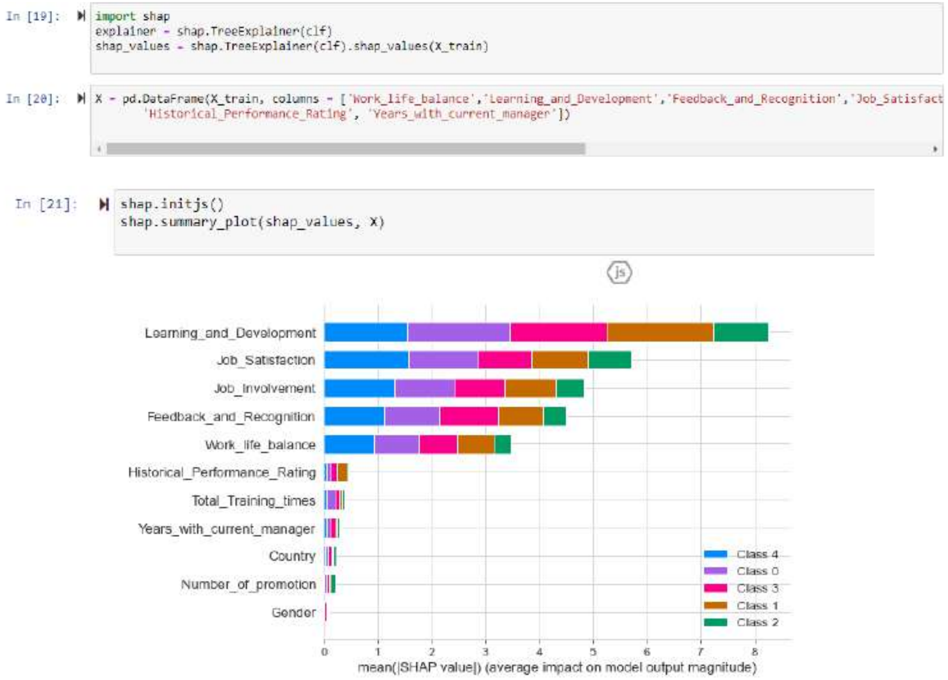


```
In [19]:  from sklearn.metrics import confusion_matrix

           cf_matrix = confusion_matrix(y_valid, clf.predict(X_valid))
           cf_matrix
```

```
In [51]:  from sklearn.metrics import classification_report
           print(classification_report(y_valid, y_pred, zero_division = 0))

              precision    recall  f1-score   support

           0       0.97      0.84      0.90       153
           1       0.82      0.87      0.84       261
           2       0.85      0.92      0.89       496
           3       0.82      0.68      0.74       161
           4       0.95      0.93      0.94       179

    accuracy                           0.87      1250
   macro avg       0.88      0.85      0.86      1250
weighted avg       0.87      0.87      0.87      1250
```

FIGURE 46: Confusion matrix and Classification Report of the LGBM Model predicting "Turnover Risk"

Ultimately, a Feature Importance process has been carried out for this model using the 'Shap' (Shapley Additive explanations) values method to understand the relative single contribution of each variable to the "Turnover Risk" prediction performed by the model. In figure 47, then, the set of features used in the model have been ordered based on how much they have impacted the final prediction: the x-axis plots the average of the absolute SHAP value of each feature, divided by class labels, while y-axis lists the name of the features themselves. In this model, HR managers could grasp valuable insights by ascertaining that the employee factor "Learning&Development" corresponds to the most important in the prediction of the turnover risk levels of the workforce, while attributes regarding promotion, training, country of origin and performance, yet included in the prediction computations, have contributed for the least part compared to the first fout for order of importance.

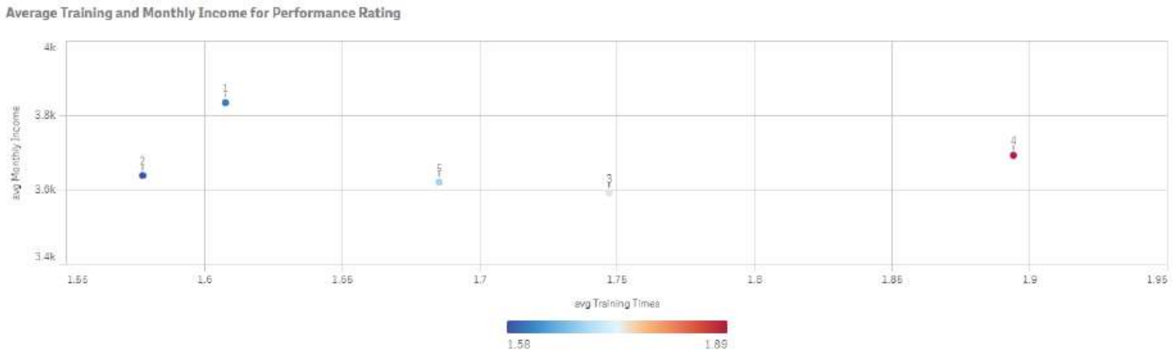


```
In [19]:  import shap
          explainer = shap.TreeExplainer(clf)
          shap_values = shap.TreeExplainer(clf).shap_values(X_train)

In [20]:  X = pd.DataFrame(X_train, columns = ['Work_life_balance','Learning_and_Development','Feedback_and_Recognition','Job_Satisfact
                   'Historical_Performance_Rating', 'Years_with_current_manager'])

In [21]:  shap.initjs()
          shap.summary_plot(shap_values, X)
```

FIGURE 47: Shap Values' Feature Importance for the LGBM Model predicting "Turnover Risk"

3.3.3 Performance Appraisal Analysis

The third area of analysis in which I've drawn my attention to project a set of empirical HR analyses utilizing my simulated dataset regards performance management. Primarily, I should remind that for this crucial employee indicator, the labels don't vary within the continuous numerical range 1 to 5 like in the past two analysis' examples, but, in ordinal sequence, can assume the following values: "Very Low", "Low", "Medium", "High" and "Excellent". Moving to the fact-based results, the first dashboard outlined in figure 48 is made of two interesting graphs. The first one is a scatter plot mapping the five performance ratings of the workforce with reference to their average salary earnings and the average number of training released for each performance category. Although the differences consist of a few decimal points, most probably because of the great totality observations taken into consideration, a crystal-clear pattern emerges in the figure: people inside the firm that over the last couple of years have performed at a "High" level coincide to employees that took part at the averagely highest number of training sessions. Training seems to be then a good practice to incentivize workers to outperform their normal level of productivity, efficiency and team collaboration, some of the many sub-indicators impacting on overall individual performance. In this direction, the opposite phenomenon occurs for the worst-performing employees of the first two levels, which have been offered fewer chances to get some training. Still, it's important to denote a bias on which HR top managers should get quite alarmed: these latter low performing employees, at the same time, are receiving averagely better pay than all the other performance employee clusters. This means that a great deal of payroll money spent by the company is not being capitalized on adequate performance by the workforce. The other line chart in the dashboard, on the other hand, shows the shape of the two gender groups' curve based on the average working hours per week and on their performance score. Some considerations could be easily grasped: male part of the labor force seems to work more, on average two hours more, than the female part for each performance category. This could be the direct cause of the great number of females among the part-time contracts signed by the enterprise. What's more, both curves are slightly raising from performance score "Very Low" to "Excellent", presupposing that the best performers, from this dashboard, correspond to the workers averagely working more hours per week.



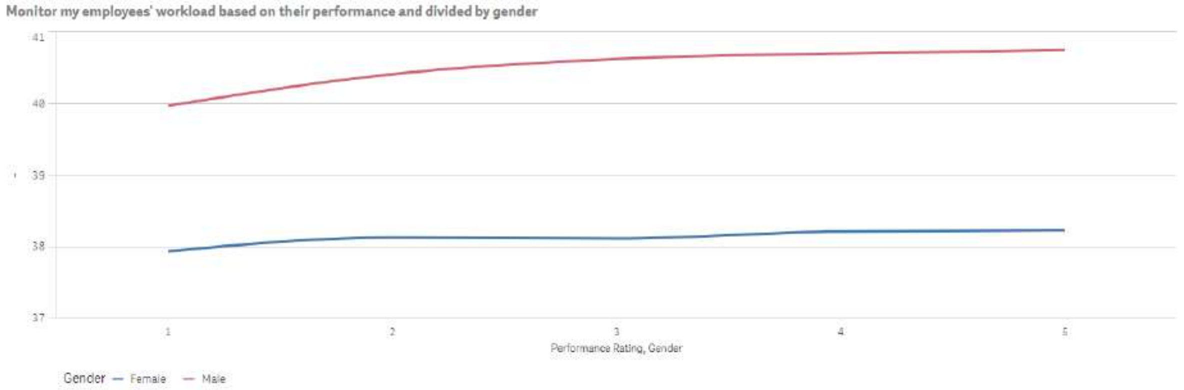
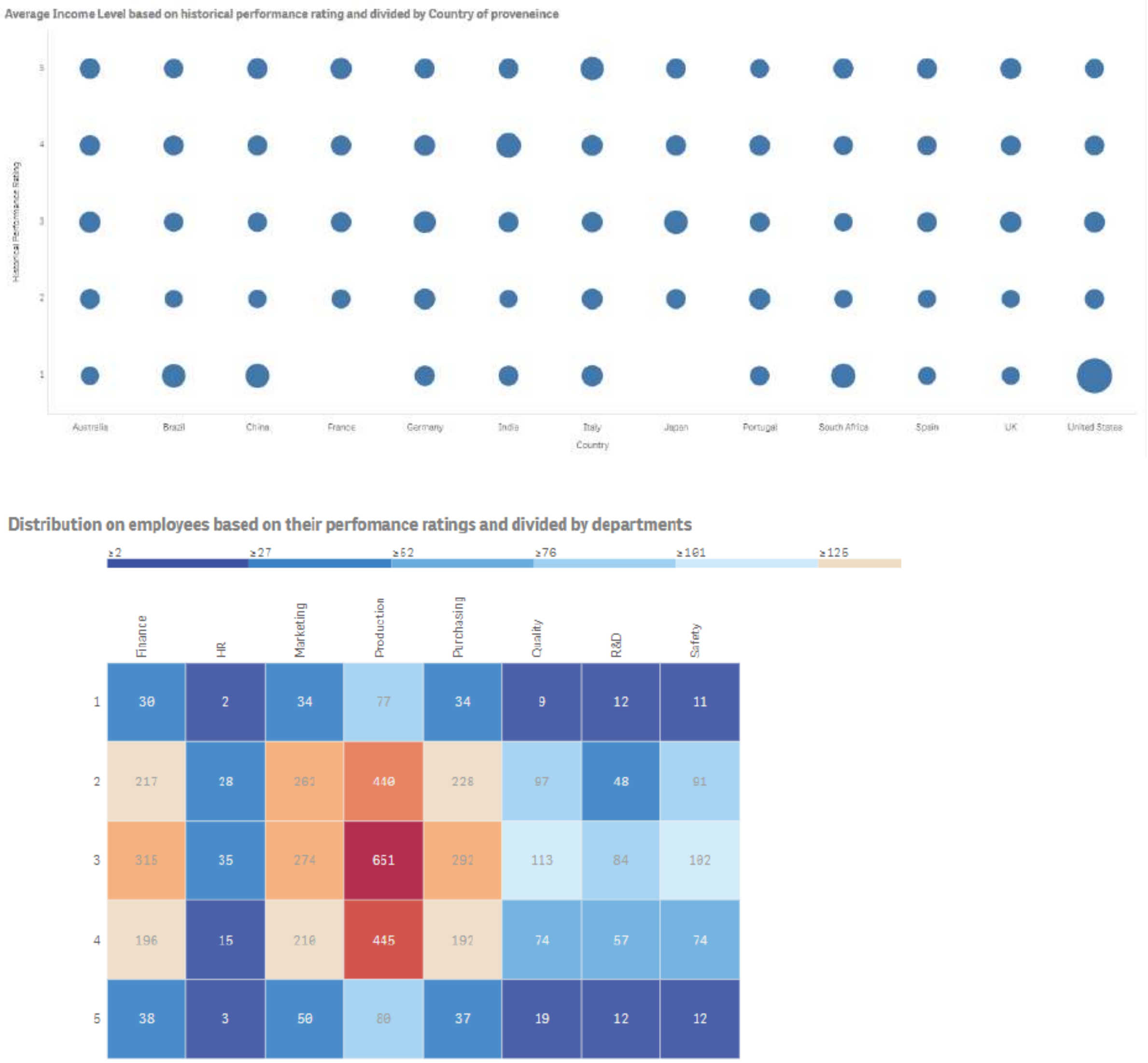


FIGURE 48: Average training and income by performance scatter plot and employees workloads by gender line chart

On this next dashboard represented in figure 49, a blue-colored grid search is presented to carry out a diversity analysis emphasizing, for each performance rating of employees, the difference in average income for clusters of nationalities. Therefore, in this representation, the bigger will be the points, the bigger will be the average salary level for that determinate ethnic group of employees. If everything worked smoothly inside the firm's walls, then people expected to be paid more for their commitment and quality of work should be the best-performing workers, though within specific ranges dictated by managerial positions. However, these descriptive outcomes are negatively striking as, looking at the three countries Brazil, China and above all United States, a dangerous payroll bias shows up, making managers realize that, in these countries, a number of employees, really underperforming the adequate standards (as they belong to the slot "Very Low"), turn out to receive a proportional enormous amount of money. The HR department should more deeply analyze this situation, understanding not only the causes that led to these figures but also how to not repeat the occurring of a situation of this kind further on time. The second graph, corresponding to a heat-map, plots the distribution of employees based on their performance score and divided by departments. in this case, fortunately for C-suite executives, there are no critical numbers to stress out, except getting the negative understanding that for each department, there are actually more employees performing at the rating "Low" than employees whose performance was classified "High": this trend should be drastically inverted to positively influence the overall business output of the firm.



	Finance	HR	Marketing	Production	Purchasing	Quality	R&D	Safety
1	30	2	34	77	34	9	12	11
2	217	28	262	440	228	97	48	91
3	315	35	274	651	292	113	84	102
4	196	15	210	445	192	74	57	74
5	38	3	50	80	37	19	12	12

FIGURE 49: diversity-country analysis grid chart and departments distribution by performance level heat-map

The final comprehensive dashboard shows, through figure 50, on the top, a mekko chart plotting the average employees' tenure at the company divided into boxes according to the management level and the performance rating associated with each worker. Regarding the executive level, the most important for effective conduct of business's company, managers could worryingly notice how among the few in that position (a total of 7) there is no one as guiding examples for all the other under-level employees that are performing at best. Other numbers are pretty variable since the average level of years of service in the company for the performance employees' clusters do not follow a similar pattern across each managerial position isolated in this analysis. Interestingly, for what concerns the "Staff" positions, employees are all averagely at the company since far fewer years, and top performers in this context are even the ones arrived, averagely, the earlier: analysts should monitor their performance scoring to see whether they're really constant and could be considered as an investment-promised asset of the company for the future, or if they're doing good just because they are making such an effort to make a profitable first impression n their superiors. Below in the same dashboard, lastly, a bar chart analyzes the trends of performance levels based on the number of promotions reached by employees. What is particularly curious and relieving, is that, as the number of promotions grows, advanced employees are found to slightly improve their medium or good level of performance and, above all, significantly bring their mediocre levels up to a sufficient or good score. Having said that, therefore, for each employee who has never been promoted or has just one only time, the probabilities to lie in the segment of "Low" and "Very Low" performers are very likely and should be somehow adjusted by HR practitioners.

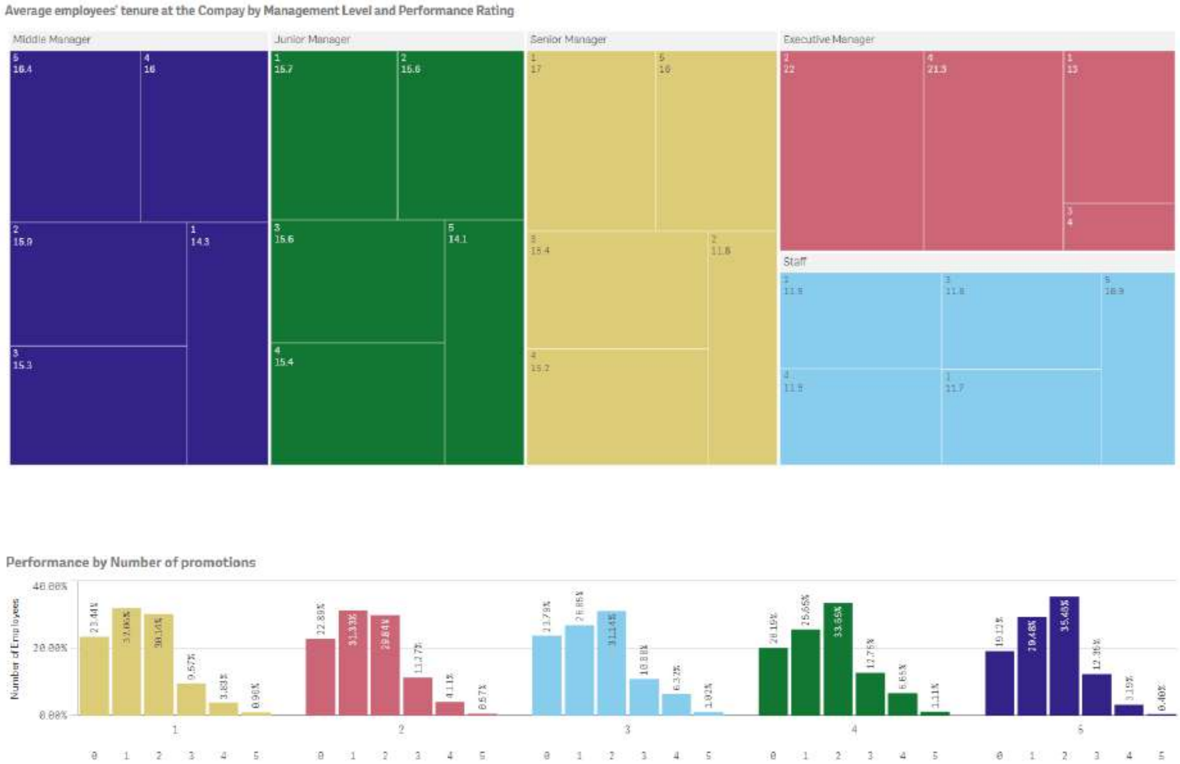
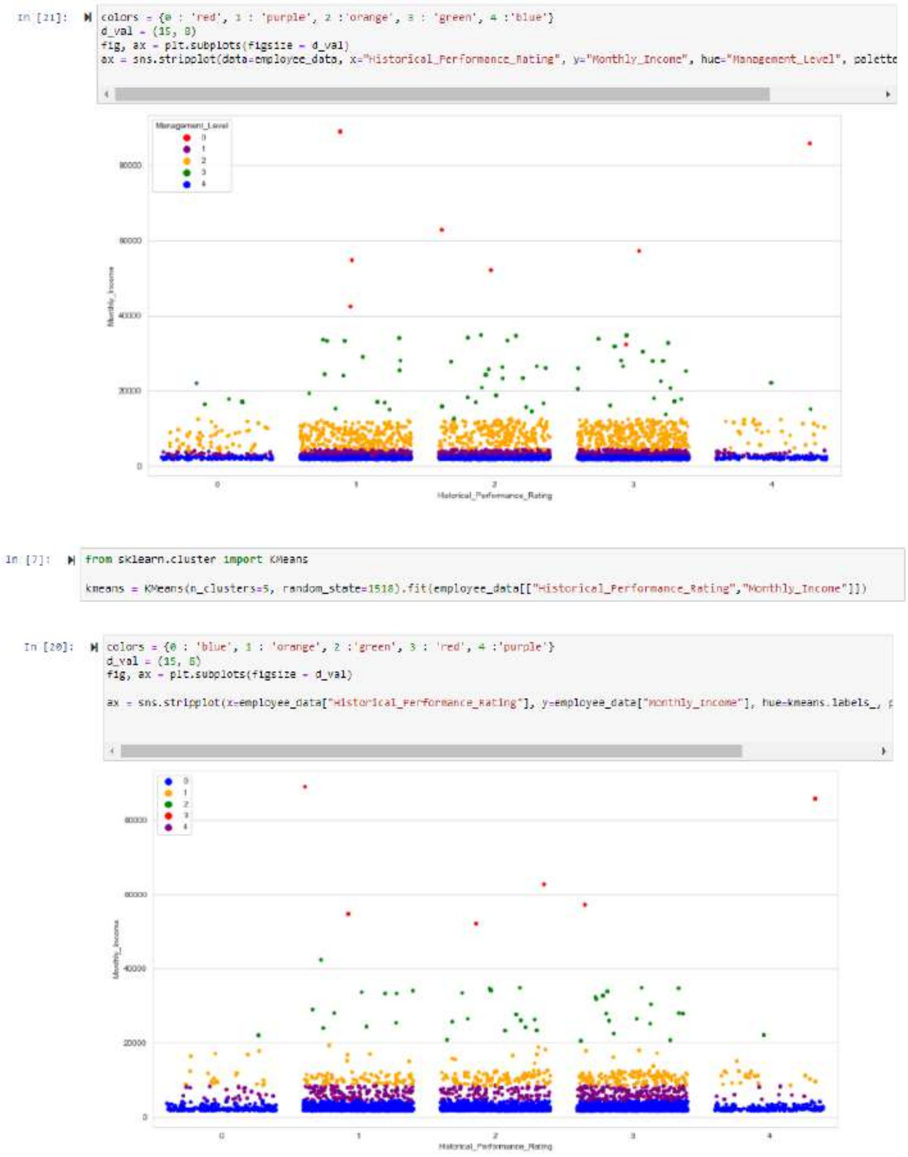


FIGURE 50: Employee average tenure mekko chart and performance by promotion bar charts

Lastly, a key cluster analysis and a clustering machine learning analysis have been deployed in figure 51: by plotting points in the graph, representing employees, according to their associated respective "Historical Performance Rating" score, on the x-axis, and "Monthly Income" level, on the y axis, a comparison between the two techniques could be made to understand were potential HR biases are happening in the company with respect to how managerial positions are distributed based on the relevant and influencing two employee factors. As a matter of fact, like for the analysis performed in the employees engagement section, the first classic cluster analysis just reflects the current situation of the managerial roles workforce's levels with reference to their past performance score and pay of each specific employee. This first graph must be then confronted by the HR analysts with the second one, that rather shows how the unsupervised K-means algorithm intelligently thought of a fairer classification of the five management levels, present in this company's organizational structure, and based on the variance and relationships among the plotted points. This comparison could be of great help to decision-making managers, who can identify some anomalous data pattern, evaluating, among the entire workforce, who should be declassed or upgraded, or who is in condition to receive an increase in pay because of the performance level shown. The K-means, in generally, as seen in the figure, would lead to a complete reorganization of the managerial structure inside the company, as, according to its classes separation, nowadays there are too many employees on the medium "Middle manager" level that should cover an inferior role such as "Junior managers"; moreover, we could notice that in the first world-picturing graphs the red points, corresponding to the top managerial position "Executive Managers" count eight points, while in the graph below, the red points are just sx, meaning that out of eight top managers, two are considered, according to a combination of their performance level and salary rate, to not belonging to the top managerial category. This analysis could be extended for each class, so as for managers to be enabled to take appropriate considerations and evaluate whether to change something.



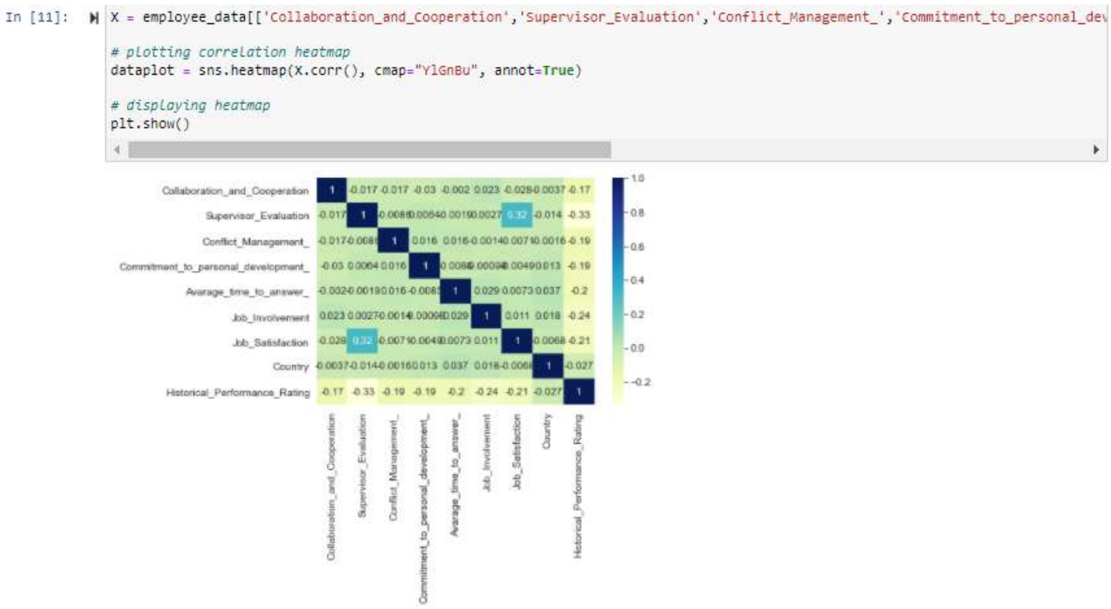
```
In [21]:  colors = {0 : 'red', 1 : 'purple', 2 :'orange', 3 : 'green', 4 :'blue'}
          d_val = (15, 8)
          fig, ax = plt.subplots(figsize = d_val)
          ax = sns.stripplot(data=employee_data, x="Historical_Performance_Rating", y="Monthly_Income", hue="Management_Level", palett
```

```
In [7]:  from sklearn.cluster import KMeans

         kmeans = KMeans(n_clusters=5, random_state=1518).fit(employee_data[["Historical_Performance_Rating","Monthly_Income"]])
```

```
In [20]:  colors = {0 : 'blue', 1 : 'orange', 2 :'green', 3 : 'red', 4 :'purple'}
          d_val = (15, 8)
          fig, ax = plt.subplots(figsize = d_val)

          ax = sns.stripplot(x=employee_data["Historical_Performance_Rating"], y=employee_data["Monthly_Income"], hue=kmeans.labels_, p
```

FIGURE 51: Real-world situation cluster analysis versus K-Means clustering classification for “Management Level” based on variables “Historical Performance Rating” and “Monthly Income”

The upcoming predictive findings will be key in making the company rethink of its managerial strategies and take actions to keep high performers inside the company’s walls, as well as proactively intervene on those employees factors that will finally be found to impact on the performance trend of this considered HR population. To start, in these HR analytics predictive modeling examples, numeric values were used in place of text categorical variable labels in such a way that the performance ratings, in increasing order, scaling up from ”Very Low” to “Excellent” rate as explained before, have been converted from 0 to 4. This allowed the algorithm to read the value and do the computations in order to make the prediction requested. Starting from the necessary phase of data pre-processing, also for this analysis I have applied a Person's Correlation, deployed using an easy-to-interpret heat-map in figure 52, to get aware of which features could be turned out to be more strongly linked, in terms of linear patterns, to the response variable that I'm going to predict, "Historical Performance Rating". At the end of this process, I will test the accuracy of my performance machine-learning models that will be consequently fitted with the following set of employee variables: "Collaboration&Cooperation", "Supervisor Evaluation", "Conflict Management", "Commitment to personal development", "Average time to answer", "Job Involvement" and "Job Satisfaction"; within this performance framework, feature selection values, in contrast to previous cases of analysis, have been curiously discovered to present quite a similar degree of negative linearity with "Historical Performance Rating", ranging between -0.17 and -0,24, with the only "Supervisor Evaluation" weighting more than other indicators (-0,33).

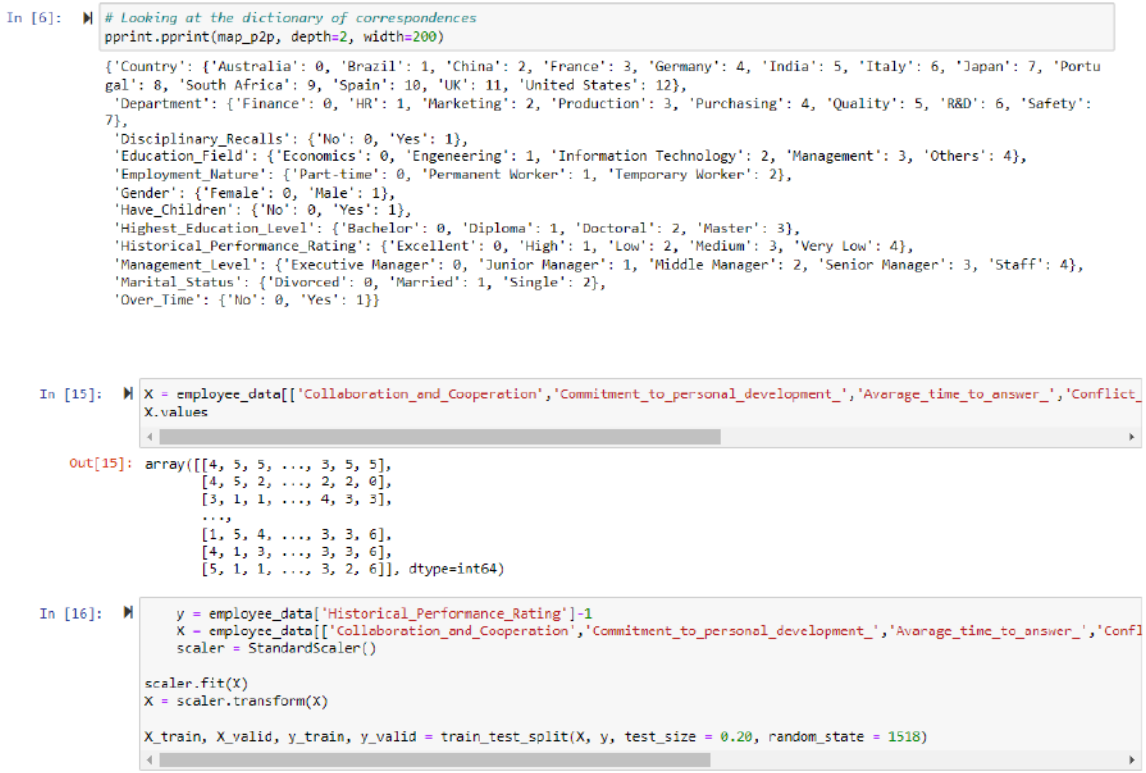


```
In [11]:  X = employee_data[['Collaboration_and_Cooperation','Supervisor_Evaluation','Conflict_Management_','Commitment_to_personal_dev

           # plotting correlation heatmap
           dataplot = sns.heatmap(X.corr(), cmap="YlGnBu", annot=True)

           # displaying heatmap
           plt.show()
```

Collaboration_and_Cooperation 1 -0.017 -0.017 -0.03 -0.002 0.023 -0.028 -0.0037 -0.17
Supervisor_Evaluation -0.017 1 -0.0068 0.0064 -0.0019 0.0027 0.32 -0.014 -0.33
Conflict_Management_ -0.017 -0.0068 1 0.016 0.016 -0.0014 -0.0071 -0.0016 -0.19
Commitment_to_personal_development_ -0.03 0.0064 0.016 1 -0.0085 -0.00098 -0.0049 0.013 -0.19
Avarage_time_to_answer_ -0.002 -0.0019 0.016 -0.0085 1 0.029 0.0073 0.037 -0.2
Job_Involvement 0.023 0.0027 -0.0014 -0.00098 0.029 1 0.011 0.018 -0.24
Job_Satisfaction -0.028 0.32 -0.0071 -0.0049 0.0073 0.011 1 -0.0068 -0.21
Country -0.0037 -0.014 -0.0016 0.013 0.037 0.018 -0.0068 1 -0.027
Historical_Performance_Rating -0.17 -0.33 -0.19 -0.19 -0.2 -0.24 -0.21 -0.027 1
Collaboration_and_Cooperation
Supervisor_Evaluation
Conflict_Management_
Commitment_to_personal_development_
Avarage_time_to_answer_
Job_Involvement
Job_Satisfaction
Country
Historical_Performance_Rating
1.0
0.8
0.6
0.4
0.2
0.0
-0.2

FIGURE 52: Heatmap Correlation Analyses for "Historical Performance Rating" for Feature Selection process

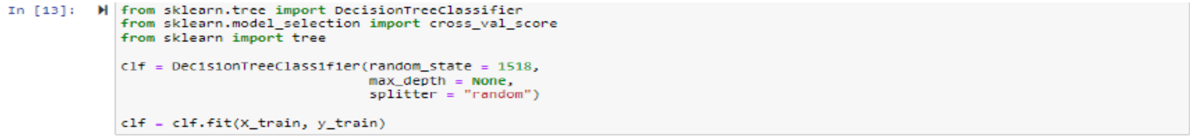
In the second part of following figure 53, I then proceed to divide my entire dataset in an 80% corresponding to the train set, which will contain all the independent variables identified not he the previous feature selection process, without including my response variable “Historical Performance Rating” which will constitute the remaining 20% pf the validation set on which the effectiveness of prediction of my machine-learning models will be ascertained after the learning phase. That being said, when I applied, as explained in the methodology, the ‘Label_encoder’ function to convert categorical variables, differently from the other turnover and job involvement and satisfaction dimensions, to “Historical Performance Rating” labels have been assigned numerical values; it’s then important to keep in mind, when interpreting the predictions’ results for this analyses, that the labels were inverted by order: as it is visible in the same superior part of the figure, “Excellent”, corresponding to the best ranking rate for performance, has been automatically associated to the value “0” and not “4”, that reciprocally will be referred in the subsequent machine computations with the tuple “Very Low”, the lowest for this ordinal variables.



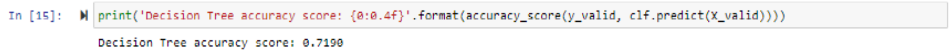
```
In [6]:  # Looking at the dictionary of correspondences
         pprint.pprint(map_p2p, depth=2, width=200)

         {'Country': {'Australia': 0, 'Brazil': 1, 'China': 2, 'France': 3, 'Germany': 4, 'India': 5, 'Italy': 6, 'Japan': 7, 'Portu
         gal': 8, 'South Africa': 9, 'Spain': 10, 'UK': 11, 'United States': 12},
          'Department': {'Finance': 0, 'HR': 1, 'Marketing': 2, 'Production': 3, 'Purchasing': 4, 'Quality': 5, 'R&D': 6, 'Safety':
         7},
          'Disciplinary_Recalls': {'No': 0, 'Yes': 1},
          'Education_Field': {'Economics': 0, 'Engeneering': 1, 'Information Technology': 2, 'Management': 3, 'Others': 4},
          'Employment_Nature': {'Part-time': 0, 'Permanent Worker': 1, 'Temporary Worker': 2},
          'Gender': {'Female': 0, 'Male': 1},
          'Have_Children': {'No': 0, 'Yes': 1},
          'Highest_Education_Level': {'Bachelor': 0, 'Diploma': 1, 'Doctoral': 2, 'Master': 3},
          'Historical_Performance_Rating': {'Excellent': 0, 'High': 1, 'Low': 2, 'Medium': 3, 'Very Low': 4},
          'Management_Level': {'Executive Manager': 0, 'Junior Manager': 1, 'Middle Manager': 2, 'Senior Manager': 3, 'Staff': 4},
          'Marital_Status': {'Divorced': 0, 'Married': 1, 'Single': 2},
          'Over_Time': {'No': 0, 'Yes': 1}}


In [15]:  X = employee_data[['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_answer_','Conflict_
          X.values

Out[15]: array([[4, 5, 5, ..., 3, 5, 5],
                [4, 5, 2, ..., 2, 2, 0],
                [3, 1, 1, ..., 4, 3, 3],
                ...,
                [1, 5, 4, ..., 3, 3, 6],
                [4, 1, 3, ..., 3, 3, 6],
                [5, 1, 1, ..., 3, 2, 6]], dtype=int64)

In [16]:      y = employee_data['Historical_Performance_Rating']-1
              X = employee_data[['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_answer_','Confl
              scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

FIGURE 53: Data Preparation with train and test split for predicting “Historical Performance Rating”

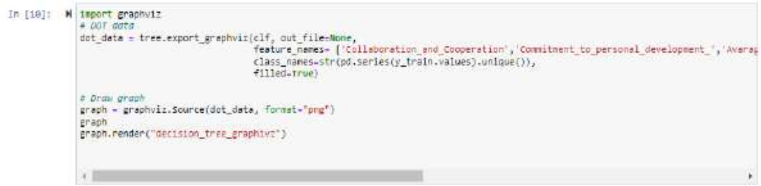
The first model that I report in this experimental performance framework represent a Decision Tree Classifier: in figure 54 is represented and it could be noticed that I didn't give further specific instructions leaving invariant the maximum growth of the tree's leaves (putting "None" on the requested 'max_depeth' parameter), and selecting a random way of splitting each leaves nodes of the tree ("random" was the order given when asked to specify the type of 'splitter. Besides, I have added an explicative image of the first three levels of the tree's depth.



```
In [13]:  from sklearn.tree import DecisionTreeClassifier
          from sklearn.model_selection import cross_val_score
          from sklearn import tree

          clf = DecisionTreeClassifier(random_state = 1518,
                                       max_depth = None,
                                       splitter = "random")

          clf = clf.fit(X_train, y_train)
```



```
In [15]:  print('Decision Tree accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          Decision Tree accuracy score: 0.7190
```



```
In [10]:  import graphviz
          # DOT data
          dot_data = tree.export_graphviz(clf, out_file=None,
                                          feature_names= ['Collaboration_and_Cooperation','Commitment_to_personal_development_','Averag
                                          class_names=str(pd.series(y_train.values).unique()),
                                          filled=True)

          # Draw graph
          graph = graphviz.Source(dot_data, format="png")
          graph
          graph.render("decision_tree_graphviz")
```

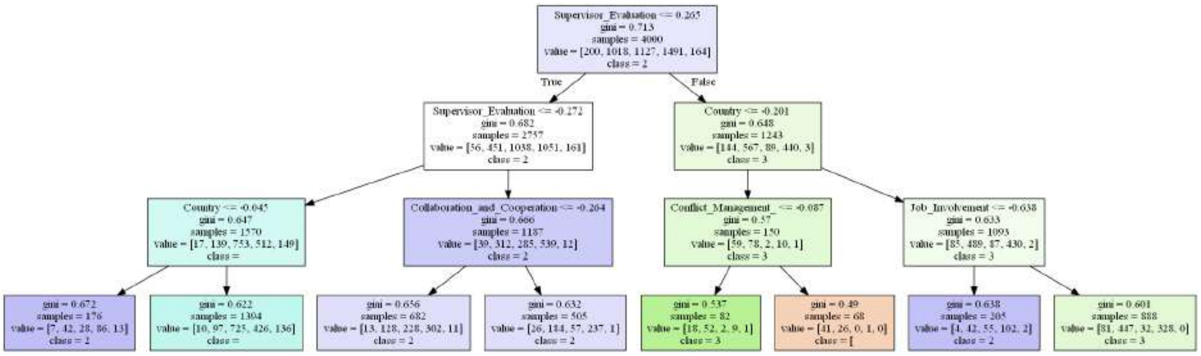
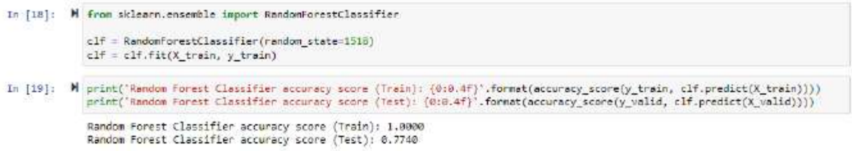

Supervisor_Evaluation <= 0.265
gini = 0.713
samples = 4000
value = [200, 1018, 1127, 1491, 164]
class = 2
True
False
Supervisor_Evaluation <= -0.272
gini = 0.682
samples = 2757
value = [56, 451, 1038, 1051, 161]
class = 2
Country <= -0.201
gini = 0.648
samples = 1243
value = [144, 567, 89, 440, 3]
class = 3
Country <= -0.045
gini = 0.647
samples = 1570
value = [17, 139, 753, 512, 149]
class =
Collaboration_and_Cooperation <= -0.204
gini = 0.666
samples = 1187
value = [39, 312, 285, 539, 12]
class = 2
Conflict_Management_ <= -0.087
gini = 0.57
samples = 150
value = [59, 78, 2, 10, 1]
class = 3
Job_Involvement <= -0.638
gini = 0.633
samples = 1093
value = [85, 489, 87, 430, 2]
class = 3
gini = 0.672
samples = 176
value = [7, 42, 28, 86, 13]
class = 2
gini = 0.622
samples = 1394
value = [10, 97, 725, 426, 136]
class =
gini = 0.656
samples = 682
value = [13, 128, 228, 302, 11]
class = 2
gini = 0.632
samples = 505
value = [26, 184, 57, 237, 1]
class = 2
gini = 0.537
samples = 82
value = [18, 52, 2, 9, 1]
class = 3
gini = 0.49
samples = 68
value = [41, 26, 0, 1, 0]
class = [
gini = 0.638
samples = 205
value = [4, 42, 55, 102, 2]
class = 2
gini = 0.601
samples = 888
value = [81, 447, 32, 328, 0]
class = 3

FIGURE 54: Decision Tree Classifier to predict "Historical Performance Rating"

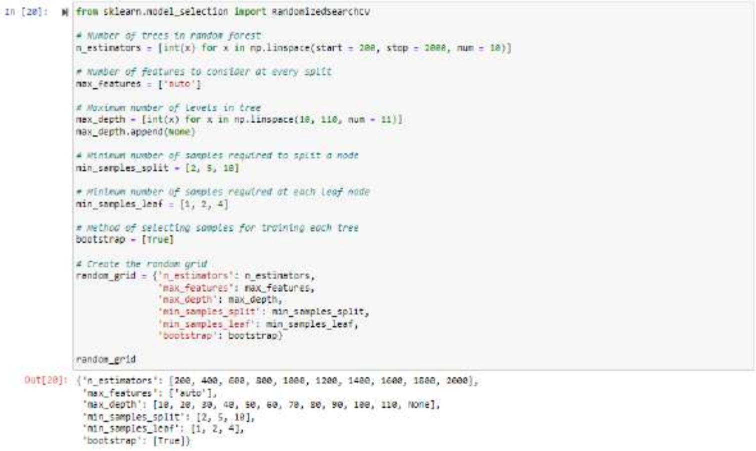
Afterward, checking the accuracy score reaching 0.71, as I wasn't totally satisfied with the model, and for this reason, I've proceed with the other analyses, I passed to implement a random forest, firstly not defining any parameters value, to verify if it could increase accuracy while preventing the model to overfit the samples. The model, illustrated in figure 55, as it could have been foreseen, raised the overall accuracy up to 0.77.



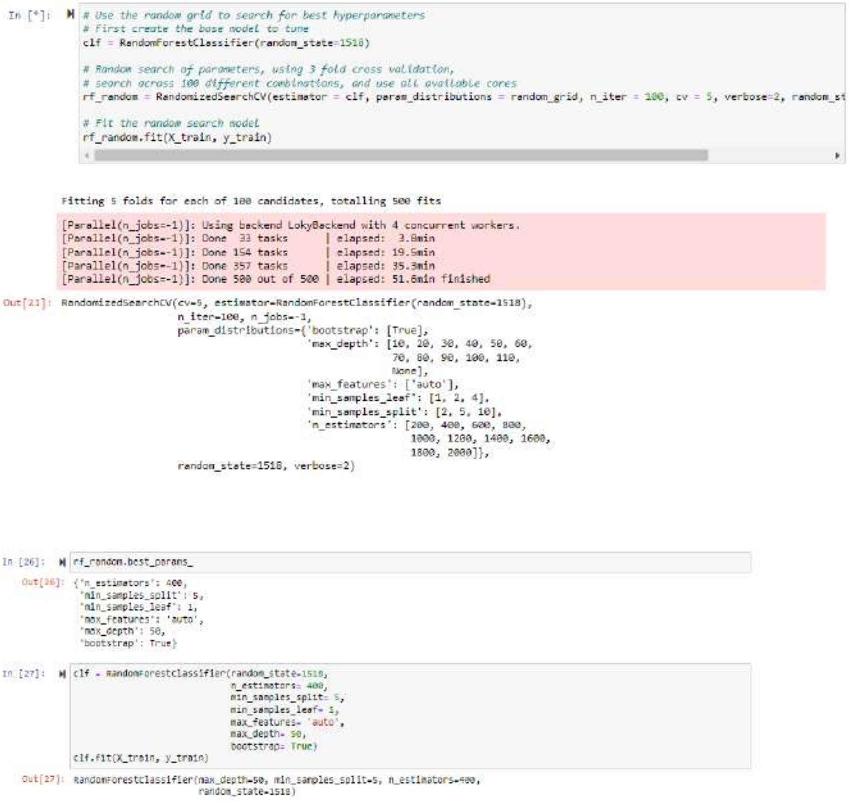
```
In [18]:  from sklearn.ensemble import RandomForestClassifier

          clf = RandomForestClassifier(random_state=1518)
          clf = clf.fit(X_train, y_train)

In [19]:  print('Random Forest Classifier accuracy score (Train): {0:0.4f}'.format(accuracy_score(y_train, clf.predict(X_train))))
          print('Random Forest Classifier accuracy score (Test): {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          Random Forest Classifier accuracy score (Train): 1.0000
          Random Forest Classifier accuracy score (Test): 0.7740
```

FIGURE 55: Basic Random Forest Classifier to predict “Historical Performance Rating”

With an additional computational effort, as shown in figure 56, I subsequently looked for an advanced hyperparameter tuning of the model operating a “Randomized Search” by giving to the algorithm the identical parameters of the Turnover Analysis Random Forest of the previous section. This request led the algorithm to process 100 iterations of different parameters combinations in accordance with the parameters random grid, from which the model, after almost one hour of processing computations, finally returned, using the module ‘rf.random_best.params_’, the most performing parameters combinations as output of the Randomized Search: a forest made of 400 trees, with 5 minimum candidates to classify at each split node, and just one employee sample required at each leaf node and a limit of 50 leaves at which stopping the growth of the trees.



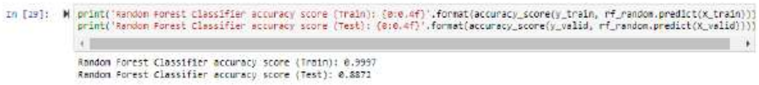
```
In [20]:  from sklearn.model_selection import RandomizedSearchCV

          # Number of trees in random forest
          n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]

          # Number of features to consider at every split
          max_features = ['auto']

          # Maximum number of levels in tree
          max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
          max_depth.append(None)

          # Minimum number of samples required to split a node
          min_samples_split = [2, 5, 10]

          # Minimum number of samples required at each leaf node
          min_samples_leaf = [1, 2, 4]

          # Method of selecting samples for training each tree
          bootstrap = [True]

          # Create the random grid
          random_grid = {'n_estimators': n_estimators,
                         'max_features': max_features,
                         'max_depth': max_depth,
                         'min_samples_split': min_samples_split,
                         'min_samples_leaf': min_samples_leaf,
                         'bootstrap': bootstrap}

          random_grid

Out[20]:  {'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000],
           'max_features': ['auto'],
           'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
           'min_samples_split': [2, 5, 10],
           'min_samples_leaf': [1, 2, 4],
           'bootstrap': [True]}
```



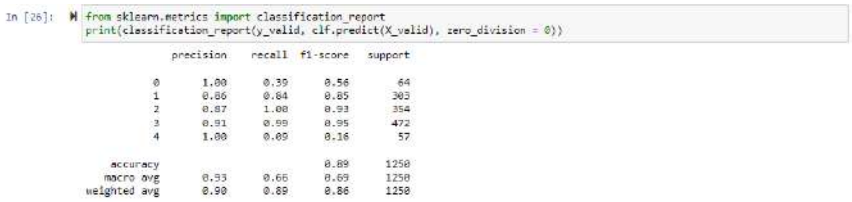
```
In [*]:  # Use the random grid to search for best hyperparameters
         # First create the base model to tune
         clf = RandomForestClassifier(random_state=1518)

         # Random search of parameters, using 3 fold cross validation,
         # search across 100 different combinations, and use all available cores
         rf_random = RandomizedSearchCV(estimator = clf, param_distributions = random_grid, n_iter = 100, cv = 5, verbose=2, random_st

         # Fit the random search model
         rf_random.fit(X_train, y_train)

        Fitting 5 folds for each of 100 candidates, totalling 500 fits

        [Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
        [Parallel(n_jobs=-1)]: Done  33 tasks      | elapsed:  3.8min
        [Parallel(n_jobs=-1)]: Done 154 tasks      | elapsed: 19.5min
        [Parallel(n_jobs=-1)]: Done 357 tasks      | elapsed: 35.3min
        [Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed: 51.6min finished

Out[21]: RandomizedSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1518),
                            n_iter=100, n_jobs=-1,
                            param_distributions={'bootstrap': [True],
                                                 'max_depth': [10, 20, 30, 40, 50, 60,
                                                               70, 80, 90, 100, 110,
                                                               None],
                                                 'max_features': ['auto'],
                                                 'min_samples_leaf': [1, 2, 4],
                                                 'min_samples_split': [2, 5, 10],
                                                 'n_estimators': [200, 400, 600, 800,
                                                                  1000, 1200, 1400, 1600,
                                                                  1800, 2000]},
                            random_state=1518, verbose=2)


In [26]:  rf_random.best_params_

  Out[26]: {'n_estimators': 400,
            'min_samples_split': 5,
            'min_samples_leaf': 1,
            'max_features': 'auto',
            'max_depth': 50,
            'bootstrap': True}

In [27]:  clf = RandomForestClassifier(random_state=1518,
                                      n_estimators= 400,
                                      min_samples_split= 5,
                                      min_samples_leaf= 1,
                                      max_features= 'auto',
                                      max_depth= 50,
                                      bootstrap= True)
         clf.fit(X_train, y_train)

  Out[27]: RandomForestClassifier(max_depth=50, min_samples_split=5, n_estimators=400,
                                  random_state=1518)
```

FIGURE 56: Random Forest Randomized Search

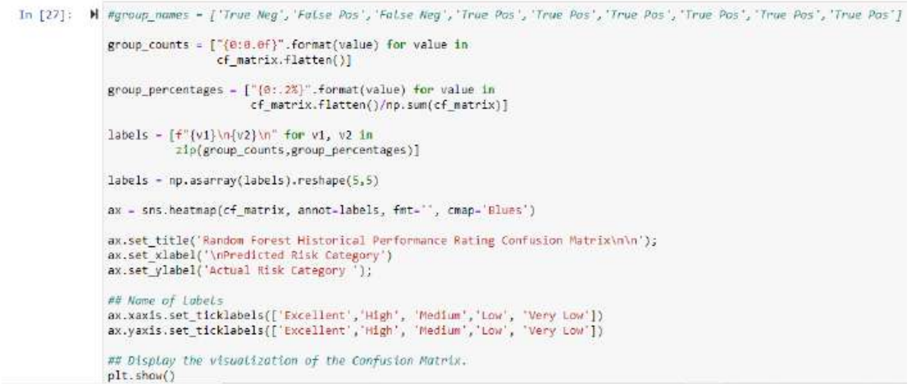
The overall accuracy rate of this enhanced model has been satisfactorily brought up to more than 10 percentage points. In figure 57, other than this score, are also deployed the final classification report, which analyzes more precisely the different components of the accuracy indicator, "precision", "recall", "f1-score" and "support", that I've already explained in the first case of engagement analysis, as well as the confusion matrix of the model from which it can be noticed that the vast majority of errors has been committed when the algorithm has classified the fifth label of the performance feature, corresponding in this analysis to the "Very Low" rate: recognizing the limitations of analysis, the model seems to be quite reliable in successfully predicting all the performance values but this lowest level, where only 5 samples over 57 have been correctly classified in its computations.



```
In [29]:  print('Random Forest Classifier accuracy score (Train): {0:0.4f}'.format(accuracy_score(y_train, rf_random.predict(X_train)))
          print('Random Forest Classifier accuracy score (Test): {0:0.4f}'.format(accuracy_score(y_valid, rf_random.predict(X_valid)))

          Random Forest Classifier accuracy score (Train): 0.9997
          Random Forest Classifier accuracy score (Test): 0.8871
```



```
In [26]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

              precision    recall  f1-score   support

           0       1.00      0.39      0.56        64
           1       0.86      0.84      0.85       303
           2       0.87      1.00      0.93       354
           3       0.91      0.99      0.95       472
           4       1.00      0.09      0.16        57

    accuracy                           0.89      1250
   macro avg       0.93      0.66      0.69      1250
weighted avg       0.90      0.89      0.86      1250
```



```
In [27]:  #group_names = ['True Neg','False Pos','False Neg','True Pos','True Pos','True Pos','True Pos','True Pos','True Pos']

          group_counts = ["{0:0.0f}".format(value) for value in
                          cf_matrix.flatten()]

          group_percentages = ["{0:.2%}".format(value) for value in
                               cf_matrix.flatten()/np.sum(cf_matrix)]

          labels = [f"{v1}\n{v2}\n" for v1, v2 in
                    zip(group_counts,group_percentages)]

          labels = np.asarray(labels).reshape(5,5)

          ax = sns.heatmap(cf_matrix, annot=labels, fmt='', cmap='Blues')

          ax.set_title('Random Forest Historical Performance Rating Confusion Matrix\n\n');
          ax.set_xlabel('\nPredicted Risk Category')
          ax.set_ylabel('Actual Risk Category ');

          ## Name of Labels
          ax.xaxis.set_ticklabels(['Excellent','High', 'Medium','Low', 'Very Low'])
          ax.yaxis.set_ticklabels(['Excellent','High', 'Medium','Low', 'Very Low'])

          ## Display the visualization of the Confusion Matrix.
          plt.show()
```

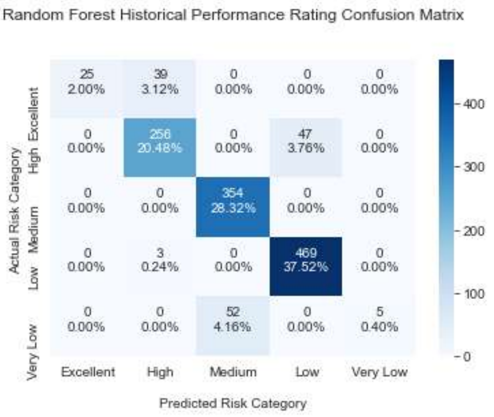
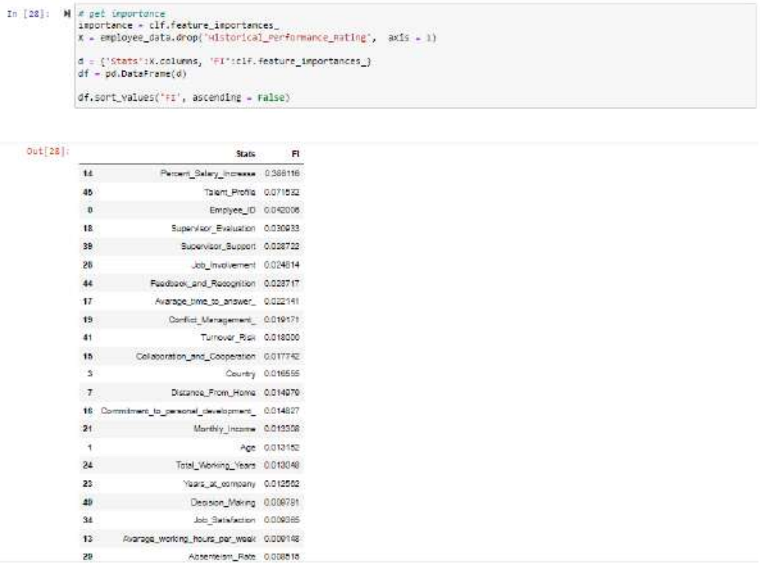


Figure 57: Random Forest predicting "Historical Performance Rating" new relative accuracy score, classification report and confusion matrix

In the next figure 58, it's illustrated a process that was not exhibited across the set of other analyses: a "Feature Importance" analysis has been performed to understand which variables revealed to be more contributing to decide where to classify the sample on the prediction made by the last model of random forest presented. Investigating the comparative impact of each feature, the function 'clf.feature_importance' automatically computed the importance score of each variable in the training phase, returning an ordinate list scaled by significance. Evaluating the findings, I notice that "Percent Salary Increase", despite not being considered voluntarily on the training phase as an independent variable, would actually be contributing quite a lot, outdistancing other features, in predicting the values of "Historical performance Rating" (0.38). However, as for “employee_id”, a dimension whose observations were just increasing ordinal numbers representing the ID number of each employee in the Excel sheet, this Feature Importance’s outcomes could even misguide analysts in evaluating how to build their model. In fact, as explained in the previous section, analysts should be careful to interpret the meaning of some varibale’s patterns using their HR contextual knowledge of their workforce at hand. In this sense, analysts would reject to take “Percent Salary Increase” in consideration as a variable to fit a machine learning model as they would be aware that those specific dimensions’ values were direct consequence of the same performance: the percentage of additional pay, in my dataset simulated, have been generated according to the performance rating of each employee, and that’s the reason why the system identifies this variable as pretty explainable of the prediction of performance. On the other hand, the other figures confirm that the data selected as input to the model have been critically and meaningfully utilized for the prediction.



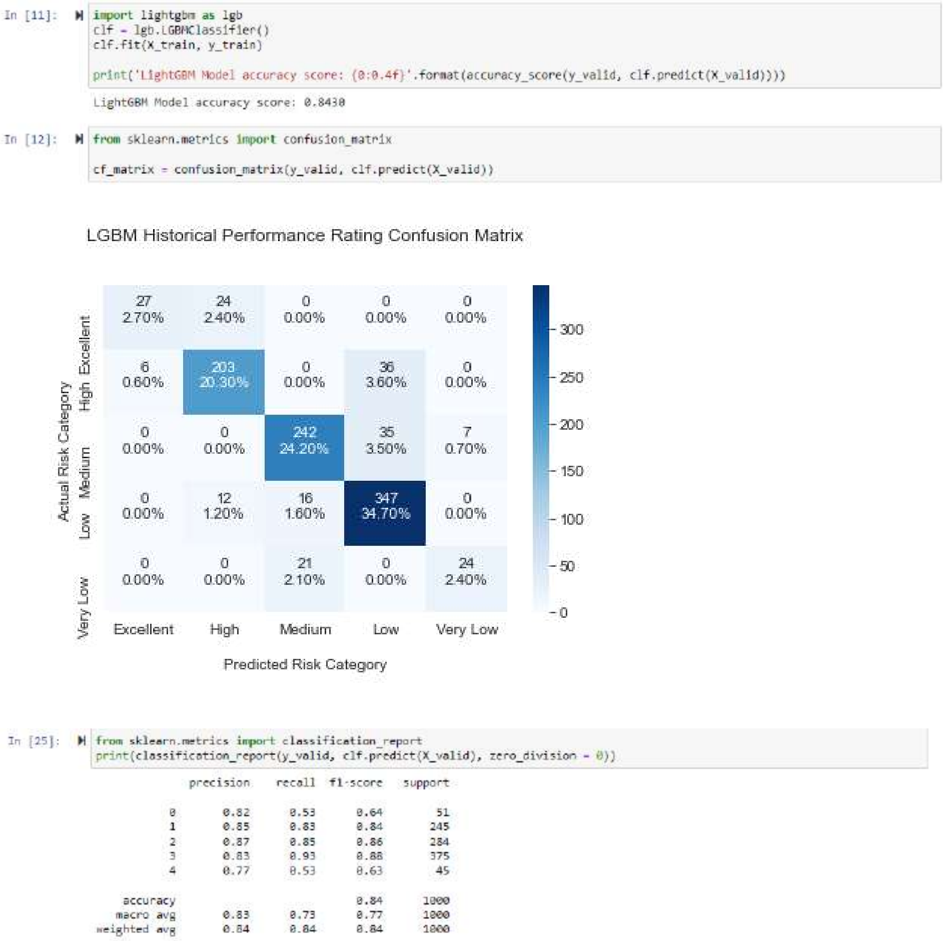
```
In [28]:  # get importance
          importance = clf.feature_importances_
          X = employee_data.drop('Historical_Performance_Rating',  axis = 1)

          d = {'Stats':X.columns, 'FI':clf.feature_importances_}
          df = pd.DataFrame(d)

          df.sort_values('FI', ascending = False)
```

Out[28]:

	Stats	FI
14	Percent_Salary_Increase	0.388116
45	Talent_Profile	0.071532
0	Emplyee_ID	0.042006
18	Supervisor_Evaluation	0.030933
39	Supervisor_Support	0.028722
26	Job_Involvement	0.024814
44	Feedback_and_Recognition	0.023717
17	Avarage_time_to_answer_	0.022141
19	Conflict_Management_	0.019171
41	Turnover_Risk	0.018000
15	Collaboration_and_Cooperation	0.017742
5	Country	0.016555
7	Distance_From_Home	0.014979
16	Commitment_to_personal_development_	0.014827
21	Monthly_Income	0.013308
1	Age	0.013152
24	Total_Working_Years	0.013048
23	Years_at_company	0.012562
40	Decision_Making	0.009781
34	Job_Satisfaction	0.009365
13	Avarage_working_hours_per_week	0.009148
29	Absenteism_Rate	0.008516

Figure 58: Random Forest Feature Importance for predicting “Historical Performance Rating”

The last machine-learning model performed for this field of analysis returned the most accurate without even trying to optimize the parameter of the algorithm , which was a Light Gradient Boosting Model, known to be greatly efficient in improving random forests with distributing rewards and penalties at each model’s iteration for each node split made. Figure 59 reports the model that first has been fitted for learning the data pattern, as usual, and then was tested with predicting the values of "Historical Performance Rating" response variables on the validation set, with an overall accuracy of 0.84. Below, besides the final results, are also presented the classification report and the confusion matrix to realize where in particular the model came up against difficulties and made some errors in specific label classifications.



```
In [11]:  import lightgbm as lgb
          clf = lgb.LGBMClassifier()
          clf.fit(X_train, y_train)

          print('LightGBM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          LightGBM Model accuracy score: 0.8430

In [12]:  from sklearn.metrics import confusion_matrix

          cf_matrix = confusion_matrix(y_valid, clf.predict(X_valid))
```

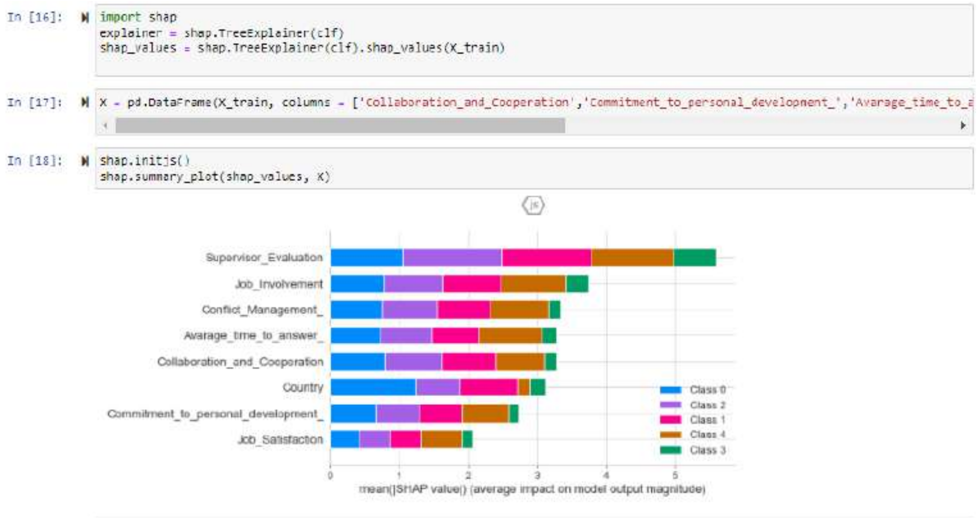
LGBM Historical Performance Rating Confusion Matrix

	Excellent	High	Medium	Low	Very Low
Excellent	27 2.70%	24 2.40%	0 0.00%	0 0.00%	0 0.00%
High	6 0.60%	203 20.30%	0 0.00%	36 3.60%	0 0.00%
Medium	0 0.00%	0 0.00%	242 24.20%	35 3.50%	7 0.70%
Low	0 0.00%	12 1.20%	16 1.60%	347 34.70%	0 0.00%
Very Low	0 0.00%	0 0.00%	21 2.10%	0 0.00%	24 2.40%

```
In [25]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

              precision    recall  f1-score   support

           0       0.82      0.53      0.64        51
           1       0.85      0.83      0.84       245
           2       0.87      0.85      0.86       284
           3       0.83      0.93      0.88       375
           4       0.77      0.53      0.63        45

    accuracy                           0.84      1000
   macro avg       0.83      0.73      0.77      1000
weighted avg       0.84      0.84      0.84      1000
```

FIGURE 59: Basic LGBM Classifier to predict "Historical Performance Rating" with relative accuracy score, classification report and confusion matrix.

To conclude this section, a "Shap values" Feature Importance process has been performed and deployed in the following figure 60 for this LGBM model that was built to forecast the levels of performance of the workforce at hand. By ordering in the y-axis the features used, according to their relative contribution to individual performance levels' predictions, the analysis seems to returns an equilibrate situation, in which only the greater impact of “Supervisor Evaluation” sticks out: HR top managers now are aware that in processing the forecasts on performance ratings, this employee feature has represented the most important factors, so they would wisely have a special attention on it.



```
In [16]:  import shap
           explainer = shap.TreeExplainer(clf)
           shap_values = shap.TreeExplainer(clf).shap_values(X_train)

In [17]:  X = pd.DataFrame(X_train, columns = ['Collaboration_and_Cooperation','Commitment_to_personal_development_','Avarage_time_to_a

In [18]:  shap.initjs()
           shap.summary_plot(shap_values, X)
```

FIGURE 60: ‘Shap Values’ Feature Importance for the LGBM Model predicting “Historical Performance Rating”

3.3.4 Talent profile Analysis

This last section of analysis is dedicated, for what concerns the descriptive part, to examining more closely if a solid pool of talents are currently present inside the company, reasoning around how to adjusts the current state of HR policies to intensify the efforts to get a comfortable situation for them to develop; on the other hand, from the predictive perspective, since talents represent a human invaluable asset on which company should invest for assuring a bright future, it will be evaluated, by running direct analysis on my data, if machine-learning algorithms, recognizing their model limitations, could effectively help HR recruiters in identifying personnel who is most likely to become a great performer with high-ranked traits. Jumping into the analyses, on figure 61 it could be visualized as a dashboard made of some bar charts that offer some overview numbers of the “Talent Profile” trend across the workforce, which rates are hypothetically derived by gathering soft skills assessment and other pulse surveys, according to some relevant employee factors. From the second graphs it could be grasped that talents are distributed almost equally among the two gender types, with male slightly presenting better percentage numbers in the best-ranked level that should certify that employees are effectively talented. Overall, the company should realize that the two categories most represented by its labor force are the second, that indicate the employees are still distant from being developing the right skills, and the third level, that is the middle way level. Unfortunately, in addition, for both the male and female population, he worst level of talent exceeds the best one , though by a few percentage points. From this panoramic picture, then, the HR department should not be worried as there seem to be no critical first-hand problems but neither should stop looking outside to make sure the next recruiting efforts will result in valuable candidates. That being said, on the second bar chart HR managers could focus on evaluating the situation of training delivered to the most promising personnel. Even if from these figures emerge a positive trend for which there have been less top-level talented employees who got no training at all in comparison to the other talent profile levels , the company's training offer seems to be still scarce as, considering always at the highest talenting scores, they proportional percentage of people that have globally received more than 3 training sessions is even lower than other talents' clusters; training represent in fact one the company's strategy to develop skills of employees, in particular this should be proven to worth for talented employees. Moreover, promotions are also a product of a series of training and other affecting employee's career growth factors, for this reason the figures reported in the third and last bar chart are not surprising as they depict a situation pretty like the previous one, with even less differences among the five talenting categories in which e entire workforce is discriminated. Rather, logically, far more top-talented employees, in proportion, should be expected to have been promoted two, three or more times than other lower talented workers at the disposal of the firm.

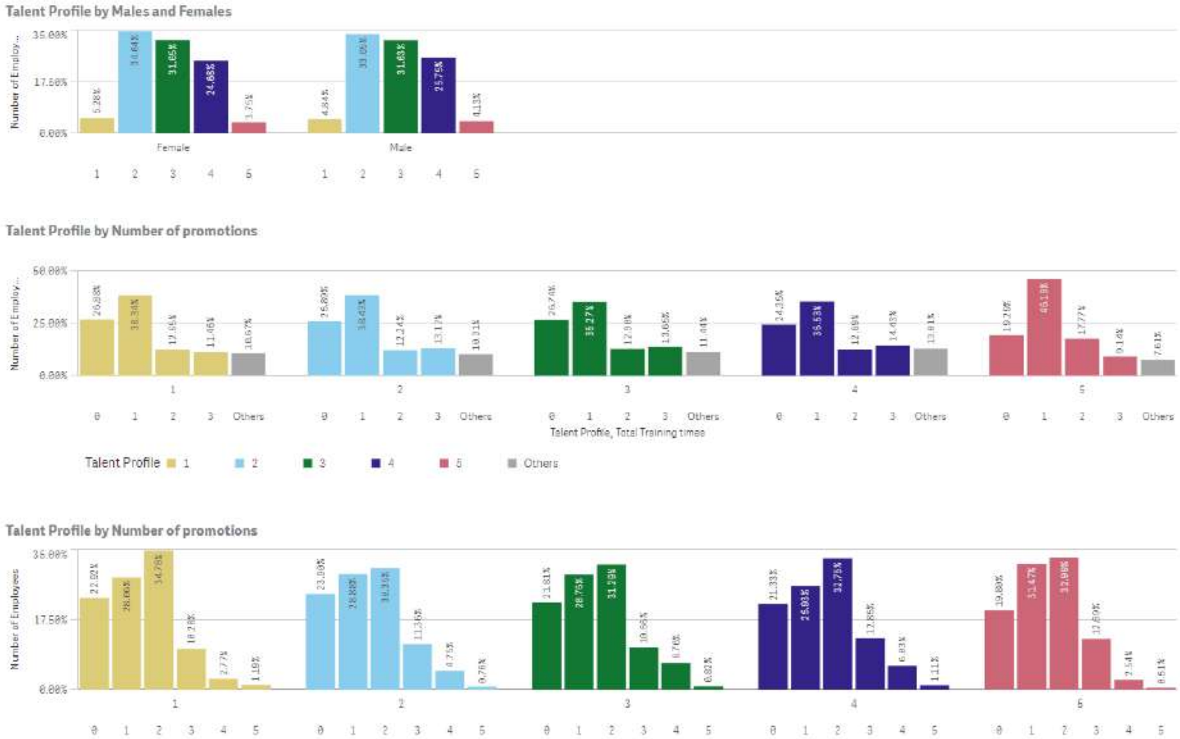
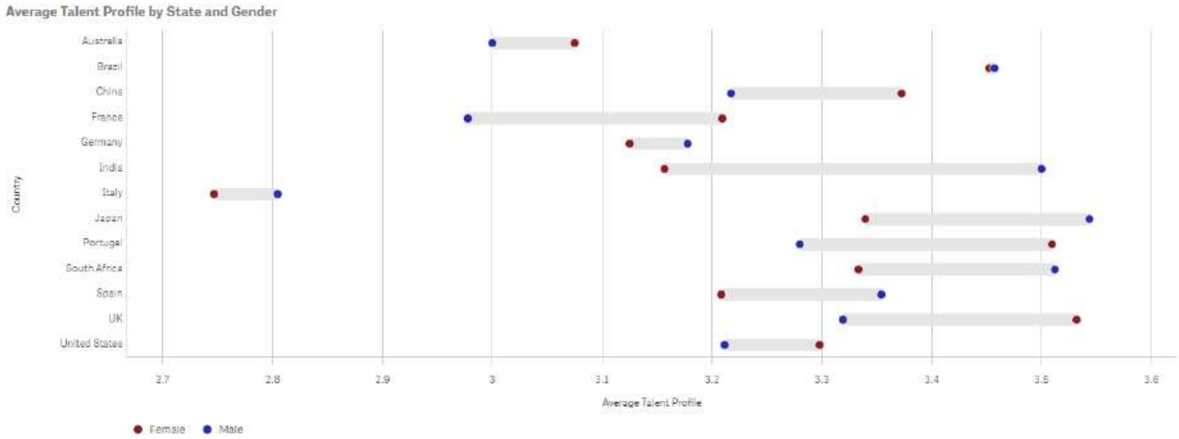


FIGURE 61: “Talent Profile” by Gender, Training and Promotion bar charts

The dashboard coming after in figure 62 deploys a diversity detailed study to evaluate how it varies the average level of talent profiles within the workers that are here divided by Country of origin. Italian employees are shockingly found to be quite behind other employee clusters, recording average levels that go from 2.75 for female to 2.8 for male, while expatriates employees line up to higher level from at least 3 averagely to even 3.5 for states like Japan, Portugal, India, South Africa and UK, where the highest proportion of talents is concentrated. Italy’s trend is due mostly to the fact that local employees are far more than other, since they represent around the 75% of the whole workforce, and, since talent profiles corresponding to the top fifth level are less in proportion to other levels, their average score is pulled behind. Overall, however, this analyzed company has been provided with a clear picture of how diversity is positively impacting the talent pool building inside the labor force: the numbers given by this research describe in fact a satisfactory situation in which talent recrutement from almost every non-local countries is turning out to have been successful, while new strategies should be addressed to improve the future selection of more talented local candidates. Then, as denoted by many theoretical studies on the positive benefits of having a diverse workforce for the overall business performance, this firm represents a case in point in this direction. Going further, the succeeding scatter plot, which examines different grades of talenting worker profiles based on their average length of service on the company and their average level of income, emphasizes two main underlying patterns, one positive and one negative. From one perspective, as a matter of fact, least-promising employees, with the score of "1", correspond to that, other than staying the longest time so far at the company, are also gaining more than other clusters, although their soft skill profile don't match what this firm is mostly needing according to a evaluation if its future requirements and projects; on the other hand, just below this category, is plotted the cluster of employees representing the identified key-future employees that the company is most mentioned to retain to assign them great responsibilities in a medium and long-term scenario. Rather, perhaps additional efforts should be directed to the second best-talented group of employees, scoring "4" at the assessment surveys and that are performing well, as they have lower salary earnings and were averagely insediated within the company's walls from shorter time, even if could be so simply because they are averagely younger than other groups; this last assumption should be then verified to proceed with further in-depth analyses to improve the retention effects of the enterprise and to show the brand itself as more attracting from external candidates.



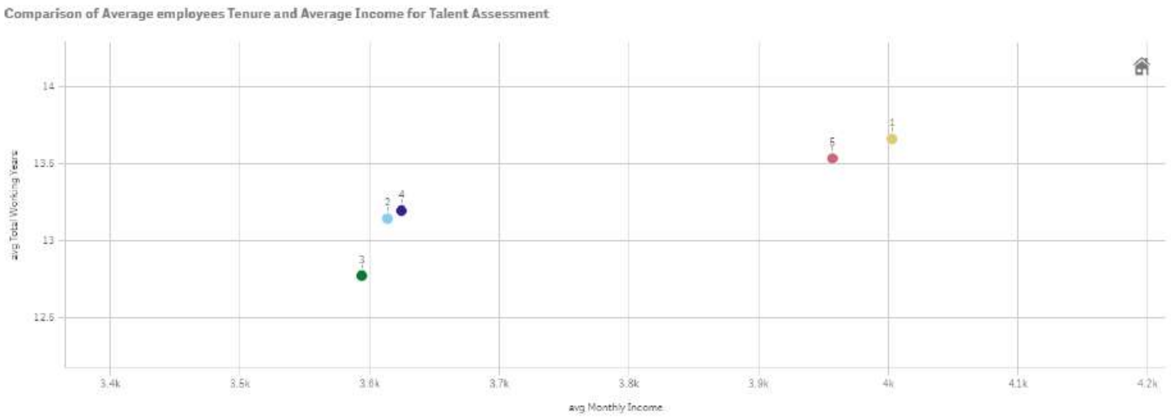


FIGURE 62: Country-diversity talents' distribution plot and Average talents' Tenure and Income scatter plot

The dashboard presented on the following Figure 63 highlights a very optimistic and efficient scenario of the company regarding how percentage rises in wages have been distributed across the same employee clusters which have been discriminated by their talent profile score and their managerial position. HR policies in this case seem to be working well as the greatest percentage of additional income has been conferred to most-talented employees in a perfect descending order of talenting categories. This is a logical consequence of how my data has been simulated, but for a hypothetically real company, would represent a state-of-the-art payroll management based on fair meritocracy and objective manager's judgment, as those talent profile indicators have been naturally affected by employees' past performance scores. Ultimately, the last bar chart analyzes a very specific HR metrics that gives valuable insights on how successfully have been carrying out he recent HR recruitment campaigns aimed at hiring promising highly-skilled young people. This KPI is called "Quality of Hire" verifies if employees who have been hired over the last 3 years and whose talent rate was at the top levels, then "4" or "5" out of "5", have also been showing an impressive performance lying among "High" or "Excellent" levels, discriminating the results based on gender. The findings are easy-to-interpret: a total of 0,31 (the majority of which corresponding to employees with maximum talenting score), including male and female components, that present very similar figures, with respect to all the last three years hires, have been turned out to be successful selections prelevated by the HR department from the labor market. This means the direction undertaken by the firm is right and will supposedly pay for the efforts committed for the purpose. At last, we suggest companies pay more attention to a problem that there are no discrepancies among incentives

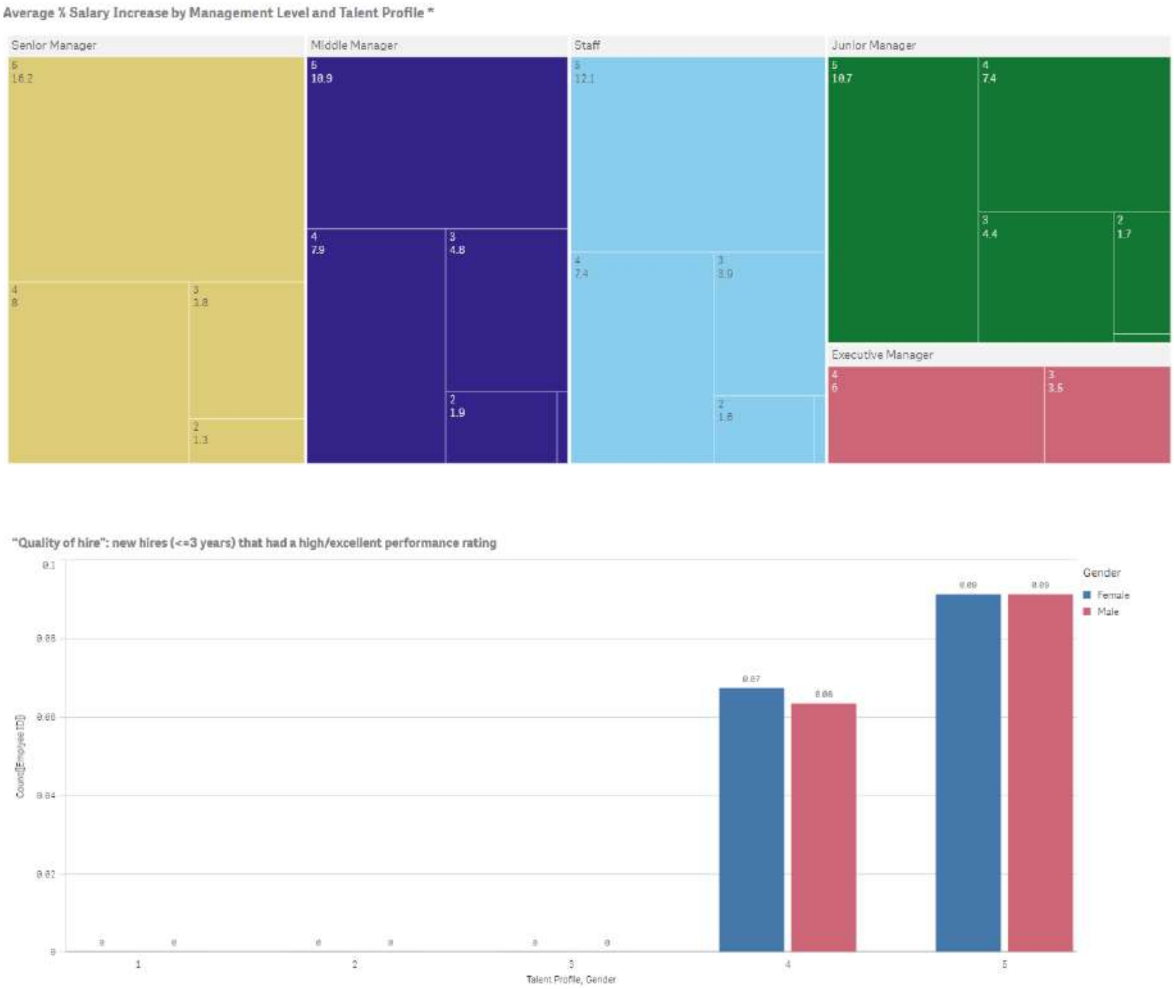
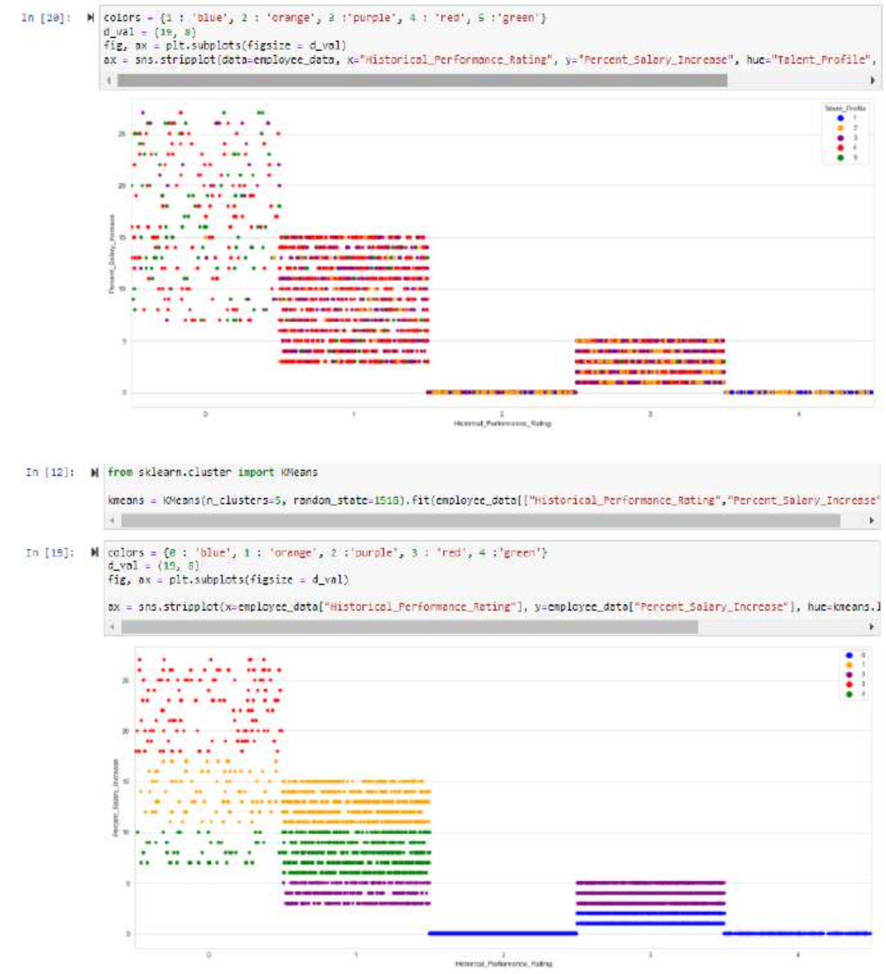


FIGURE 63: Average salary increase by management position mekko chart and "Quality of hire" bar chart

To conclude the descriptive part of analysis for this section, I decided once again to graphically compare two different but similar techniques. In figure 64, the second graph deployed was built through the launch of an intelligent K-means clustering algorithm whose goal was to divide my entire workforce dataset into a five clusters based on some particular characteristics, in this case based on the "Historical Performance Rating" and "Percent Salary Employee" factors. This gives management a clearer perspective of how most talented employees should be identified if these two were the only factors to be taken under consideration. The right above graph is a product of a normal cluster analysis that reports a faithful picture of how talents are at the moment actually distributed among the workforce in accordance with their performance level and percentage of last additional pay. Remembering that that performance rating “excellent” correspond to the associated value “0”, it could be noticed the relevant difference provided by the confront between the two graphs; while in the K-means classification, most-promising employees correspond to the red cluster associated with a top-performance score and a very considerable recent pay rise, on the other hand the real-word situation, depicted thanks to the cluster analysis, shows that there is quite a confusion in clearly distinguishing the five groups based on the two employee attributes. this consideration could drive HR top-managers either to be more careful next time in establishing an informed pay rise structure planning or to dìmore deeply analyze the talent profile trend inside their organization delving into other affecting features that were not taken into account in this analysis.



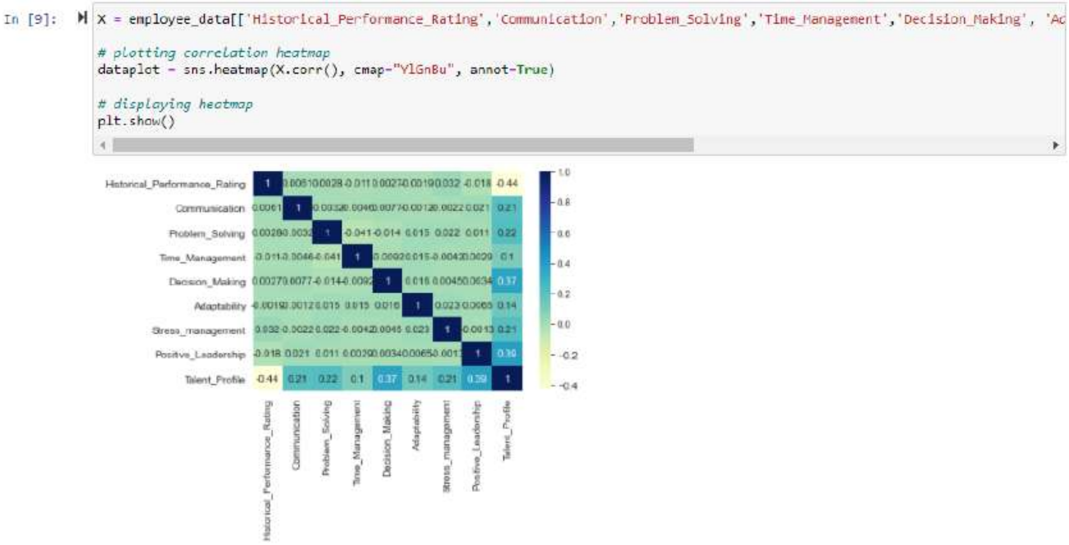
```
In [20]: colors = {1 : 'blue', 2 : 'orange', 3 :'purple', 4 : 'red', 5 :'green'}
         d_val = (19, 8)
         fig, ax = plt.subplots(figsize = d_val)
         ax = sns.stripplot(data=employee_data, x="Historical_Performance_Rating", y="Percent_Salary_Increase", hue="Talent_Profile",
```

```
In [12]: from sklearn.cluster import KMeans

         kmeans = KMeans(n_clusters=5, random_state=1518).fit(employee_data[["Historical_Performance_Rating","Percent_Salary_Increase"
```

```
In [19]: colors = {0 : 'blue', 1 : 'orange', 2 :'purple', 3 : 'red', 4 :'green'}
         d_val = (19, 8)
         fig, ax = plt.subplots(figsize = d_val)

         ax = sns.stripplot(x=employee_data["Historical_Performance_Rating"], y=employee_data["Percent_Salary_Increase"], hue=kmeans.l
```

FIGURE 64: Cluster analysis versus K-Means clustering classification for "Talent Profile" based on the variables "Percent Salary Increase" and "Historical Performance Rating"

Having finished the descriptive part of analyses, I approached the data preparation processes to get my algorithm ready to make the requested predictions by plotting a Person Correlation analysis showing the strength of linearity patterns among the variables which I arbitrarily selected, keeping in mind the assumptions made on elaborating my simulated dataset, to give to my further model in the learning phases. In figure 65, three main independent employee features appear quite correlated to the response variable "Talent Profile", showing the strongest degrees resulted by the whole set of analyses of this type run up to this point: "Historical Performance Rating", showing a score of 0.44 of negative correlation, and "positive Leadership" and "Talent Profile" that figures as positively correlated to the same dependent variable with a score of 0.39 and 0.37. All the other features under examination all share a positive linear pattern of correlation with the talent dimension. Still, importantly, I found that higher past performance rates are associated with lower “Talent Profile” scores. This negative pattern of relationship between the two has no logical meaning: in fact, as for the analyses implemented before, “Historical Performance Rating” labels have been labeled, in the data pre-processing stage, in an inverse order, converting the lowest score “Very Low” to the integer number “4”, considered the highest in this logistic model, and the top value “Excellent” to the number “0”. For this reason, the machine-learning models should always be interpreted firstly by the analysts who have built and processed them, because this could be a fatal mistake. Finally, we could then argue that performance, like the other independent variables in the model, is positively correlated to the talent measure.

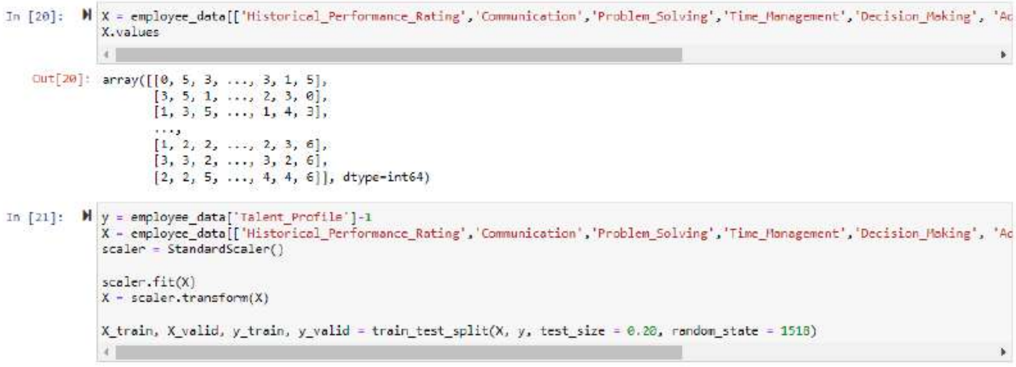


```
In [9]:  X = employee_data[['Historical_Performance_Rating','Communication','Problem_Solving','Time_Management','Decision_Making', 'Ac

         # plotting correlation heatmap
         dataplot - sns.heatmap(X.corr(), cmap-"YlGnBu", annot-True)

         # displaying heatmap
         plt.show()
```

	Historical_Performance_Rating	Communication	Problem_Solving	Time_Management	Decision_Making	Adaptability	Stress_management	Positive_Leadership	Talent_Profile
Historical_Performance_Rating	1	0.0061	0.0028	-0.011	0.0027	0.0019	0.032	-0.018	-0.44
Communication	0.0061	1	0.0032	0.0046	0.0077	0.0012	0.0022	0.021	0.21
Problem_Solving	0.0028	0.0032	1	-0.041	-0.014	0.015	0.022	0.011	0.22
Time_Management	-0.011	-0.0046	-0.041	1	0.0092	0.015	-0.0043	0.0029	0.1
Decision_Making	0.0027	0.0077	-0.014	-0.0092	1	0.016	0.0045	0.0034	0.37
Adaptability	-0.0019	0.0012	0.015	0.015	0.016	1	0.023	0.0065	0.14
Stress_management	0.032	-0.0022	0.022	-0.0043	0.0045	0.023	1	-0.0013	0.21
Positive_Leadership	-0.018	0.021	0.011	0.0029	0.0034	0.0065	0.0013	1	0.39
Talent_Profile	-0.44	0.21	0.22	0.1	0.37	0.14	0.21	0.39	1

FIGURE 65: Heatmap Correlation Analyses for “Talent Profile”

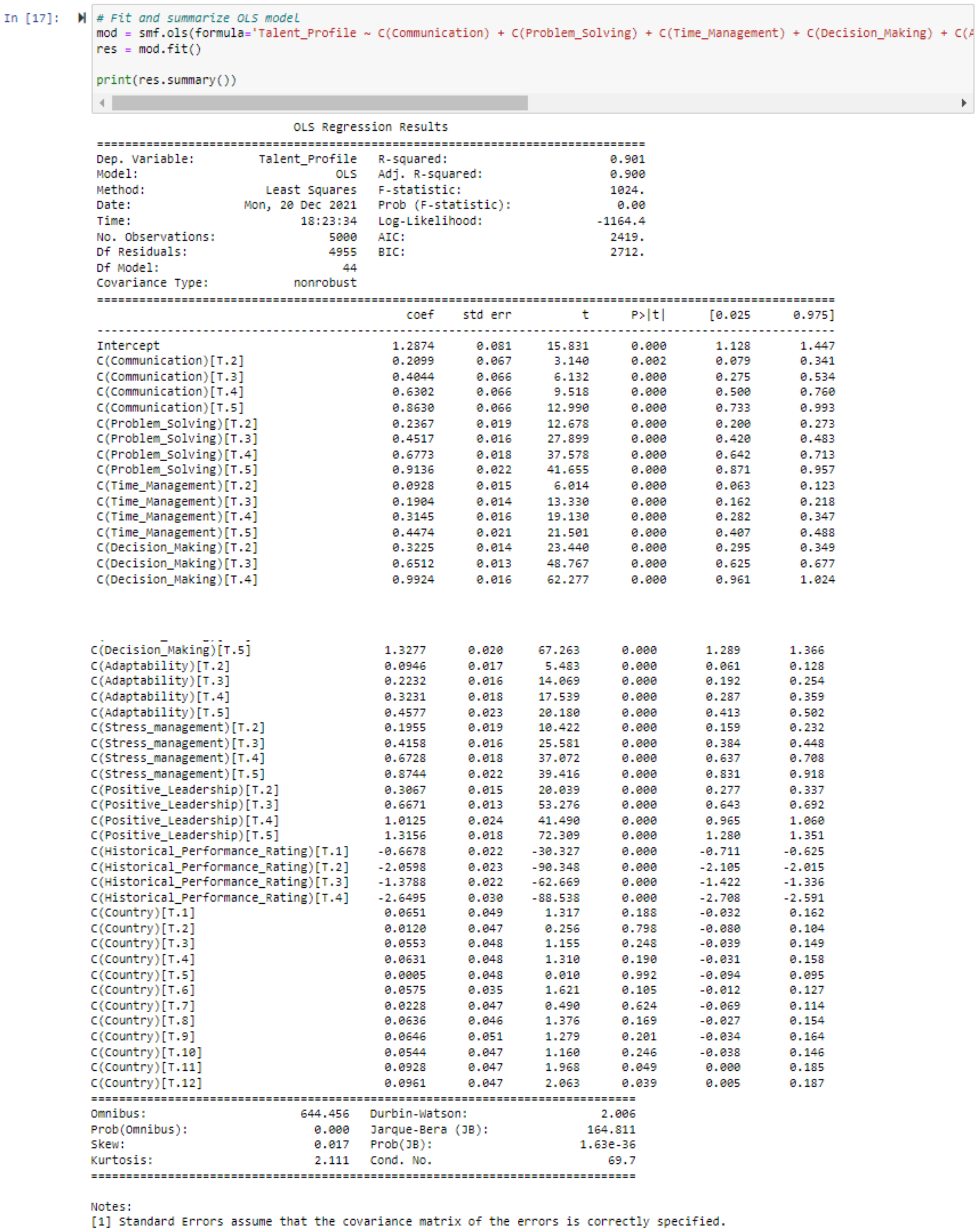
As a premise, it should be clear that this whole predictive case of analyses has been designed to help HR recruiters to evaluate on which specific employees the company should heavily invest and make additional and careful action plan to develop their careers: these employees will be in fact identified according to the estimations that will result from the machine-learning computations, that will understand the most impacting employee work-related factors and personal traits that have led scrutinized talents, currently present in the company, to be considered so. Later on the analysis, I then chose the same set of variables to create the train test “x”, that my algorithms will need to lean the data patterns indispensable to make the prediction of the labels belonging to the dimension “Talent Profile”, the response variable which is removed from the created “y” test set on purpose as it could be noticed in figure 66, where is operated the usual function “train_test_split” to divide my dataset as anticipated.



```
In [20]:  X = employee_data[['Historical_Performance_Rating','Communication','Problem_Solving','Time_Management','Decision_Making', 'Ac
          X.values

Out[20]:  array([[0, 5, 3, ..., 3, 1, 5],
                 [3, 5, 1, ..., 2, 3, 0],
                 [1, 3, 5, ..., 1, 4, 3],
                 ...,
                 [1, 2, 2, ..., 2, 3, 6],
                 [3, 3, 2, ..., 3, 2, 6],
                 [2, 2, 5, ..., 4, 4, 6]], dtype=int64)

In [21]:  y = employee_data['Talent_Profile']-1
          X = employee_data[['Historical_Performance_Rating','Communication','Problem_Solving','Time_Management','Decision_Making', 'Ac
          scaler = StandardScaler()

          scaler.fit(X)
          X = scaler.transform(X)

          X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size = 0.20, random_state = 1518)
```

FIGURE 66: Data Preparation with train and test split for predicting “Talent Profile”

The first predictive model that I run corresponds to a Logistic Regression classifier, illustrated in figure 67, whose regressors have all been classified into categories, instead of single regressors. For this reason, the model results are more difficult to read, but, on the other side, more precise on its outcomes. In general, the “R-squared” measure returned sensational results, indicating that for a percentage over 90%, indicating the strength of the relationship between the model and the dependent variable, the model is able, with the selected independent variables and considering its limitations, to explain the generation of the researched variable “Talent Profile”. Also, all the p-values of the variables intercept’ coefficients are all statistically meaningful, except for “Country'' whose variance is not enough correlated to the talenting indicator: the algorithm successfully predicts variation in Talent levels based on these variables, that could then be deemed as talent predictors affecting the development of their labor force. The same issue of negative correlation is highlighted for “Historical Performance Rating” variable that should be submitted to the careful examination of the HR analysts before passing through the decision-making C-suite level of the company.

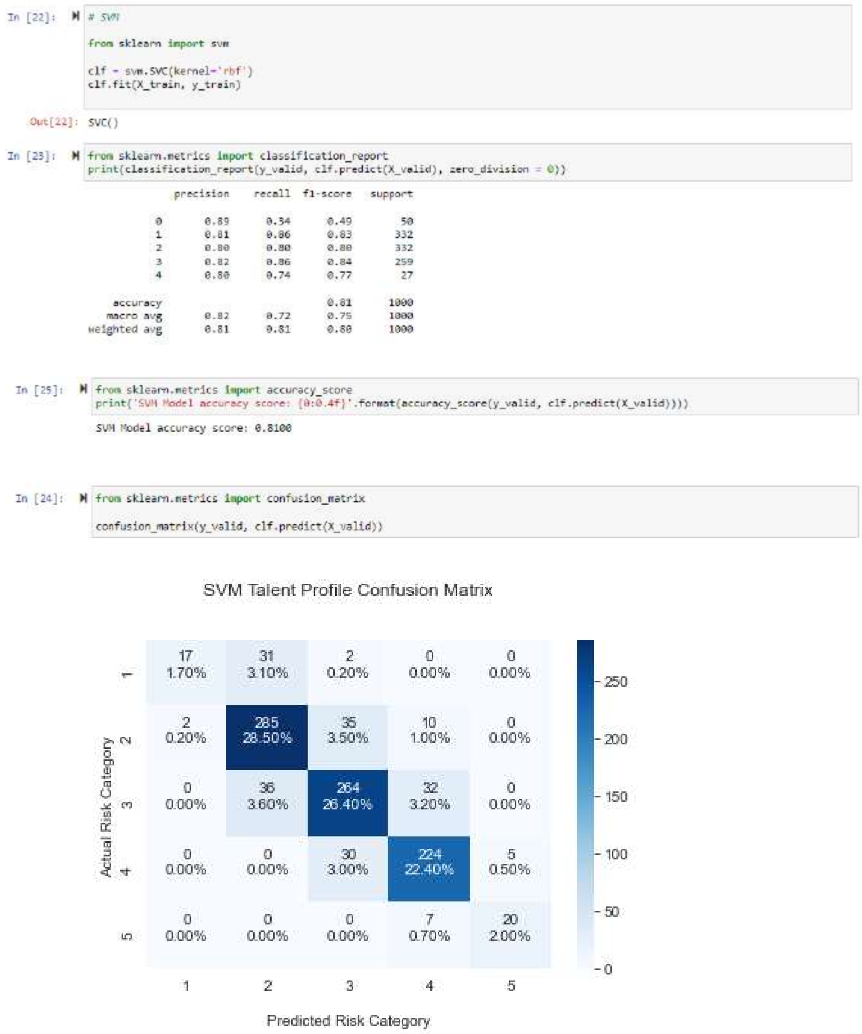


```
In [17]:  # Fit and summarize OLS model
           mod = smf.ols(formula='Talent_Profile ~ C(Communication) + C(Problem_Solving) + C(Time_Management) + C(Decision_Making) + C(A
           res = mod.fit()

           print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Talent_Profile   R-squared:                       0.901
Model:                             OLS   Adj. R-squared:                  0.900
Method:                  Least Squares   F-statistic:                     1024.
Date:                 Mon, 20 Dec 2021   Prob (F-statistic):               0.00
Time:                         18:23:34   Log-Likelihood:                -1164.4
No. Observations:                 5000   AIC:                             2419.
Df Residuals:                     4955   BIC:                             2712.
Df Model:                           44
Covariance Type:             nonrobust
========================================================================================================
                                            coef    std err          t      P>|t|      [0.025     0.975]
--------------------------------------------------------------------------------------------------------
Intercept                                 1.2874      0.081     15.831      0.000       1.128      1.447
C(Communication)[T.2]                     0.2099      0.067      3.140      0.002       0.079      0.341
C(Communication)[T.3]                     0.4044      0.066      6.132      0.000       0.275      0.534
C(Communication)[T.4]                     0.6302      0.066      9.518      0.000       0.500      0.760
C(Communication)[T.5]                     0.8630      0.066     12.990      0.000       0.733      0.993
C(Problem_Solving)[T.2]                   0.2367      0.019     12.678      0.000       0.200      0.273
C(Problem_Solving)[T.3]                   0.4517      0.016     27.899      0.000       0.420      0.483
C(Problem_Solving)[T.4]                   0.6773      0.018     37.578      0.000       0.642      0.713
C(Problem_Solving)[T.5]                   0.9136      0.022     41.655      0.000       0.871      0.957
C(Time_Management)[T.2]                   0.0928      0.015      6.014      0.000       0.063      0.123
C(Time_Management)[T.3]                   0.1904      0.014     13.330      0.000       0.162      0.218
C(Time_Management)[T.4]                   0.3145      0.016     19.130      0.000       0.282      0.347
C(Time_Management)[T.5]                   0.4474      0.021     21.501      0.000       0.407      0.488
C(Decision_Making)[T.2]                   0.3225      0.014     23.440      0.000       0.295      0.349
C(Decision_Making)[T.3]                   0.6512      0.013     48.767      0.000       0.625      0.677
C(Decision_Making)[T.4]                   0.9924      0.016     62.277      0.000       0.961      1.024


C(Decision_Making)[T.5]                   1.3277      0.020     67.263      0.000       1.289      1.366
C(Adaptability)[T.2]                      0.0946      0.017      5.483      0.000       0.061      0.128
C(Adaptability)[T.3]                      0.2232      0.016     14.069      0.000       0.192      0.254
C(Adaptability)[T.4]                      0.3231      0.018     17.539      0.000       0.287      0.359
C(Adaptability)[T.5]                      0.4577      0.023     20.180      0.000       0.413      0.502
C(Stress_management)[T.2]                 0.1955      0.019     10.422      0.000       0.159      0.232
C(Stress_management)[T.3]                 0.4158      0.016     25.581      0.000       0.384      0.448
C(Stress_management)[T.4]                 0.6728      0.018     37.072      0.000       0.637      0.708
C(Stress_management)[T.5]                 0.8744      0.022     39.416      0.000       0.831      0.918
C(Positive_Leadership)[T.2]               0.3067      0.015     20.039      0.000       0.277      0.337
C(Positive_Leadership)[T.3]               0.6671      0.013     53.276      0.000       0.643      0.692
C(Positive_Leadership)[T.4]               1.0125      0.024     41.490      0.000       0.965      1.060
C(Positive_Leadership)[T.5]               1.3156      0.018     72.309      0.000       1.280      1.351
C(Historical_Performance_Rating)[T.1]    -0.6678      0.022    -30.327      0.000      -0.711     -0.625
C(Historical_Performance_Rating)[T.2]    -2.0598      0.023    -90.348      0.000      -2.105     -2.015
C(Historical_Performance_Rating)[T.3]    -1.3788      0.022    -62.669      0.000      -1.422     -1.336
C(Historical_Performance_Rating)[T.4]    -2.6495      0.030    -88.538      0.000      -2.708     -2.591
C(Country)[T.1]                           0.0651      0.049      1.317      0.188      -0.032      0.162
C(Country)[T.2]                           0.0120      0.047      0.256      0.798      -0.080      0.104
C(Country)[T.3]                           0.0553      0.048      1.155      0.248      -0.039      0.149
C(Country)[T.4]                           0.0631      0.048      1.310      0.190      -0.031      0.158
C(Country)[T.5]                           0.0005      0.048      0.010      0.992      -0.094      0.095
C(Country)[T.6]                           0.0575      0.035      1.621      0.105      -0.012      0.127
C(Country)[T.7]                           0.0228      0.047      0.490      0.624      -0.069      0.114
C(Country)[T.8]                           0.0636      0.046      1.376      0.169      -0.027      0.154
C(Country)[T.9]                           0.0646      0.051      1.279      0.201      -0.034      0.164
C(Country)[T.10]                          0.0544      0.047      1.160      0.246      -0.038      0.146
C(Country)[T.11]                          0.0928      0.047      1.968      0.049       0.000      0.185
C(Country)[T.12]                          0.0961      0.047      2.063      0.039       0.005      0.187
==============================================================================
Omnibus:                       644.456   Durbin-Watson:                   2.006
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              164.811
Skew:                            0.017   Prob(JB):                     1.63e-36
Kurtosis:                        2.111   Cond. No.                         69.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

FIGURE 67: OLS Logistic Regression model to predict “Talent Profile”

Among the models performed, a Support Vector Machine is the last algorithm presented for this section. In figure 68, it’s been shown the accuracy score with respect to the test dataset, that, reaching 0,81, represents a good result even though not as performant as the precedent logistic regression. In the same way, the classification report and the confusion matrix have been implemented to give a more precise idea of how errors were distributed during the classification tasks carried out by the model, though, even in this case, hyperplanes could not be graphically defined.



```
In [22]:  # SVM

          from sklearn import svm

          clf = svm.SVC(kernel='rbf')
          clf.fit(X_train, y_train)

Out[22]:  SVC()

In [23]:  from sklearn.metrics import classification_report
          print(classification_report(y_valid, clf.predict(X_valid), zero_division = 0))

                        precision    recall  f1-score   support

                     0       0.89      0.34      0.49        50
                     1       0.81      0.86      0.83       332
                     2       0.80      0.80      0.80       332
                     3       0.82      0.86      0.84       259
                     4       0.80      0.74      0.77        27

              accuracy                           0.81      1000
             macro avg       0.82      0.72      0.75      1000
          weighted avg       0.81      0.81      0.80      1000


In [25]:  from sklearn.metrics import accuracy_score
          print('SVM Model accuracy score: {0:0.4f}'.format(accuracy_score(y_valid, clf.predict(X_valid))))

          SVM Model accuracy score: 0.8100


In [24]:  from sklearn.metrics import confusion_matrix

          confusion_matrix(y_valid, clf.predict(X_valid))
```

Figure 68: Basic SVM classifier model fit and relative classification report; "Talent Profile" prediction

Overall, within this broad analytical framework, the descriptive segments of analyses have enabled a deep focus on various trends of my dataset workforce contributing to reach counter-intuitive understandings based on data visualization tools and allowing for testing potentially biased HR initiatives currently underway. In second place. rather than keep developing mere reportstic activities, I implemented, for each HR areas of interests, a set of modeling predictive applications for the C-suite level to get aware of the most impacting employee factors on which to make predictions. Therefore, by knowing in advance how to intervene in their human population to improve their level of performance, engagement, retention and talent, they would consequently be in power to reap considerable overall business benefits.

3.4 Discussion and Implications

This final section concludes the research thesis by discussing the interpretation and evaluation of the HR analytics proposed model, highlighting its limitations and ultimately explaining some implications derived by the implementation of an HR analytics initiatives inside a company.

3.4.1 Model Evalutation and Interpretation

Overall, within the unfolding of the four sets of descriptive and predicted analyses presented in this research work, hypothetical HR managers of my invented company have been equipped with important insights on how start designing next strategic moves: considering the reliable machine-learning predictions performed, they have been enabled to envision HR areas of intervention on which to address future requirements as well as workforce's past and present detached problems. Addressing the model evaluation, in this work, I employed, with the use of many data-analytics packages and through an extensive data reviewing effort, five different machine learning algorithms for each of the four HR areas of interest taken into examination. Among the models performed, from the panoramic of the experimental analyses results, including the application of stratified cross-validation and Grid or Randomized Searches to shape hyperparameter tuning processes directed at improving their accuracy, it could be clearly seen that SVM, random forests and light gradient boosting models, have proven to outperform other classifiers in delivering their estimations. These have provided, in fact, the most accurate results after being trained with my simulated employee dataset. Generally, all the confusion matrices reported, showing good scores of metrics like precision, recall, and F1-scores as evaluation criteria, demonstrates that all the models were sufficiently efficient. In particular, out of the more than twenty-five models built, Support Vector Machines elaborated for the Job Involvement and for the Job Involvement predictions have obtained the best performance rate, reaching the beauty of 100% af accuracy. Globally, it could be then stated that the whole experimental part of this work adds a sense of credibility, reliability and validity to the entirety of the thesis thanks to the empirical findings achieved in conducting the front-line designed analyses; I the wish this attempt would incentivize the corporate world in changing its approach towards HR data treatment. Additionally, the goddess of the analyses results has been also greatly enriched by a robust methodology and process of analysis description that have given a logical sequence to the various analytical phases (from data construction, passing through HR techniques research reviews, to model building): its function was also to allow private single firms, owning similar data, to replicate the same framework of analytics endeavors and possibly get comparable results across other different contexts. Besides, as in performing the analyses it has not been followed an existing formalized HR analytics process, it's important to say something about the peculiarities of model interpretation. As a matter of fact, for delivering this experimental analyses, it has been required to collect a specific type and amount of data and to delve into previous analytics research and machine learning applications in HR context to understand how to most suitably build my analysis framework in accordance to the desirable outcome that I previously designed and that I wanted to reached. As an instance, it has been studies in the HR literature which were the most impacting factors for turnover trend, which then have been considered relevant in the process of sample construction for first generating, and then examining and predicting, the employee attribute "Turnover Risk"; all these theoretical assumptions based on a personal and critical literature review have therefore influenced the nature of the expected primary HR analytical results of the experiment. Moreover, in specific moments along the process of analysis, many analysis design choices and evaluation of intermediary outputs, have been subjected to the author's judgment to better envision which were going to be the next set of activities necessary to reach determinate HR analytical findings: look for new data, modify the underlying assumptions of certain data, run another algorithm, try improving the current model, and so on. An example of design choice coincided, other than with the construction of data above-mentioned, with how part of the analytical outcomes of descriptive analyses have eventually been deployed: HR dashboards have been presented in an aggregation data visualization structure rather than in detailed individual employee reports. Ultimately, in all these situations, domain-specific HR knowledge have resulted to be essential in fitting the required analytical procedure with the theoretical HR conjectures undertaken when considering the employee dataset.

3.4.2 Model Limitations

Analyzing the study's limitations, it could result complex for organizations to directly embrace this proposed model in their HR analytics application-frame for a number of reasons. In the first place, because in this project it has been used a restricted number of data sources and no data collection methods at all; the employee dataset analyzed was in fact created through a precise data sampling method, that, though remaining as much faithful as possible to realistic observations, can't totally reflect in all its aspects a real-world database of a firm. As a case in point, this data construction endeavor didn't give any indication about how any type of unstructured data could be effectively used in the machine learning predictions. These types of data are usually extracted by interview video-recordings, by workers' comments from the engagement and pulse surveys or by contents found in social media platforms, and are increasingly being leveraged by the corporate world, as they can provide valuable information on how the employees feel about the workplace. In addition, the dataset created was pretty balanced and perfectly suitable for the algorithms selected, that, with other different data, could not work the same way. Since in the real world it's not rare to find imbalanced HR datasets, the universality applicability of the work should be reasonably extended to address these gaps. Looking at the predetermined theoretical assumptions, this project could have investigated other more employee factors affecting most relevant workforce trends; in the same way, the objectives envisioned at the beginning of the experimental analysis design could have been directed to solve other various employee issues that have not been taken under examination throughout this analytical framework, such as recruitment&selection, training&developement, career planning and diversity trends, making the algorithmic application more comprehensive. Another limitations regards instead the sacrifice of some operational efficiency and overall prediction accuracy when it came to build a model or improve an existing one: in part of the hyperparameter tuning processes carried out, as a matter of fact, I didn't try on purpose to minimize the prediction error rate, using an exploratory "Grid Search", since it would have required way more processing power and higher cost in terms of time for completing the great number of possible combinations for each single parameter. This has worthed in particular for the random forest and lgb models, for which it has been applied a "Randomized Search" whose scope was to find the optimal hyper parameters of the model within some parameter ranges that I gave. Overall, then, this project would demand further exploration to possibly raise each single model's accuracy score. Concerning the descriptive section of analyses, moreover, this work could have been enriched and expanded with the application of the many advanced statistical techniques through the statistical programming language of R or MatLab: however, in this very framework of analysis, apart from the deployment of descriptive visualization tools, it has been purposely emphasized the predictive means of analysis, as HR departments, though still being technologically behind than others, already produce some descriptive reportistics within their limited HR analysis pipeline. Having said that, the last consideration, among the identified limitations based on the author's personal critique, corresponds to not having inserted the analytical findings into a "ROI"-based perspective framework in power to offer relevant measurement tools to assist shareholders with a more data-driven decision making. In particular, this different context could have prefigured a step-by-step procedure for companies to integrate their current or new HR Analytics application to their financial budgeting system, therefore keeping track and predicting the actual and future levels of earnings and savings in HR processes accomplished thanks to the meaningful managerial insights derived by the analyses. However, it must be clear that such an advanced approach based on financial measurement criteria, that could apparently augment the possibilities of practical adoption of HR Analytics application, would demand an active collaboration between HR domain experts and data analysts inside a firm.

3.4.3 HR Analytics Implications

The implications for a successful implementation of an HR analytics endeavor like the one it has been presented throughout this work are copious. First of all, analytical, and, in this case, machine-learning specific skills must be acquired in advance to assure the whole HR analytics process leads to effective and actionable results.

Necessarily, to gain valuable information from data, a HR practitioner taking on this HR Analytics' cause should be experienced enough to smoothly address some crucial activities regarding data collection, selection, pre-processing and transformation and to grasp insights to better solve other incoming data-related issues. These employees should be either specifically trained or researched in the labor market, making a visionary upskilling investment, to reach such a level of analytical confidence and embeddedness in the company. In addition, context-specific HR knowledge are necessary when it comes to treating people-related data. First, these skills allow analysts to interpret the results of the driven analyses in light of the surrounding HR company's setting: they should assess whether the analytical outputs match the targeted HR analytical issues addressed and decide accordingly if something in the analytical processes and activities should be modified to look for different information and people-related insights. Secondly, when it comes to inform company's management, not used to comfortably read visualization charts or algorithms' scores neither to understand data processes, communication's ability, other than HR competencies, are proven to be fundamental in effectively converting the analytical findings into the HR specific knowledge: closing the gap between computational schemes and HR domains using legitimate financial terms and adequate metrics, an overall improved decision-making is achieved as data findings have been appropriately contextualized and correlated to business outcomes.

Having ensured a sound dialogue between analysts and decision makers, another implication is about proactively enforcing HR analytics activities to better align with business objectives. In order to do so, it is essential to realize, as this thesis clearly suggests, that human capital trends greatly affect overall company performance, positing HR to a new business-core role: through HR analytics means, predicting employee trends, companies could gain insightful and strategic information on how to make HR policies at hand more individually tailored to mitigate forecasted workforce risks and leverage favorable opportunities. Moreover, concerning the sphere of ethical implications, while this work has been focused on an artificial employee dataset, in reality companies should put additional efforts in not risking to overcross workforce privacy limits; firms should enrich their HR analytics framework with appropriate and severe legal and societal devised terms that regulate the activities directed at acquiring and analyzing employees data. A special attention should be given to data extracted by external sources, like from social media employee profiles, that, without capturing a reliable perception of the employees lifestyles, could even offer findings leading to misleading decisions. Ultimately, then, companies, based on their employee findings, should first enact intervention planning made by an adjustment process of current policies and new focused initiatives implementations, like the introduction of a training programme. Subsequently, HR managers would develop impact assessment indicators, tracking workforce perceptions, behaviors and performance measures, by which evaluating if their interventions have been effective, paying off the investments.

Reflecting on the analytical results, from the theoretical point of view, companies could even possibly notice that from the application of such computational means of analyses, they would start questioning their basic HR theoretical assumptions: for example, investigating the employee engagement trend inside their organization, managers may arrive to data-driven results that would lead them to uncover challenges to their settled theoretical assumptions on this theme, modifying the way the company approaches to these issues.

Conclusions

The conclusive part of this research shall begin with a premise: the goal of such an experimental thesis, characterized by a resonate exploratory nature, was trying to draw an HR Analytics application pathway to use more advanced computational methods thus shifting the critical role played by HR management from securing efficiency to unlocking the strategic measurable impact of people–related factors to the organizational bottom line. Overall, it has been stressed how HR departments need to be supported by a data-driven organizational culture that allows them to become more intrinsically aligned with organizational strategies. My thesis offers then an analytical guiding prototype, bringing application cases of descriptive and, above all, predictive techniques and tools to improve the way HR managers interpret and read workforce Big Data, which are increasingly scaling as a powerful engine for any type of company. In particular, throughout this work I've not been developing any new computation techniques nor I've been focused on proposing innovating theoretical hypothesis by bringing into question fundamental theory-based constructs according to analytical findings obtained; rather, my only intention has been to enforce companies to adopt a more evidence-based approach to people management by showing, with some empirical case of analyses based HR-like sets of data, the potential of predictive machine-learning algorithms, and of other HR analytics standards. Unfortunately, nowadays, most of the academics developed for the evolving field of HR Analytics still hinge on a theoretical point of view with qualitative case studies, not providing any relevant insights on how to implement such emergent analyses and, consequently, without showing the real cause-effect association between the benefits gained and these analytical practices. This has created solid barriers for the application of these systems. Conversely, this work acts as a bridge between literature research and methodological practice, recommending the acceptance of such predictive computational methods by showing how effectively they perform in predicting employees' work-related features. As a matter of fact, the endeavor profused in my analyses was inserted in a close to real-world strategic context, necessary to guide HR practitioners to develop an analytical design application of similar scale. As such, companies should use the proposed methodological approach as a roadmap to evaluate their current HR Analytics operational framework or organize it from scratch, with the ambitious scope of adding critical value to their HR processes and aspire to reach a more enlightened management of their workforce. Ultimately, this thesis would then hopefully contribute to convince the corporate world to adopt HR analytics solutions for enabling a more data-driven decision-making in the context of HR initiatives. Company leaders, within their competitive and resource-constrained settings, can no longer keep ignoring the power of such advanced analytical tools and persist adjusting problematic workforce situations once they have already occurred: leveraging more advanced techniques to make predictions on the identified employee-related factors that mostly drive overall business performance, companies could acquire compounded competitive advantage in the near future.

BIBLIOGRAPHY

Analytics, M. (2016). The age of analytics: competing in a data-driven world. McKinsey Global Institute Research.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. Human Resource Management Journal, 26, 1-11.

Arellano, C., DiLeonardo, A., & Felix, I. (2017). Using people analytics to drive business performance: A case study. McKinsey Quarterly, 3, 114-119.

Bassi, L. (2011). Raging debates in HR analytics. People and Strategy, 34(2), 14.

Berral-García, J. L. (2016, July). A quick view on current techniques and machine learning algorithms for big data analytics. In 2016 18th international conference on transparent optical networks (ICTON) (pp. 1-4). IEEE.

Bersin, J., & Chamorro-Premuzic, T. (2019). New ways to gauge talent and potential. MIT Sloan management review, 60(2), 1.

Bhuva, K., & Srivastava, K. (2018). Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 5(3), 568-577.

Boudreau, J. W., & Ramstad, P. M. (2002). Strategic HRM measurement in the 21st century: From justifying HR to strategic talent leadership.

C. Derose, (2013), The Atlantic, How Google Uses Data to Build a Better Worker

Cappelli, P. (2017). There's no such thing as big data in HR. Harvard Business Review.

Chartered Institute of Personnel and Development, (June 2018), Global Research, People analytics: driving business performance with people data

Chartered Institute of Personnel and Development,(May 2017), Research Report, Human capital analytics and reporting: exploring theory and evidence

Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. Harvard business review, 88(10), 52-58.

DiClaudio, M. (2019). People analytics and the rise of HR: how data, analytics and emerging technology can transform human resources (HR) into a profit center. Strategic HR Review.

Dulebohn, J. H., & Johnson, R. D. (2013). Human resource metrics and decision support: A classification framework. Human Resource Management Review, 23(1), 71-83.

E. Ledet, K. McNulty, D. Morales, and M. Shandell, (2020), McKinsey&Company, How to be great at people analytics

Edwards, M. R., & Edwards, K. (2019). Predictive HR analytics: Mastering the HR metric. Kogan Page Publishers.

Fink, A. A., & Sturman, M. C. (2017). HR metrics and talent analytics. The Oxford handbook of talent management, 375-390.

Fitz-Enz, J. (2010). The new HR analytics. American Management Association.

Fitz-Enz, J., & John Mattox, I. I. (2014). Predictive analytics for human resources. John Wiley & Sons.

Gao, X., Wen, J., & Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. Mathematical Problems in Engineering, 2019.

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2021). A review of machine learning applications in human resource management. International Journal of Productivity and Performance Management.

Green, D. (2017). The best practices to excel at people analytics. Journal of Organizational Effectiveness: People and Performance.

Guenole, N., & Feinzig, S. (2018). The business case for AI in HR: With insights and tips on getting started. IBM Smarter Workforce Institute

Guo, X. Enterprise Human Resources Management Based on Cluster Analysis.

Gupta, R., Tanwar, S., Tyagi, S., & Kumar, N. (2020). Machine learning models for secure data analytics: A taxonomy and threat model. Computer Communications, 153, 406-440.

Hamilton, R. H., & Sodeman, W. A. (2020). The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. Business Horizons, 63(1), 85-95.

Harvard Business school publishing, (2017), HR ANALYTICS BUSTING SILOS AND DELIVERING OUTCOMES

Heric, M. (2018). HR new digital mandate. Digital technologies have become essential for HR to engage top talent and add value to the business. Retrieved August, 20, 2019.

J. Baier, J.l Caye, R. Strack, P. Kolo, A. Kumar, F. Ruan, B. Morton, A. Ariganello, J. Jauregui, L. van Wees, T. Burner, W. Wong , (2021), Boston Consulting Group, The Future of People Management Priorities

J. Ferrar, C. Styr, A. Ktena, (2020), Delivering Value At Scale, Insight222 People Analytics Program

J. Rocca, (2019), Ensemble methods: bagging, boosting and stacking. Understanding the key concepts of ensemble learning, TowardsDataScience

Jain, R., & Nayyar, A. (2018, November). Predicting employee attrition using xgboost machine learning approach. In 2018 International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 113-120). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

K. Waddell, (2016), The Atlantic, The Algorithms That Tell Bosses How Employees Are Feeling

King, K. G. (2016). Data analytics in human resources: A case study and critical review. Human Resource Development Review, 15(4), 487-495.

Lawler III, E. E., Levenson, A., & Boudreau, J. W. (2004). HR metrics and analytics–uses and impacts. Human Resource Planning Journal, 27(4), 27-35.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. Journal of Business Ethics, 160(2), 377-392.

Leonardi, P., & Contractor, N. (2018). Better people analytics. Harvard Business Review, 96(6), 70-81.

Levenson, A. (2011). Using targeted analytics to improve talent decisions. People and Strategy, 34(2), 34..

Levenson, A., & Pillans, G. (2017). Strategic workforce analytics, Corporate Research Forum, Concentra. Retrieved July, 3, 2019

Liu, L., Akkineni, S., Story, P., & Davis, C. (2020, April). Using HR analytics to support managerial decisions: a case study. In Proceedings of the 2020 ACM Southeast Conference (pp. 168-175).

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. The International Journal of Human Resource Management, 28(1), 3-26.

Martin, K. (2019). Ethical implications and accountability of algorithms. Journal of business ethics, 160(4), 835-850.

MHR International UK Registered office: Peterbridge House, The Lakes, Northampton NN4 7HB. No. 1852206 England, HR ANALYTICS: A PRACTICAL GUIDE TO THE FUTURE OF HR

Mishra, S. N., Lama, D. R., & Pal, Y. (2016). Human Resource Predictive Analytics (HRPA) for HR management in organizations. International Journal of Scientific & Technology Research, 5(5), 33-35.

Momin, W. Y. M., & Mishra, K. (2015). HR analytics as a strategic workforce planning. International Journal of Applied Research, 1(4), 258-260.

Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytics age. European Journal of Operational Research, 241(3), 583-595.

Nocker, M., & Sena, V. (2019). Big data and human resources management: The rise of talent analytics. Social Sciences, 8(10), 273.

Opatha, H. H. D. P. J. (2020). HR Analytics: A Literature Review and New Conceptual Model. International Journal of Scientific and Research Publications, 10(6), 130-141.

P. Gupta, A. Sharma and R. Jindal, (2016), WIREs Data Mining and Knowledge Discovery, Scalable machine-learning algorithms for big data analytics: a comprehensive review

Pape, T. (2016). Prioritising data items for business analytics: Framework and application to human resources. European Journal of Operational Research, 252(2), 687-698.

Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., & Lekakos, G. (2018). Big data and business analytics ecosystems: paving the way towards digital transformation and sustainable societies. Information Systems and e-Business Management, 16(3), 479-491.

Patel, S. (2017). Integrating Machine Learning Techniques for Big Data Analytics. International Journal of Advanced Research in Computer Science, 8(5), 2760-2763.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Decision Support Systems, 134, 113290.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 1-16.

Rasmussen, T., & Ulrich, D. (2015). Learning from practice: how HR analytics avoids being a management fad. Organizational Dynamics, 44(3), 236-242.

Reddy, P. R., & Lakshmikeerthi, P. (2017). HR analytics-An effective evidence based HRM tool. International Journal of Business and Management Invention, 6(7), 23-34.

Schwarz, J. L., & Murphy, T. E. (2008). Human capital metrics: An approach to teaching using data and metrics to design and evaluate management practices. Journal of management education, 32(2), 164-182.

Shrivastava, S., Nagdev, K., & Rajesh, A. (2018). Redefining HR using people analytics: the case of Google. Human Resource Management International Digest.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017, November). Evaluation of machine learning models for employee churn prediction. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 1016-1020). IEEE.

Sullivan, J. (2013). How Google is using people analytics to completely reinvent HR. TLNT: The Business of HR, 26.

T. Shin, (2021), Understanding Feature Importance and How to Implement it in Python. Learn one of the most practical data science concepts,TowardsDataScience

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15-42.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2008). Linear and logistic regression analysis. Kidney international, 73(7), 806-810.

Tursunbayeva, A., Pagliari, C., Di Lauro, S., & Antonelli, G. (2021). The ethics of people analytics: risks, opportunities and recommendations. Personnel Review.

Ulrich, D., & Dulebohn, J. H. (2015). Are we there yet? What's next for HR?. Human Resource Management Review, 25(2), 188-204.

V. Fernandez, E. Gallardo-Gallardo, (2020), TechTalent-Lab, Department of Management, Universitat Politècnica de Catalunya – BarcelonaTech, Terrassa, Spain, 162-186, Tackling the HR digitalization challenge: key factors and barriers to HR analytics adoption

W. Cascio and J. Boudreau, (2017), Journal of Organizational Effectiveness: People and Performance, human capital analytics: why are we not there?, Vol. 4 No. 2, 2017 pp. 119-126 H

Wang, L., & Alexander, C. A. (2016). Machine learning in big data. International Journal of Mathematical, Engineering and Management Sciences, 1(2), 52-61.

WHITE PAPER Osservatorio HR Innovation Practice, (2016), HR ANALYTICS & BIG DATA DRIVEN INNOVATION: COSA SIGNIFICA E COME IMPOSTARE UNA ROADMAP DI INNOVAZIONE

X. Wang, (2021), How to interpret and explain your machine learning models using SHAP values, Mage

Zhang, X. D. (2020). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). Springer, Singapore.

SITOGRAPHY

https://analyticsindiamag.com

https://jupyter.org/

https://www.anaconda.com

https://www.qlik.com/it-it/

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Wikipedia contributors. (2022, January 12). English Wikipedia. In
Wikipedia, The Free Encyclopedia. Retrieved 17:21, January 17, 2022, from https://en.wikipedia.org/w/index.php?title=English_Wikipedia&oldid=1065245002

www.analyticsvidhya.com