Enhancing Emergency Healthcare Delivery Through Predictive Modelling: A Machine Learning Approach

Francesco Mancini

30844686

Submitted in partial fulfilment of the requirements of the degree

of

MSc Information Management and Digital Business– Big data in Business

Henley Business School – University of Reading

September 2023

Abstract

The dissertation delves into the realm of healthcare data analysis, employing machine learning techniques to predict disease outcomes within Emergency Department (ED) settings. Its overarching objective is to enhance patient care, optimize resource allocation.

A diverse and extensive dataset, sourced from reputable healthcare repositories, forms the bedrock of this study. However, data validity presented its own set of challenges, notably inconsistent data formats. Rigorous data preprocessing and cleansing addressed these issues, ensuring the reliability and accuracy of subsequent analyses. Additionally, machine learning techniques were chosen and implemented with precision; despite these rigorous measures, however, the study is cognizant of the inherent potential for biases in healthcare data, rooted in healthcare access, socio-economic disparities, and regional practices.

Machine learning models exhibit formidable predictive capabilities, offering the promise of data-driven decision-making in healthcare. For binary classification, the prediction of the diagnosis associated with abdominal pain and Bayes-optimized Random Forest model emerged as the most powerful, achieving an F1-score of 0.87 for class 0 and 0.86 for class 1. Moreover, in the three-level multiclass classification, which involved predicting the diagnoses associated to abdominal pain and chest pain, again the Bayes-optimized Random Forest model again stood out. It achieved precision scores of 0.84 (class 0) and 0.85 (class 2) and a recall of 0.86 (class 0) and 0.81 (class 2). The F1-score reached 0.85 (class 0) and 0.83 (class 2), showcasing its effectiveness in classifying these diagnoses.

The research underscores several vital recommendations for future research. First, while machine learning models show immense potential, their practical implementation in real-world healthcare settings necessitates rigorous validation through collaboration with healthcare institutions. Comparative studies pitting these models against human clinicians can yield invaluable insights. Efforts to enhance model transparency and user-friendliness should remain paramount, fostering trust among healthcare practitioners. Ethical dimensions, including data privacy and bias mitigation, warrant deeper exploration to ensure responsible AI deployment in healthcare.

Future work offers exciting prospects. Deep learning models could elevate predictive capabilities, potentially encompassing a broader spectrum of diagnoses. Addressing data representativeness concerns by integrating external data sources is imperative to improve equity in predictions. Collaborative efforts with healthcare institutions are essential for validating model practicality. Designing intuitive user interfaces to facilitate human-AI interactions and prioritizing ethical considerations will shape the future of healthcare AI.

In conclusion, this research paves the way for data-driven decision-making in healthcare, where tailored care, efficient resource allocation, and improved patient outcomes are within reach. Its findings and recommendations lay a strong foundation for future work, guiding the transformation of healthcare practices and policies.

Table of contents

List of tables

List of figures

Introduction

Motivation of the Research

Healthcare, as one of the cornerstones of societal well-being, stands at a juncture in the modern era. With the advent of technology and the accumulation of vast amounts of patient data, there arises an opportunity to revolutionize the way we approach healthcare. The motivation behind this research lies in harnessing the power of data-driven insights to transform healthcare practices, making them more patient-centric, efficient, and effective.

In today's healthcare landscape, both patients and practitioners face myriad challenges. Patients seek personalized and timely interventions, while healthcare providers strive to allocate resources judiciously and improve overall healthcare outcomes. Yet, the complexities of healthcare, the diversity of patient populations, and the interplay of countless variables make achieving these goals a formidable task.

This research aims to address these challenges by employing advanced machine learning techniques to predict disease outcomes. The underlying premise is simple: by leveraging data, it can be possible to gain a deeper understanding of disease patterns and develop tailored interventions. Such an approach has the potential to significantly enhance patient care, optimize resource allocation, and shape healthcare policies for the betterment of society.

Background of the Research

Healthcare has long been a dynamic and multifaceted field, driven by the constant pursuit of improving patient outcomes, reducing the burden of disease, and enhancing the quality of care. The history of healthcare is marked by remarkable advances in medical science; however, nowadays the transformation of healthcare has extended beyond medical breakthroughs, and it encompasses the way data are collected, analysed, and used for clinical decision-making.

In the early days of modern medicine, healthcare was largely a discipline guided by experience, intuition, and a limited understanding of diseases. Clinical decisions were informed by a physician's training and expertise, while this approach yielded valuable insights, it was inherently limited by the human capacity to process vast amounts of data and discern subtle patterns.

As healthcare data became increasingly digitized, thus, the potential for data analytics and machine learning in healthcare emerged. Shifting the paradigm to the development of data-driven healthcare applications, ranging from predictive analytics for disease diagnosis to personalized treatment recommendations.

Today, the healthcare landscape is characterized by an explosion of health data, generated not only in clinical settings but also through wearable devices, genetic testing, and patient-reported outcomes representing both challenges and opportunities.

In this context, this research seeks to contribute to the ongoing evolution of healthcare by leveraging advanced machine learning techniques to enhance disease prediction, and, from a broader point of view, healthcare resource allocation. This work aligns with the broader movement towards evidence-based, patient-centric care, offering a glimpse into the future of healthcare, where data-driven insights empower clinicians and benefit patients.

Aims and Objectives of the Research

This research aspires to serve as a bridge between the realms of healthcare and data science, with the aim of harnessing the potential of advanced machine learning techniques.

The primary objective is to develop robust predictive models for disease outcomes. By scrutinizing healthcare data, research aims to uncover patterns and risk factors associated with various medical conditions. These predictive models are intended to be practical tools for healthcare practitioners, facilitating early disease progression prediction, treatment optimization, and ultimately, improved patient outcomes.

The goal imagined for the potential implementation of the tools is to offer personalized treatment recommendations, driven by patient-specific data, including demographic information, clinical history, and other personal factors.

These objectives serve as guiding principles for this research, grounded in the substantial potential of data-driven healthcare transformation.

Structure of the Dissertation

The dissertation follows a structured format designed to facilitate a systematic understanding of the research processes, methodologies, and findings.

In this opening chapter, the research's motivation has been explored, emphasizing the significance of data-driven healthcare in contemporary society, and providing a backdrop for the study by delving into the context of healthcare data analysis and the role of machine learning.

The following chapter is dedicated to literature review delving into a comprehensive examination of the healthcare landscape, with a particular focus on the primary care structure in the UK and the intricate operations of Emergency Departments (EDs). Additionally, is analysed existing research in healthcare data analytics, examining various methodologies and their applications.

Third chapter, moreover, provides an in-depth look at the research methodology offering insights into the data sources used in the research, highlighting their quality and relevance. Moreover, it explains the feature engineering methods applied to enhance the predictive capabilities of the models. Additionally, it discusses the selection and implementation of machine learning techniques, considering their relevance to the healthcare domain. Furthermore, it explains the model validation techniques employed to assess the reliability of the developed models.

Fifth chapter presents the research outcomes, showcasing the specificities of the data employed in model building and the performance of predictive models in disease outcome prediction and is followed by the chapter in which research's validity is critically assessed, covering data validity, methodological rigor, model performance, and ethical considerations. The discussion chapter, moreover, evaluates the research's potential impact on healthcare practices and discusses the research's limitations, highlighting areas for future work.

In conclusion the final chapter summarizes the entire research and concludes by summarizing the research's contributions to healthcare and provides recommendations for future studies.

Literature review

Background

A model for public health services shows a concerted effort to protect and enhance population health with the goal of providing the greatest possible service to the people. All citizens have access to high-quality medical care regardless of their financial situation thanks to political-legislative actions, programmes, and strategies targeted at public health forums as well as by creating organisational frameworks that would facilitate the delivery of medical services that the public has expressed a need for (*Bunaciu*, 2016). Comparing the advantages of public and private healthcare systems also showed that the former fosters a healthy workforce, lowers future costs, and influences people to make wiser decisions. It also lowers overall healthcare and administrative costs and helps standardise services (*Kaabi et al.*, 2022). In the UK National Health Service (NHS), which is made up of a large variety of groups specialised in various sorts of patient services, as shown in *Figure 1*, is charged with managing the country's healthcare system and every 36 hours treat more than 1 million patients (*Department of Health and Social Care*, 2013), and over the last 15 years, there has been a 40% increase in the number of patients visiting emergency departments (EDs) (*NHS England,* 2020).

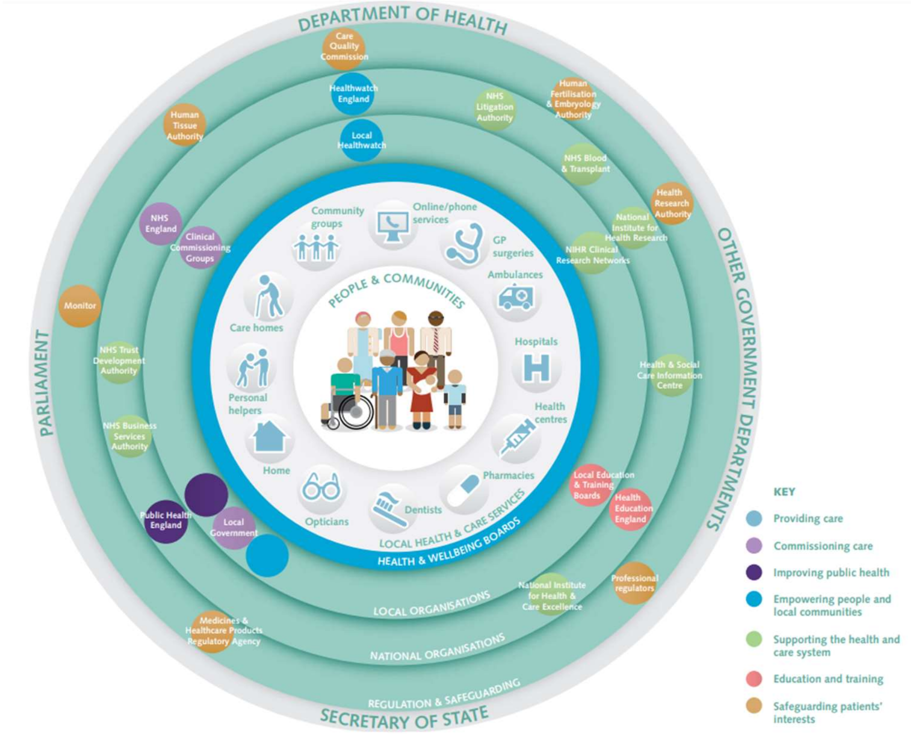

DEPARTMENT OF HEALTH
Care Quality Commission
Healthwatch England
Local Healthwatch
Human Tissue Authority
NHS Litigation Authority
Human Fertilisation & Embryology Authority
NHS Blood & Transplant
Health Research Authority
National Institute for Health Research
NIHR Clinical Research Networks
NHS England
Clinical Commissioning Groups
Monitor
NHS Trust Development Authority
NHS Business Services Authority
Health & Social Care Information Centre
PARLIAMENT
OTHER GOVERNMENT DEPARTMENTS
Community groups
Online/phone services
GP surgeries
Ambulances
Hospitals
Health centres
Pharmacies
Dentists
Opticians
Home
Personal helpers
Care homes
PEOPLE & COMMUNITIES
LOCAL HEALTH & CARE SERVICES
HEALTH & WELLBEING BOARDS
Local Education & Training Boards
Health Education England
Public Health England
Local Government
National Institute for Health & Care Excellence
Professional regulators
Medicines & Healthcare Products Regulatory Agency
LOCAL ORGANISATIONS
NATIONAL ORGANISATIONS
REGULATION & SAFEGUARDING
SECRETARY OF STATE
KEY
Providing care
Commissioning care
Improving public health
Empowering people and local communities
Supporting the health and care system
Education and training
Safeguarding patients' interests

Figure 1: The healthcare system in England (Department of Health and Social Care, 2013)

Structure of primary care in UK

The first requirement that the NHS must address is the increase in circumstances where citizens have access to professional healthcare staff (*NHS*, 2019). The primary care process typically entails people signing up for services and scheduling appointments with a general practitioner (GP) of their choice who offers a wide range of medical services, including: making diagnoses, providing treatments or referrals to specialists when necessary, prescribing medications, immunisations, screenings, mental health support, and managing chronic diseases. In fact, in emergency situations, individuals need immediate access to health services offered by the NHS, requiring them to interact with many entities both physically and digitally.

First, by dialling 111 or 999, the person has the option of using two services that are accessible whenever needed. The former offers help for circumstances that are not immediately life-threatening and is for non-emergency medical advice and guidance (*NHS UKb*, no date). The latter is only used in cases of sudden, life-threatening crises where prompt medical care or assistance is needed (*NHS UKa*, undated).

If the citizen's need has not been met and, therefore, the intervention of specialised personnel is required, the citizen has the option of accessing physical structures designed to handle such emergency situations, which are specifically referred to, according to the variety of situations they can handle, as Urgent Treatment Centres (UTCs) and Accident & Emergencies departments (EDs).

The former are GP-led facilities that are open at least 12 hours a day, seven days a week, and offer consultations that can be scheduled through 111 or with a GP referral. They are equipped to recognise and treat many of the most common illnesses for which patients visit EDs; they are designed to relieve hospital strain so that other components of the system can concentrate on the most urgent situations (*NHS England*, no date). A multidisciplinary team of doctors, nurses, paraclinical practitioners, and administrators work in the latter, which is a 24/7 healthcare facility equipped to provide comprehensive emergency care and operates in a fast-paced environment with unpredictable patient volumes and decision-making (*Seow*, 2013); given that many people who are unsure of where to turn when they need urgent care or guidance make coming there their first choice, it reflects the foundation of the primary care services offered by the NHS (*NHS England*, no date).

Operation of an Emergency department

Most patients who visit the ED are unselected when they arrive (*RCEM*, 2017), even though this tendency has changed since the COVID 19 pandemic given that the hospitals have increased the way of access to the EDs due to increased infection prevention and control increasing the proportion of heralded patients (*NHS England a*, no date). Some patients may have already seen a doctor or have been directed to the department after a pre-hospital evaluation like NHS111. Thus, EDs receive two different types of patients: those who are unheralded and so arrive at the facility voluntarily and those whose arrival is announced by ambulance operators or through prior interactions the patient has had with other components of the healthcare system. Regardless of the type of patient the facility happens to receive, its primary goal is to manage their clinical care.

According to ED guidelines, patients must undergo a first assessment process once they arrive, ideally within 15 minutes, that involves a quick evaluation of their condition in accordance with regionally approved practises. Patients with planned appointments through NHS 111 at a co-located UTC are exempt from further examinations if seen within 30 minutes of their appointment time, unless their health has gotten worse. The patient's principal complaint or probable diagnosis and the severity of their disease serve as the two key guiding principles for the examination.

Acuity denotes the severity and urgency of the patient's condition; in particular, if no formal triage system is utilised, acuity is given a score between 1 and 5, and the patient is then guided to the proper categories (such as resuscitation, majors, or minors) based on the implicit triage provided in *Table 1*.

Table 1: Implicit triage after initial assessment (NHS England, 2022)

Acuity score	***ED coding***
1 Immediate emergency care	1 Resuscitation
2 Very urgent emergency care	3 Majors
3 Urgent emergency care	4 Minors
4 Standard emergency care	
5 Low acuity emergency care	

Initial assessments are designed to help patients as they go through the healthcare system, thus even while initial assessment models might differ throughout the nation, they are all built to:

- Ensure that patients with the most serious illnesses receive the highest priority care.
- Properly assess accidents and illnesses that don't pose a life-threatening threat to make sure that these patients receive the required attention and are seen by the right clinician in a timely manner.
- To aid in the prevention and management of infections, do not overcrowd emergency rooms.
- Identify patients who could be vulnerable and take urgent safety concerns into account.

Furthermore, *Figure 2* highlights how professional decisions made during the initial stage of a patient's treatment can result in them being sent off-site to a suitable service, or staying in the hospital to use the ED or another service, like a UTC, when there is one co-located. This is because the initial assessment can include streaming, triage, Rapid assessment and Treatment (RAT), among other things.

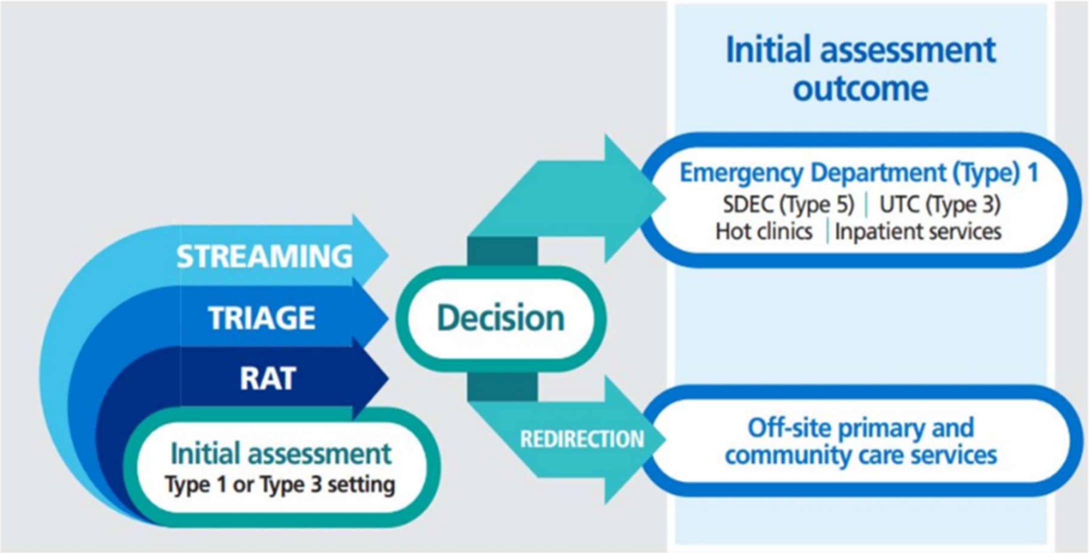

Initial assessment outcome
STREAMING
TRIAGE
RAT
Initial assessment
Type 1 or Type 3 setting
Decision
Emergency Department (Type) 1
SDEC (Type 5) | UTC (Type 3)
Hot clinics | Inpatient services
REDIRECTION
Off-site primary and community care services

Figure 2: Initial assessment flow (NHS England, 2022)

When a patient enters the ED, the first clinical activity they see is streaming, which aims to quickly guide them to the most appropriate service based on their symptoms, primary complaint, and acuity. This procedure usually entails a quick clinical evaluation that includes the patient's medical history and certain fundamental observations. Its results could range from additional evaluation inside the ED (resuscitation room, majors, minors), or UTC, to transfer to other services like a Specialty Diagnostic and Emergency Centre (SDEC) or specialty assessment units (medical, surgical, gynaecology, children's and other), or off-site rerouting.

In order to manage demand and flow throughout the healthcare institution, triage is also the clinical process of prioritising patients and is a key step undertaken before completing a comprehensive examination. Triage can be done on its own or following the first streaming of patients upon arrival and involves a comprehensive face-to-face assessment that includes making observations and using the available triaging tools to support the decision-making process. If streaming hasn't already happened, it becomes the patient's initial evaluation. It often takes longer than streaming and assigns a priority to each patient. Ideally, it should start as soon as feasible after the patient arrives.

Instead, the sickest patients typically receive rapid assessment and treatment (RAT), a thorough initial evaluation technique that combines streaming and triage methods. Because RAT typically takes longer than streaming or triage, frequently requiring 20 to 30 minutes per patient, insufficient resources may result in lineups.

Actual situation of EDs in UK

Due to the COVID19 pandemic and increased demand for flu medications peaking at the same time, EDs have faced the most difficult period in NHS history due to a perfect storm of pressures affecting the entire healthcare system and causing problems at the front door. Nineteen out of every twenty beds are occupied in hospitals, which are busier than they were prior to the epidemic. Additionally, compared to pre-pandemic levels, up to 14,000 beds are occupied by patients who are clinically ready to go, and the number of the most urgent ambulance calls has occasionally increased by one third (*NHS England*, 2023). Workers have also been impacted by these expectations, having to work in an environment that is more demanding; as a result, the service's survival has been threatened.

Increased duration of stay in EDs provides serious obstacles to patient flow, as schematically depicted in *Figure 3*, and has a substantial impact on the efficiency of the healthcare system. Long-term stays in the ED make it more difficult to diagnose and treat new patients promptly and worsen congestion, which has a cascading effect on patient care delays, lengthened wait times, and reduced patient safety (*GIRFT*, 2021). In order to reduce the burden on the healthcare system, improve patient flow, and guarantee prompt and effective emergency care delivery, it is crucial to address the variables that contribute to lengthy ED stays (*NHS England*, 2023).

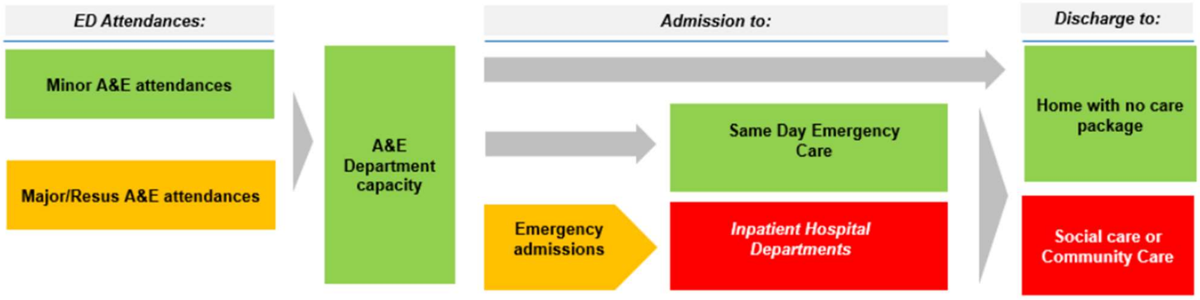


Figure 3: EDs patient flow (NHS England, 2023)

Triage process in emergency services

The triage phase acts as a crucial turning point, as is seen from the examination of the structure of primary care in the UK. But when there are delays in the triage procedure, it affects the whole ED, slowing down patient flow and delaying prompt access to the right care. Therefore, a detailed investigation of the current triaging procedures is required given that the study's goal is to develop a tool that specifically operates at this stage of the process to increase its efficiency.

Traditional triage methods

Traditional triage techniques have a long history, are based on staff expertise, and are still crucial to hospitals in the process of prioritising patients' medical requirements when they arrive (*Wang et al*., 2022). In order to get around challenges like the high demand for in-person clinical services, hospital personnel, such as nursing staff, sometimes do this clinical evaluation approach manually (*Sánchez-Salmerón et al.*, 2022). However, it is also done online (*Eccles et al*., 2019) or remotely (*Boggan et al*., 2020). Triage systems have been standardised in the majority of advanced nations with efforts made to ensure uniformity in execution being crucial for the effective operation of modern emergency rooms (*FitzGerald et al.*, 2010).

As a result, a number of assessment scales based on the expertise of the professionals have been developed to determine the levels of urgency, such as the *Australasian Triage Scale* (ATS), which has five tiers that link patient's history, symptoms, and signs to the clinical urgency and the maximum amount of time that patients should wait (*Hodge et al.*, 2013); or the *Manchester Triage Scale* (MTS) which is a clinical risk management tool used by clinicians worldwide based on a five step process that starts from identifying the problem and arrives to the monitoring of the implementation of the selected alternative and to the evaluation of the outcome (*Mackway-Jones, Marsden and Windle*, 2014).

In addition, a number of early warning instruments, such as *Early Warning Scores* (EWS) (*Nagarajah et al*., 2022), *Modified Early Warning Score* (MEWS) (*Subbe et al*., 2001), and *National Early Warning Score* (NEWS) (*Smith et al*., 2013), are used in EDs to identify patient deterioration during emergency care. Due to their capacity to escalate patients, these systems have generated criticism and cannot be entirely relied upon for ED triage (*O'Neill et al*., 2021).

Data driven triage methods

The use of deep learning and machine learning for triage applications has significantly changed from expert-driven to data-driven strategies as a result of the advancement of artificial intelligence (*Wang et al*., 2022). In fact, *Sánchez-Salmerón* et al. (2022) found that machine learning models may outperform traditional methods in several ED triage scenarios. In particular, *Joseph et al*. (2020) discovered that even with limited data, deep learning algorithms offer a potential method to enhance triage in the scenario of 24 hour mortality prediction, but also demonstrated high predictive power in screening patients at risk of early and short-term mortality (*Klug et al*., 2020) and outperformed conventional tools in predicting critical care and hospitalisation admissions (*Kwon et al*., 2021).

To categorise the severity grades of ED patients, for instance, Nave Bayes and C4.5 algorithms were used (*Zmiri et al*., 2012). Similar to this, *Teubner et al*. (2015) employed logistic regression to predict inpatient mortality using information acquired at the ED's point of triage. Machine learning paired with NLP algorithms have also been developed to anticipate patient disposition, optimising resource allocation inside the ED using emergency triage notes (*Tahayori et al*, 2021).

Deep learning has also been used to forecast ED hospitalisations (*Arnaud et al*., 2020) and researchers have utilised machine learning to predict ED wait times, medical needs, and the main complaints of patients. Based on the patients' conditions and the nurses' descriptions when they arrived at the ED, support vector machine was used to predict patients' primary complaints during triage (*Jernite et al.,* 2011). *Sterling et al.* (2020) used machine learning to predict resource needs in the ED using nursing triage notes and clinical data from the electronic health record. A decision tree method has also been created to predict how long patients will stay in the ER (*Azari, Janeja, and Levin*, 2015), and NLP techniques have been applied to nursing triage notes to predict how patients would be handled in the emergency department (*Sterling et al*., 2019)

Conclusion

Better data and technology use has the ability to improve health by raising the calibre and decreasing the cost of health and care services. It may provide carers more control over their patients' health and welfarereduce the administrative burden on healthcare professionals, and encourage the development of novel pharmaceuticals and treatments (*National Information Board*, 2019).

Throughout this literature review, it has been examined the intricacies of clinical processes that patients undergo upon arrival at an ED giving a context to the aim of the project which is to help streamlining this process by predicting, whether they may exhibit symptoms of a particular disease in order to reduce the time required for more complex triage procedures developing a tool that can harness available data to expedite the triage process within EDs to address the absence of previous studies attempting to predict the onset of specific diseases.

The goal is to develop a predictive model that can accurately identify potential diseases or conditions by utilising the wealth of data generated during the triage procedure, including patients' medical records, complaints, and history, in order to equip healthcare professionals with the tools they need to provide patients in need with quicker, more effective, and more accurate care. The basic goal of this strategy is to change healthcare delivery through the application of data analytics and digital health.

Methodology

Introduction

This chapter outlines the methodologies employed in the development and application of predictive models with the objective of providing a clear understanding of the processes employed to derive valuable insights from real-world healthcare data and construct predictive models.

The first section addresses the initial steps related to data acquisition and preparation. Accessing and understanding the sources and quality of healthcare data has been essential for constructing reliable predictive models and collaborative partnerships with healthcare institutions are discussed as a means of obtaining access to real-world data. The focus then shifts to data cleaning, refinement, and feature selection, ensuring that the resulting datasets are suitable for analysis.

The subsequent part of the methodology delves into feature engineering, the process of transforming raw data into meaningful variables for predictive modelling. Specific attention is given to the creation of categorical variables, including patient demographics and visit characteristics, using techniques like one-hot encoding and dimensionality reduction. For numerical variables, the methodology covers the identification and handling of outliers to enhance data quality.

The concluding section of this chapter explores the methodology for constructing and evaluating predictive models and the employed approach for binary and multi-class classification tasks. It begins with the development of baseline models, such as the naive Bayes classifier, for benchmarking purposes and covers hyperparameter optimization using grid search and Bayesian methods. Finally, it covers the evaluation metrics employed to assess model performance.

Methods of data collection

Collaborative Partnership and Data Access

The project's foundation relied on a collaborative effort with the NHS Royal Berkshire Foundation Trust, which offered access to their databases through an honorary contract. The Trust's business intelligence team provided valuable insights into the vast pool of data recorded at the Reading City Hospital ED. To effectively harness the potential of this data, a dedicated period of onsite visits was undertaken to understand the structure of their Microsoft SQL database to the extend needed for performing the analysis allowing to perform the crucial data extraction procedure and access the wealth of real-world data necessary for the project.

Data Selection and Key Aspects

Once access to the Trust's databases was secured, the next critical step was selecting the most relevant features to achieve the project's objective: constructing a predictor capable of forecasting ED diagnosis for each patient.

The information chosen centred primarily on three crucial aspects:

1. Demographic Details:

- Demographic information provided crucial context for understanding patient characteristics.

2. Presenting Complaint

- Recording the patient's presenting complaint upon arrival at the ED was considered as a significant indicator for potential diagnosis.

3. Data for Target Prediction

- This included the diagnosis for the current ED visit, the HRG band serving as a proxy for the acuity level, and healthcare provider services such as arrival mode or discharge status.

Initial Data Cleaning and Refinement

The resulting datasets, once the essential features were identified, have been too big and raised concerns about privacy and technological issues for remote work. To address this, an initial data cleaning step was implemented, which involved eliminating null values and keeping only those observations with a single registered diagnosis.

This decision aimed to avoid complexities associated with multiple diagnosis given to a patient at the end of the ED triage process. While the possibility of prioritizing different registered diagnosis was considered, it was decided not to pursue this route. The primary concern was the potential for clinical mistakes arising from giving more importance to the wrong category of diagnosis or overlooking critical medical domain knowledge linkages between diagnosis that might not be apparent from an analytics point of view but were crucial from a clinical perspective.

Maintaining a comprehensive and unbiased dataset ensured that the predictor's predictions would be robust and medically acceptable.

The Resultant Datasets

Following the data cleaning and refinement, the final datasets became more manageable and conducive for further analysis; more than one dataset is mentioned, indeed, because information relating to the first two aspects listed above was extracted from a table which was subsequently merged with another table containing information relating to the last of the aspects listed above. This marked a crucial milestone in the model development process, as it provided a solid foundation for subsequent steps.

Methods of data pre-processing and feature engineering

Dependent variable

The initial phase of the analysis revolved around comprehending the primary target variable for the model, specifically the recorded diagnoses attributed to each visit made to the ED, referred to as "*Diagnosis Codes Concat*". This examination was aimed at unravelling the intrinsic nature of this variable and identifying any underlying patterns it may exhibit.

As predicted, it became clear throughout the analysis of this variable that it contained a high number of diagnostic types. Consequently, a pivotal decision needed to be made concerning the specific diagnosis focus. To navigate this choice judiciously, careful consideration was given to the frequency of occurrence for each distinct diagnosis. The selection process aimed at identifying the clinically significant diagnoses that manifested themselves with the highest frequency to select them as target for the analysis.

This comprehensive analysis highlighted a pronounced imbalance within the dataset, presenting a notable challenge that necessitates diligent attention during the model construction phase. Addressing this issue becomes imperative to develop a robust model capable of effectively managing the inherent data imbalances. Consequently, the development of appropriate sampling strategies emerges as a crucial step to mitigate the impact of this data disparity throughout the model development process.

By proactively acknowledging and conscientiously addressing this data imbalance challenge the model's ability to accurately predict and classify future instances will be fortified, thereby enabling it to perform optimally in real-world scenarios.

Independent variables

Upon completing the analysis of the dependant variables, the focus shifted towards the pre-processing of the independent ones, which were categorized into two main groups: numerical variables and categorical variables.

Categorical Variables

Given the nature of the information available in the hospital's databases, a huge portion of the independent variables fell into the latter category. In fact, when patients arrive at the ED, the recorded data primarily capture qualitative characteristics, including demographic and clinical information. Additionally, are collected, data relevant for visit categorization and clinically significant details, such as the mode of arrival at the hospital (represented by the variable "*ARRIVAL_MODE*") and the visit's disposition (captured by the variable "*ATTENDANCE_DISPOSAL*").

To effectively manage categorical data and ensure compatibility for subsequent analysis and modelling tasks, a general approach of one-hot encoding was adopted for these variables. One-hot encoding is a data transformation process that allows machine learning models to process categorical data more efficiently. It simplifies the representation of complex categorical variables and facilitates the utilization of mathematical and statistical techniques that require numerical inputs. This technique converts categorical attributes into binary vectors, where each category is represented by a unique binary bit. As a result, distinct categories are no longer ordinal but binary, allowing the model to treat each category independently during analysis and ensuring the absence of any false sense of ordinality.

Additionally, it has been chosen to perform one-hot encoding because it avoids introducing any unintended biases that could arise if categorical variables were treated numerically. Moreover, it ensures that the model's calculations and predictions are not skewed or influenced by any arbitrary numerical assignments to categorical attributes.

Firstly, therefore, it must be highlighted the process of handling of the information load relating to the clinical history of the patients. The first obstacle, in fact, was found in the construction of the original table and, therefore, in the way the data were recorded and stored by the hospital. The table, in fact, presented an observation for each comorbidity associated with each individual patient both in relation to when this was recorded in the system and for each change of status, which represented a major problem for the development of the project as it gave rise to a large quantity of duplicate observations that had to be eliminated.

However, the difficulties associated with the management of this data did not end with the management of the duplicates; in fact, the structure of the table itself had to be modified, as it originally presented a structure that could be described as 'problem-wise'[1] (i.e., presenting one observation per problem) in contrast to the 'patient-wise' structure required for the development of the model. This required aggregating the comorbidities related to each individual patient in single patient-related rows.

Once a usable table was obtained for the purposes of the project, however, the analysis showed that there was not a reasonably limited number of definitions that physicians could use to describe the comorbidities relating to each patient, but, probably for reasons of clinical necessity and physician freedom, more than twenty thousand different comorbidities were enclosed in the dataset.

Considering, therefore, this to be an excessive quantity of unique values to be able to carry out a process of one hot encoding, the first approach attempted was to try to reduce the dimensionality of the variable by

[1] '*Problem*' is the term used in Royal Berkshire Hospital databases for persistent comorbidities over a significantly long period of time related to each patient.

investigating whether there were possible typos among the less frequent values by calculating the similarity between the various instances of the observations. Although there are numerous calculation methods often used to compare different texts, based either on syntactic similarities, such as *Nissim and Markert* (2003) or *Huang et al.* (2019), or on semantic similarities, such as those proposed by *Rozeva and Zerkova* (2017) or *Martinez-Gil* (2012), to address the specific data at hand, for which there was no material to use a semantic-based approach, being characterised exclusively by strings composed of single words, it was decided to use the method developed in the *'fuzzywuzzy'* project (*fuzzywuzzy*, 2020), which is based on the calculation of the Levenshtein distance to assign a similarity score to each term . This process, however, did not produce the desired result as the reduction in the dimensionality of the variable was negligible.

Thus, after analysing various possible methodologies, such as that proposed by *Dahouda and Joe* (2021) to use an embedding technique to carry out feature encoding or those proposed by *Shyu et al*. (2005), both for reasons of computational complexity in relation to the instruments available and for reasons of comprehensibility of the model it was decided to select only a sub set of problems to submit to the procedure of one hot encoding. Specifically, therefore, it was decided to select the hundred most frequent comorbidities for each of the diagnoses to be predicted by the model. It must also be specified that in order not to completely lose the information load on those comorbidities not selected for encoding, it was decided to summarise them in counter variables that considered the number of comorbidities recorded for each individual patient for each individual status (i.e., 'Active', 'Cancelled', 'Inactive', 'Resolved').

Finally, it should be made explicit that null values were handled differently for these data than the drop performed for the data on demographic characteristics and visit characteristics; in fact, due to the functioning of the 'LEFT JOIN' function of the SQL language used to extract the data, null values were generated for those patients who had no comorbidity recorded and, so, were therefore replaced with the value 'no previous problems'.

Moving on with the analysis, therefore, we approached the pre-processing of the categorical variables relating to the demographic characteristics of the patients and the characteristics relating to their visits to the hospital ED.

Specifically, we first dealt with the data collection structure in the hospital's databases. In these databases, in fact, each characteristic related to each of the aspects considered is recorded in two diverse ways. Firstly, in fact, these characteristics are coded using alphanumeric codes and, furthermore, associated with each of the variables containing alphanumeric codes is recorded an explanatory variable containing the description of the code itself.
Thus, the structure of the extracted dataset presented pairs of columns characterised by the suffixes *'_CODE*' and '*_DESC'*, (i.e., *'A_AND_E_ATTENDANCE_CATEGORY_CODE', 'A_AND_E_ATTENDANCE_CATEGORY_DESC'*).

To manage this type of dichotomy, therefore, the first step was to select the pairs of variables referring to the same factor of interest and to perform a check that for each alphanumeric code relating to the specific aspect there was a univocal description to be sure that the two variables provided the model with the same information load. Therefore, once the existence of this peculiarity had been acknowledged in order to avoid multicollinearity problems, that could compromise the performance of the model (*Leeuwenberg et al.*, 2021), it was decided to keep only one of the two variables and to proceed with the one-hot encoding procedure; specifically, it was decided to keep the variable relating to the description of the recorded phenomenon in order to safeguard the interpretability of the model.

The pre-processing of the categorical variables was therefore concluded with the analysis of the pairs of variables relating to a single characteristic that did not present the same number of alphanumerical codes and descriptions.

For each of these variables, ad hoc considerations were made to obtain a one-hot encoding process that could make the most of the information load.

Special mention must be made of the methodology used to approach the analysis of the variable concerning the complaints presented by patients on arrival at the ED. Here too, as in the case of comorbidities, we found ourselves having to deal with data in the form of words that could not be individually one-hot encoded to avoid exponentially increasing the granularity of the data with the risk of there being a negative effect on the effectiveness of the model. The process applied, therefore, was the same as that used with comorbidities in the first place. Specifically in this case, an approach was implemented aimed firstly at understanding whether instances recorded a small number of times could be considered as typing errors due to errors attributable to the conditions under which the data is entered into the system. To do this, therefore, the Levenshtein distance was calculated, again using the computational tools provided by the *fuzzywuzzy* library, to identify the cases in which this exceeded an empirically defined threshold such that the assumption that the observation was a typo was acceptable.

Finally, in this case, it can be said that this methodology resulted in a significant reduction in the dimensionality of the variable, which, combined with the aggregation of all observations that were not typo and had a frequency of less than five into a single category called 'residuals', allowed us to proceed with the one-hot encoding without undermining the robustness of the model by causing excessive granularity of the data.

Numerical variables

In concluding the analysis of pre-processing methods for the independent variables, it is crucial to discuss the approach employed for handling numeric variables. These variables constitute a minority within the dataset, encompassing only the age of patients upon arrival at the ED and the duration of their stays.

The analysis of these two variables primarily focused on understanding their distribution and identifying potential extreme values that could impact the performance of the constructed models. Concerning the patients' age, it was recorded in two different variables, but upon verification, it was found that the values reported in both columns were consistent for each observation so one of them has been dropped to avoid multicollinearity problems.

Conversely, when examining the values related to the length of stay at the EDs, the analysis revealed the presence of extreme values. These values were deemed implausible and unrealistic, suggesting errors in the databases due to mistakes made by data entry operators.

Methods of model building and evaluation

After establishing the data handling process for model input, the next step is to outline how the predictive models were crafted.

Initially, various model types were explored to address different classification tasks. This progression began with the creation of a binary classifier aiming to predict the most prevalent diagnosis in the dataset as accurately as possible. Eventually, it was developed a model encompassing four classification classes, including the null class and the three most frequent diagnoses in the dataset.

To accomplish this, the data underwent an initial transformation to create a categorical target variable, which would indicate the presence or absence of the target diagnoses for prediction. Specifically, a distinct target variable was generated for each classification task:

- For binary classification, a binary target variable was created.
- In the case of three-level classification, the target variable was defined with values of one for the most frequent diagnosis, two for the second most frequent, and zero for all other values.
- In the four-level classification, a level '3' was introduced within the target variable to represent observations related to the third most frequent diagnosis in the dataset.

However, as noted in the analysis of the dependent variable, it is crucial to address the non-negligible imbalance observed in all the created target variables. To tackle this concern, were employed sampling techniques, specifically under sampling, to enhance the performance of our models, as advocated by *Wah et al.* (2013). Simultaneously is acknowledged that oversampling techniques are an option, but it has been decided to not implement them because they are generally more suited to small datasets (*Domingues et al.,* 2018) and, furthermore, because the generation of new minority samples often fails to capture the intricate causal relationships between the features essential for disease diagnosis, as highlighted by *Luo et al.* (2021).

So once the data had been randomly extracted from the majority class and a dataset suitable for the classification process had been created, the next step was the selection and creation of models from which the best performing one would be selected.

The initial model developed was a naive Bayes classifier. This served as a benchmark for comparing results with other models. Justifying the implementation of a computationally complex model necessitates its performance surpassing that of the simplest model, the Bayes classifier.

Following this, a logistic regression model was exclusively constructed for binary classification.

After assessing the performance of these classifiers, the focus shifted to crafting more flexible models capable of effectively partitioning the vector hyperspace in which the data were embedded. Specifically, attention turned to the construction of decision trees and random forests.

Initially, the attention was directed towards crafting an optimal decision tree to manage diverse classification tasks. This involved implementing a hyper-parameter optimization algorithm based on a meticulously chosen grid of values. This grid helped identify the most effective parameter combination, primarily focusing on the f1-score, as recommended by *Mithrakumar* (2021), *Mantovani* (2018), and *Raileanu and Stoffel* (2004).

The parameters subjected to tuning encompassed the criterion for evaluating split quality in the decision tree algorithm, with options including the Gini index or entropy reduction in the data. Additionally, were adjusted the maximum tree depth and the maximum number of leaf nodes. These parameters ranged from zero to fifty in increments of five.

Following the selection of the best classifier from the array of models constructed using the chosen parameter values, the aim shifted to create models with reduced bias and variance. This was tried to be achieved by implementing bagging and boosting algorithms for the decision tree, specifically utilizing the BaggingClassifier and AdaBoostClassifier algorithms provided by the scikit-learn library.

Following this, to mitigate the limitations arising from the discretionary selection of hyperparameter values, a decision was made, drawing inspiration from works like *Wang et al.* (2006) and *Xu et al.* (2009), to employ a hyperparameter selection algorithm rooted in Bayesian techniques. This approach sought to systematically determine values within predefined boundaries, thereby reducing the risk of overlooking the truly optimal parameters in constructing the classification model. Specifically, the Bayesian search was conducted within the interval of 1-200 for each hyperparameter.

Adhering to the same principle as employed in the grid search-based decision tree construction, it was also tried to reduce bias and variance in the decision tree crafted via Bayesian optimization. This was tried to be accomplished through the application of the previously mentioned bagging and boosting algorithms.

Ultimately, employing both methodologies, grid search and Bayesian optimization, were crafted random forests for each classification task. Specifically for this model type, the following hyperparameters underwent tuning:

- The number of estimators considered in constructing the random forest, limited to 20, 30, and 40 for grid search, and sought within the range of 10-100 for Bayesian optimization.
- The criterion for assessing split quality within the decision trees forming the random forest, with options encompassing the Gini index or entropy reduction.
- The maximum depth of the individual trees, explored within the range of none to 45 in 5-step increments for grid search, and between 1 and 100 for Bayesian optimization.
- Additionally, in the implementation of Bayesian search, adjustments were made to the maximum number of feature subsets considered in the individual estimators, as well as the minimum number of split samples and sample leaves within the decision trees comprising the forests.

To comprehensively assess the performance of the various models, different metrics were considered, thus avoiding reliance solely on accuracy. Given the imbalanced nature of the data and the domain of reference, it became imperative to account for the model's proficiency in correctly predicting observations related to the various diagnoses of interest. Therefore, it was opted for the harmonized average of precision and recall, denoted by the f1-score, as the primary metric for model comparison.

Additionally, to gain a clearer insight into the actual number of positive cases correctly identified from the total predicted positives, it was recorded the recall value.

It is worth noting that, for all classification tasks except the binary ones, the performance of the models was compared referring to the weighted values of the metrics. These weighted values account for the model's performance across various classes, factoring in the number of observations within each class.

Summary

This chapter has elucidated the methodologies applied in the development of predictive models within the project’s domain. The foundational steps of data acquisition and pre-processing were detailed, emphasizing the significance of collaborative partnerships and rigorous data cleaning. Feature engineering techniques, including the creation of categorical variables and numerical data refinement, were expounded upon.

Furthermore, it elucidated the approach to developing models for binary and multi-class classification, encompassing baseline models, class imbalance handling, and hyperparameter optimization. Evaluation metrics, specifically the f1-score and recall, were highlighted for assessing model performance.

Implementation and results

Data extraction and pre-processing

As written in the section dedicated to the description of the methodology used for the development of the project, the first step taken for the development of the models was characterised by the extraction from the databases of the Royal Berkshire hospital of the data necessary for the analysis.

This process therefore resulted in two datasets with different structures containing the distinct characteristics of interest, which however had two different dimensions according to the causes described in relation to the way the dataset containing the comorbidities was extracted.

Dependant Variable

The first analysis of the variable *'Diagnosis Codes Concat'*, selected as the dependent variable precisely because it contained information on the output of individual ED visits by individual patients, showed that the five most frequent variables were those summarised in *Table 2*.

Table 2: Target diagnosis description

Diagnosis codes	Diagnosis description	Frequency
380008	Diagnosis not classifiable	57911
R10X	Abdominal and pelvic pain	4402
R074	Chest pain	3834
263248	Gastrointestinal conditions	3744

It should be noted that for the interpretation of the codes, since there is no description within the dataset, the dictionary universally adopted by the hospitals of the countries belonging to the world health organisations was used (*ICD-10 version:2010*, no date), as well as an internal hospital dictionary in which the mode of derivation of those codes which cannot be traced back to the World Health organisation's own denomination is given. This dictionary can be found in *Appendix I*.

Furthermore, it should be mentioned that of the first five codes by frequency, it was decided not to construct a predictor targeting the first code, since it represents a case of little clinical interest.

Independent Variables

In first place the dataset related to the comorbidities was analysed, which originally had 821844 rows by four columns of interest whose structure is described in *Table 3*.

Table 3: Comorbidities dataset description

	Column name	Non-Null Count
1	PROBLEM_ANNOTATED_DISPLAY	692435
2	SNOMED_SEMANTIC_TAG	622256
3	PROBLEM_LIFE_CYCLE_STATUS_DESCRIPTION	693780
4	ANON_NUMBER	821844

The high dimensionality recorded for this dataset, therefore, raised the doubt that there might be a problem with duplicate observations; a doubt that was also reinforced by the analysis of the distribution of the number of comorbidities in relation to each individual patient which, as shown in *Figure 4*, resulted in the presence of some extreme values with patients having more than 1000 comorbidities recorded.

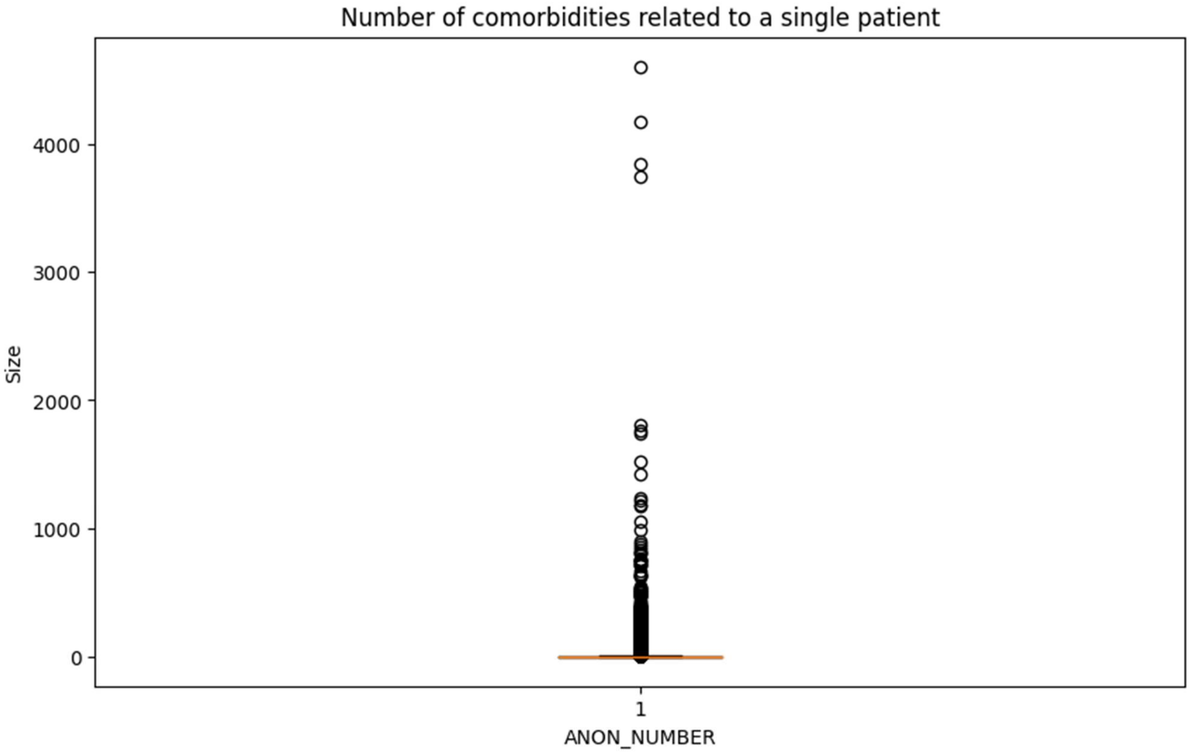


Figure 4: Distribution of comorbidities per patient

Proceeding with the analysis, therefore, it was found that the patients with duplicate values in the comorbidities recorded for them were more than one third of the dataset, to be precise, 60068, and for most of them the duplicates represented most of the recorded comorbidities. As an example, *Table 4* shows the duplicate values of the ten patients with the highest number of duplicate values recorded.

Table 4: Top ten patients with the most duplicates

Patient	Number of duplicates	Number of comorbidities registered
1	4582	4608
2	4150	4180
3	3831	3852
4	3723	3750
5	1775	1804
6	1733	1764
7	1714	1739
8	1506	1525
9	1413	1428
10	1222	1235

Thus, after cleaning the data from the presence of these duplicates, the final distribution of the comorbidity variable, although presenting extreme values, as shown in *Figure 5,* was considered significantly more realistic. See *Table 5* for descriptive values of this distribution.

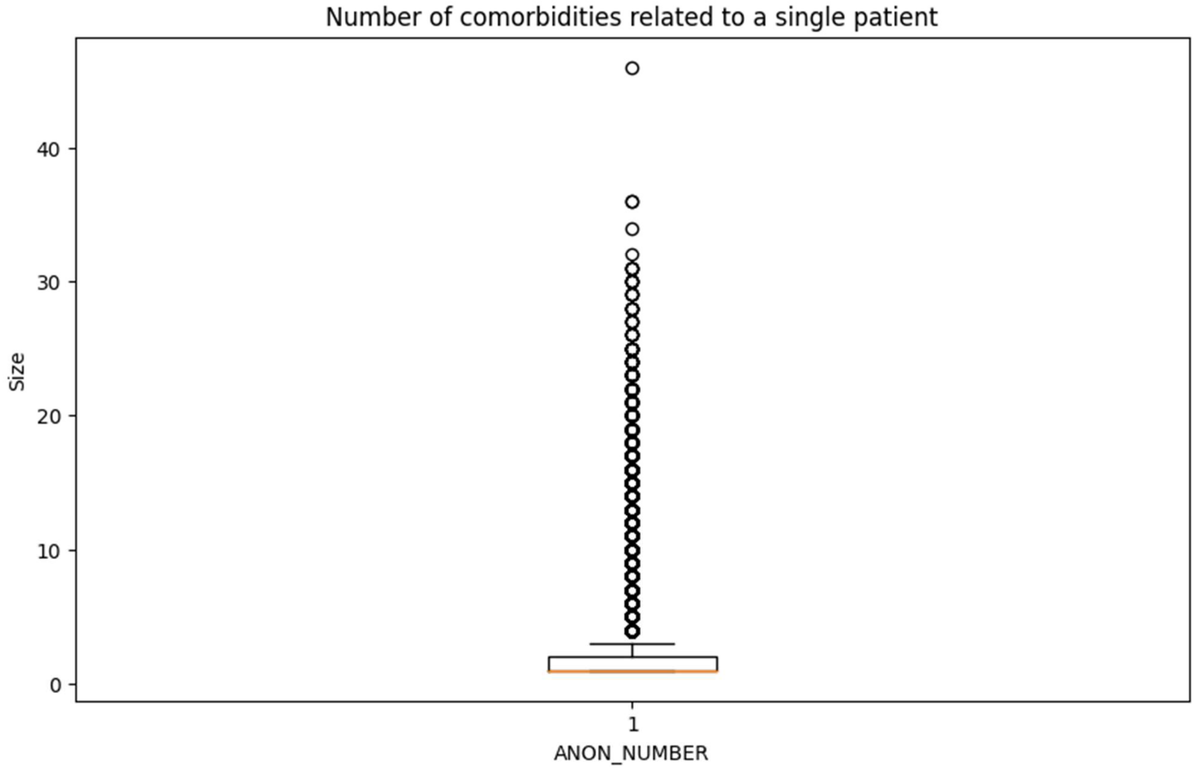


Figure 5: Distribution of comorbidities per patient after cleaning

Table 5: Descriptive statistics of comorbidities per patient after cleaning

Mean	std	min	25%	50%	75%	90%	95%	97%	99%	max
2.004013	2.467342	1	1	1	2	5	7	9	13	46

The analysis, as described, went on to identify the comorbidities most frequently associated with the diagnoses taken into analysis, which turned out to be 146, and columns recording the number of comorbidities per patient per comorbidity status were added, ultimately resulting in a dataset consisting of 142276 rows, representing the overall number of patients taken into analysis in the development of the project, and 151 columns.

Turning to the dataset concerning the other characteristics of visits to EDs, this dataset originally consisted of 275697 each representing a single visit to the ED by a patient and thirty-five columns of characteristics of these visits.

As already described, an analysis of the correspondence between the information carried by the column pairs was carried out, which resulted in the identification of those column pairs that presented the same information, of which the columns with the ending '_DESC' were selected for the one-hot encoding procedure, and those column pairs that, on the contrary, needed a more in-depth analysis.

A summary of the analysis can be found in *Table 6*.

Table 6: Summary of categories correspondence

Matching columns	**Different columns**
A_AND_E_ATTENDANCE_CATEGORY'	'A_AND_E_ARRIVAL_MODE'
'A_AND_E_ATTENDANCE_CATEGORY_DESC'	'A_AND_E_ARRIVAL_MODE_DESC'
'A_AND_E_DEPARTMENT_TYPE'	'A_AND_E_ATTENDANCE_DISPOSAL'
'A_AND_E_DEPARTMENT_TYPE_DESC'	'A_AND_E_ATTENDANCE_DISPOSAL_DESC'
'A_AND_E_INITIAL_ASSESSMENT_TRIAGE_CATEGORY'	'A_AND_E_PATIENT_GROUP'
'A_AND_E_INITIAL_ASSESSMENT_TRIAGE_CATEGORY_DESC'	'A_AND_E_PATIENT_GROUP_DESC'
'ETHNIC_CATEGORY_CODE'	'PRESENTING_COMPLAINT'
'ETHNIC_CATEGORY_CODE_DESC'	'PRESENTING_COMPLAINT_GROUP'
'PERSON_MARITAL_STATUS_CODE'	'SOURCE_OF_REFERRAL_FOR_A_AND_E'
'PERSON_MARITAL_STATUS_CODE_DESC'	'SOURCE_OF_REFERRAL_FOR_A_AND_E_DESC'
'HRG Code'	'PERSON_GENDER_CODE'
'HRG Desc'	'PERSON_GENDER_CODE_DESC'

Thus, the first pair of columns with unequal values to be analysed was the one associated with the arrival of the patients at the ED, which turned out to have a structure, as can be seen from *Table 7*, in which the differentiation between code '1' and code 2' was well defined and could be interpreted as the patient's arrival or non-arrival in the ambulance.

In the light of this clear differentiation, and to reduce the granularity of the data as much as possible, it was, however, decided to proceed with the one-hot encoding of the numeric code variable, thus resulting in only two columns.

Table 7: structure of arrival mode associated variables

Arrival mode	
Code: 1	*Code: 2*
Ambulance Air Ambulance	*Walk* *Private Transport* *Public Transport* *Police Vehicle* *Other*

On the other hand, the same could not be done for the pair of variables associated with the description of the attendance disposal for which, as shown in *Table 8*, the descriptions that distinguish the various visits to the ED even more clearly and substantially were accumulated under the same code, and it was therefore decided to retain their informative contribution to the model by proceeding with the one-hot encoding of these.

Table 8: structure of attendance disposal associated variables

Attendance disposal	
Code 1	'Admitted as IP - Same Trust' 'Admitted as inpatient' 'Admitted same Trust for CDU observation'
Code 2	'Disch for GP F/Up - to Check Progress' 'Disch for GP F/Up - to Remove Sutures' 'Disch for GP F/Up - to Refer Outpatient' 'Disch for GP F/Up - to Register with GP'
Code 3	'Discharged no Follow Up' 'Sent home via triage & advice' 'Treatment Complete' 'Treatment complete'
Code 4	'Discharged to A&E Follow-up'
Code 5	'Disch-Ref to Fracture clinic same trust'
Code 6	'Disch-Ref to Main O/P same trust'
Code 7	'Discharge to Primary Care Service' 'Disch-Ref to O/P Other trust' 'Tx to Other Trust for Admission' 'Discharged to see District Nurse'
Code 10	A&E Died in Department
Code 11	Tx to Facility in same Hosp for advice' 'Disch-Ref to Physio same trust
Code 12	Did not wait for triage' 'Did Not Wait to be seen by doctor' 'Seen by Doctor - Left before treatment
Code 13	Refused treatment
Code 14	Dead on Arrival

The next pair of variables analysed was the one relating to the type of accident that occurred to the patient as the cause of his visit to the ED, the structure of which is described in *Table 9*. In this case, although originally the number of unique values presented by the codes and the associated descriptions was different, the discrepancy between these two values, after changing all *'not known'* to *'unknown*', was represented exclusively by the accumulation under the same code of the descriptions *'unknown'* and *‘major incident'*, so in order to distinguish these two substantially different categories it was decided to proceed to one hot encoding for the descriptions. The same approach has been applied also when dealing with the pair of variables related to the source of referral for the patients' visit (*Table 10*) and for gender description, although it must be said that a minimum of manipulation with regard to this variable was made and, in fact, the only 35 observations with gender as 'unspecified' were dropped so that only two columns, one for males and one for females, were maintained in the final dataset.

Table 9: structure of patient group associated variables

Patient Group	
Code 10	Motor Vehicle (MVA)
Code 20	Assault
Code 30	Domestic
Code 40	Sport/Recreation
Code 50	Fire/Explosion
Code 60	Other
Code 80	Unknown' 'Not Known' 'Major Incident'

Table 10: structure of source of referral associated variables

Source of Referral	
Code 0	General Practitioner
Code 1	Self-Referral
Code 2	Local Authority Social Services
Code 3	Emergency Services
Code 4	Work
Code 5	Educational Establishment
Code 6	Police
Code 7	Health Care Provider(Same or Other)' 'Ambulance service - patient in transit
Code 8	Other
Code 92	Dentist' 'Community Dental Service

In conclusion, it is necessary to report the results of the dimensionality reduction process applied to the complaints presented by patients on arrival at the ED.

In fact, as described in the section describing the methodologies applied to the analysis, an attempt was made to map the 360 complaints that had a low frequency in the original dataset, specifically less than 5 instances, to identify typos.

Applying, then, an empirically selected threshold of 75 to the similarity calculation method provided by the *fuzzywyzzy* library, it was possible to identify, among the 360 with low frequency, 203 which represented more frequent complaint typos, and which were replaced with the correct complaints according to the mapping dictionary that can be found in *Appendix II*.

Having therefore explained the results of the analysis of the categorical variables, it is necessary to also outline the results of the analysis of the numerical variables which, as already mentioned, represent a decidedly smaller percentage of the total of the variables analysed.

Specifically, the two numerical variables of interest contained in the dataset are that relating to the age of the patients on arrival at the hospital and the duration in hours of their visit.

As far as the former is concerned, as shown in *Figure 6* and *Table 11*, no cleaning or pre-processing of any kind was required, as no extreme or unreasonable value was found, nor any null value.

Table 11: Descriptive statistics of patients' age at activity

Mean	***Std***	***min***	***25%***	***50%***	***75&***	***90%***	***95%***	***97%***	***Max***
36.628	28.205	0	11	32	60	79	85	88	106

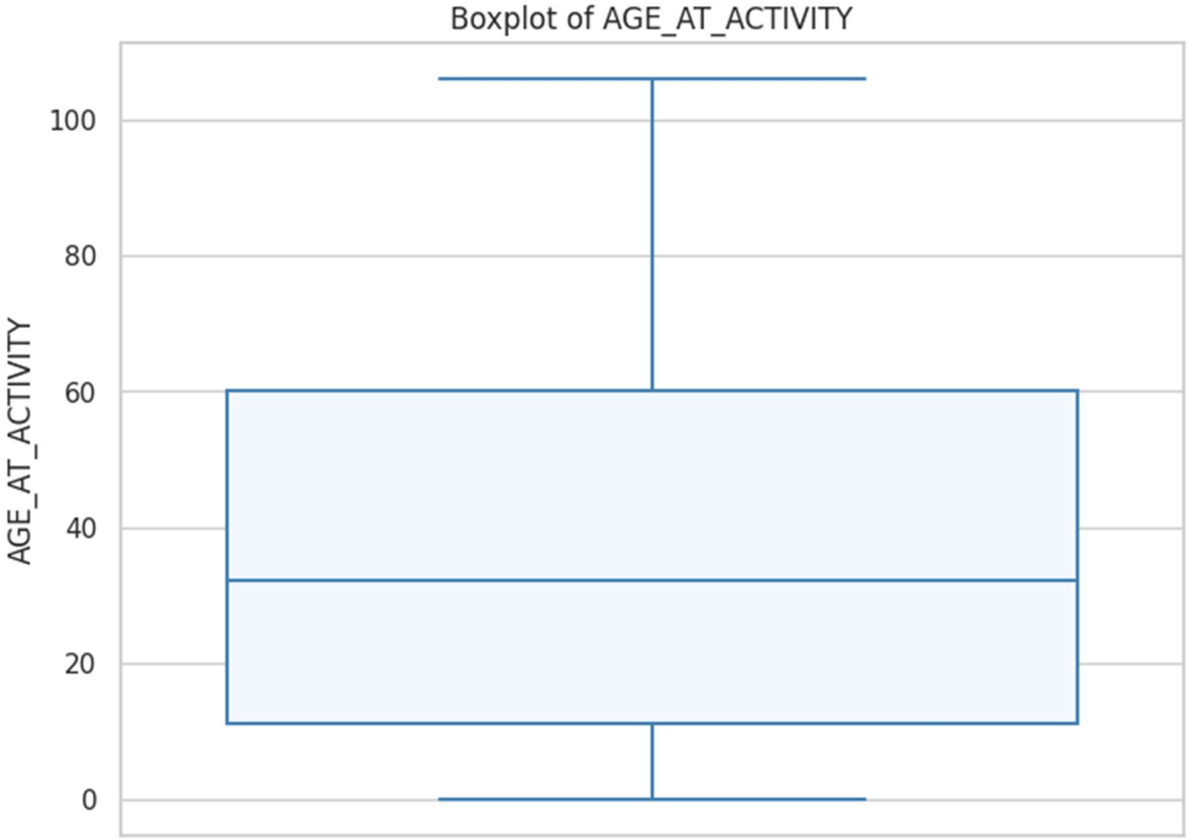


Figure 6: Distribution of patients' age at activity

On the contrary, about the length of stay in the ED, as shown in *Figure 7*, extreme and unreasonable values were found, with a maximum recorded for one patient of a stay of almost 36,000 hours. It was therefore decided not to consider in the analysis all values exceeding the 97th percentile (i.e., 8269 observations), both to safeguard the integrity of the performance of the prediction models that could have been distorted by the presence of these values, and in recognition of the possibility of errors in the data collection process.

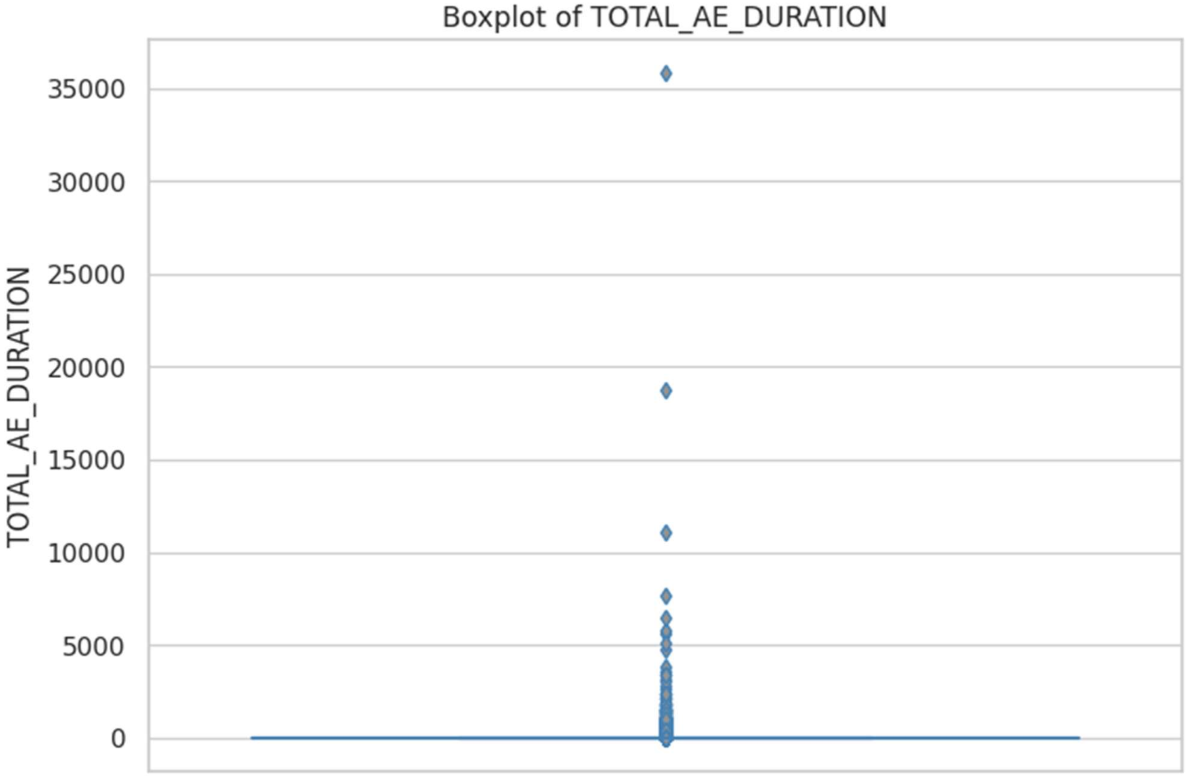


Figure 7: original distribution of duration of visits

The final distribution of the variable is, therefore, shown in *Figure 8* and *Table 12*.

Table 12: Descriptive statistics of ED visit duration

Mean	std	min	25%	50%	75%	90%	95%	97%	Max
3.254	2.974	0.016	1.66	2.73	3.80	5.45	7.716	9.8	27.76

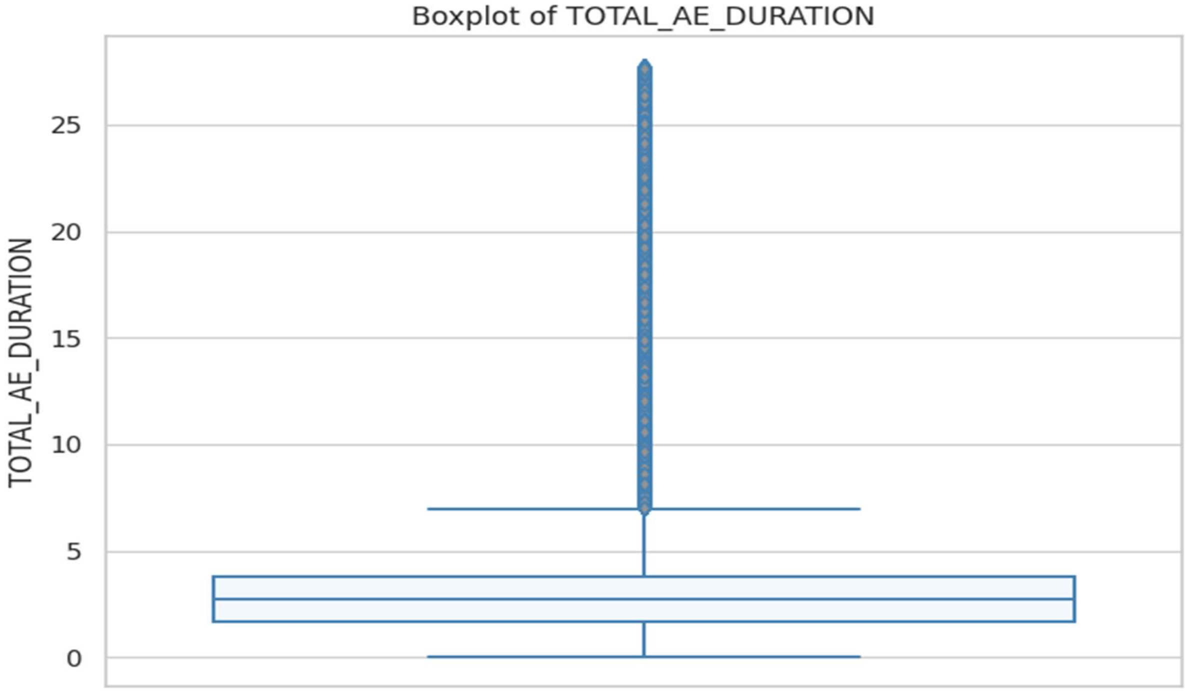


Figure 8: final distribution of duration of visits

Data Exploration

After analysing the dataset structure, it became crucial to gain a better grasp of the environment in which the model would operate post-construction and validation. To begin, about the age group with the most frequent diagnoses *was* inquired. Examining *Figure 9*, it becomes evident that as patients' age increases, the incidence of diagnoses also rises. Notably, the age group encompassing patients aged 17 to 50 exhibits the highest number of cases.

Examining the diagnosis distribution within younger age groups, it can be observed that the primary diagnoses at the ED department are gastrointestinal pain and abdominal pain, with fewer instances of chest pains. Conversely, with increasing patient age, the incidence of chest pain diagnoses systematically escalates, eventually becoming the prevailing diagnosis.

More specifically, a noticeable prevalence of gastrointestinal conditions as a diagnosis is observed within the infant age group (i.e., up to one month of age), along with a comparatively lower yet still observable prevalence of the other two diagnoses.

It is noteworthy that there is an absence of patients presenting with this type of condition in the age group spanning from 2 months to 5 years old. This occurrence may likely be attributed to the presence of alternative healthcare services, such as General Practitioners (GPs), as suggested by *Wier* (2013). These other services typically are more chosen by parents because they can offer more specialized personnel and equipment for infant care compared to a general ED that primarily caters to adults, additionally, accordig to , the presence of few instances registered by the consequent age bands can be due to the intermittent availability of such facilities (*Soliday and Hoeksel,* 2001).

Observing *Figure 10*, displaying the relative frequencies of diagnoses across various age groups, the insights derived from absolute frequencies are confirmed.

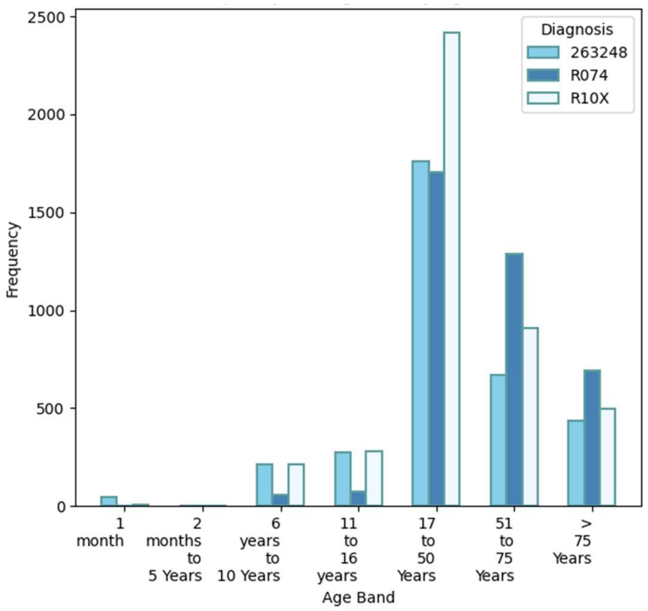


Figure 9: Frequency of diagnoses by age band

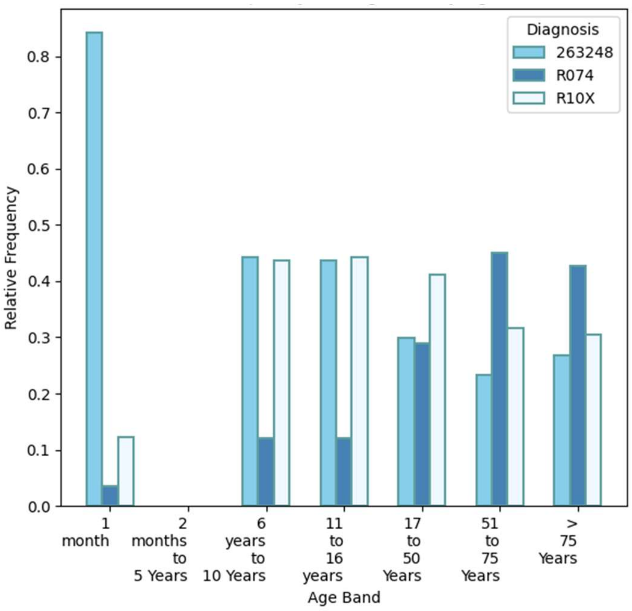


Figure 10: Relative frequency of diagnoses by age band

After conducting the initial analysis of diagnosis distribution across different age groups, focus shifted to understanding the potential outcomes associated with these diagnoses.

It is important to note a methodological difference when examining *Figure 11*. In contrast to the approach taken during data preparation for constructing the model, where were considered individual disposal descriptions, in creating this graph, it was opted for a more streamlined approach. To provide a clearer yet not overly detailed view, were used numerical codes corresponding to the disposal recorded in hospital databases, as outlined in *Table 13.*

Table 13: Mapping of disposal codes

Disposal code	Mapping
1	'Admitted',
2	'Disch for GP F/Up',
3	'Disch no F/Up',
4	'Disch to A&E F/Up',
5	'Disch to fracture clinic',
6	'Disch to Main O/P',
7	'Disch to Primary Care Service',
10	'Died in dept',
11	'Disch to physio',
12	'Left before treatment',
13	'Refused treatment',
14	'Death on arrival'

Examining *Figure 11*, it becomes evident that most instances involving a patient's diagnosis from the three conditions under scrutiny culminated in admission to the ED. Nevertheless, a minority of cases resulted in the patient's discharge, either under the care of a General Practitioner (GP) or without any further recommendations.

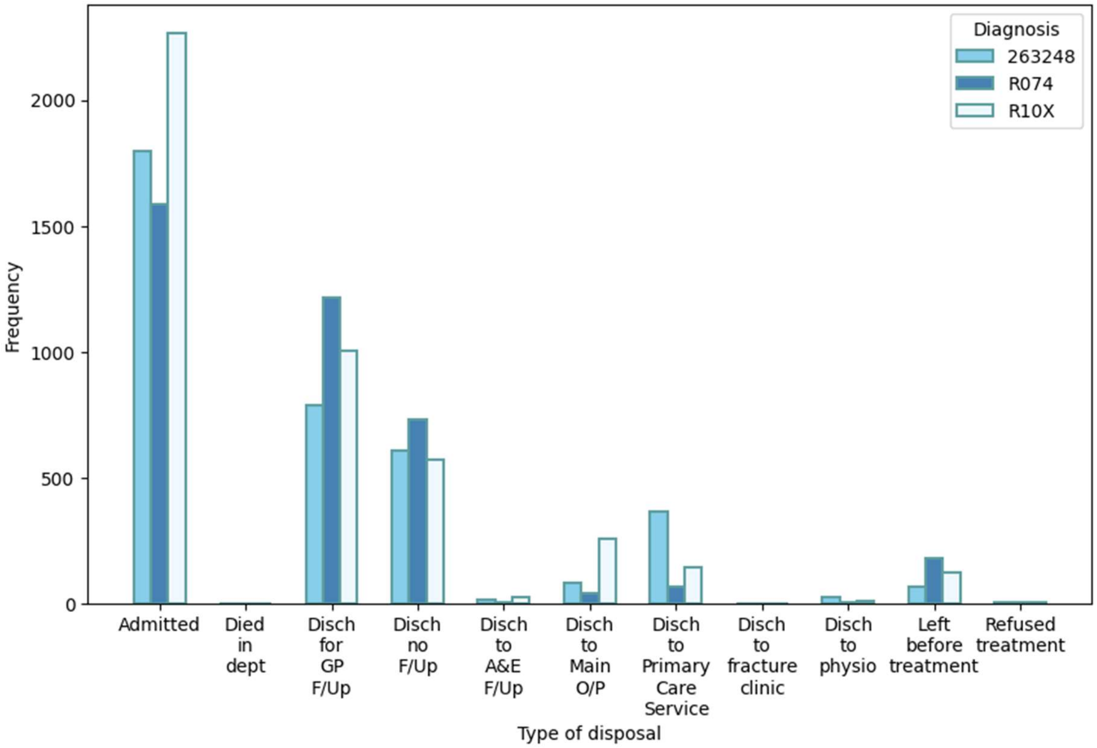


Figure 11: Type of disposal by diagnosis

Simultaneously, *Figure 12*, illustrating the triage category assigned to cases associated with these diagnoses upon arrival, reveals that the majority fell into the "urgent" category. However, it is noteworthy that they rarely reached the level of "very urgent" or necessitated immediate resuscitation procedures.

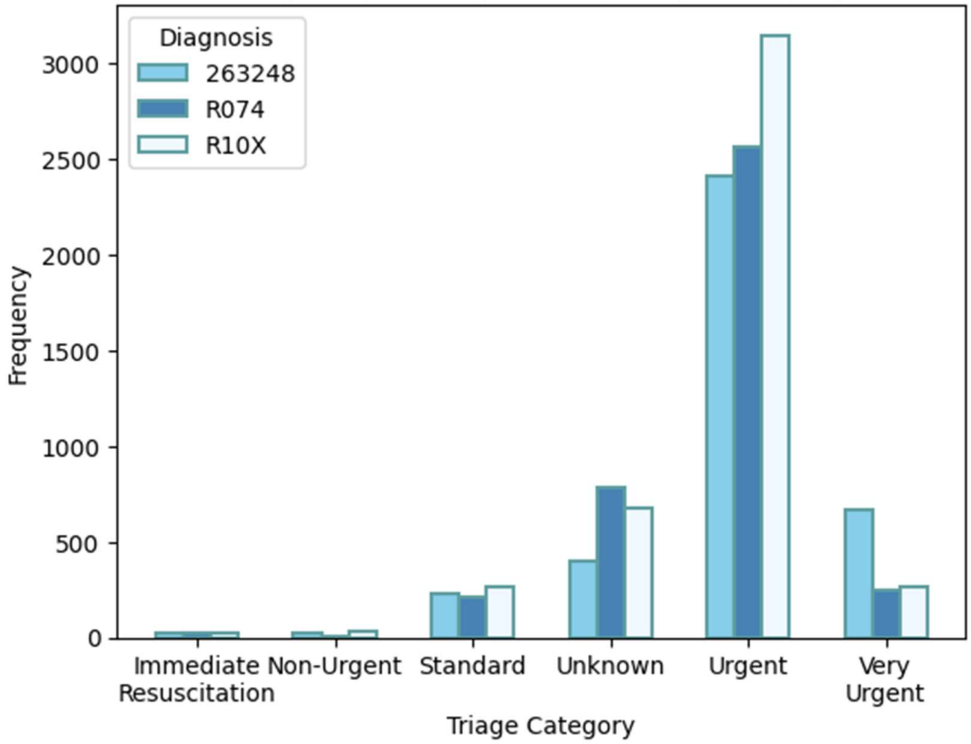


Figure 12: Frequency of diagnoses by Triage category

Following the previous analysis, the focus shifted to discerning the durations that patients diagnosed with one of the three conditions spent in the ED. The outcomes, as evident from *Figure 13, Figure 14*, and *Figure 15*, indicate that, on average, patients diagnosed with gastrointestinal pain tend to have the lengthiest stay at the ED, approximately 4.21 hours. In contrast, the shortest average stay is associated with chest pain diagnoses, at 3.79 hours.

Notably, it is intriguing to observe that the mean values in all three cases are positively skewed due to the presence of right-handed outliers. This is evident when comparing them to the median values. Even when considering the medians, gastrointestinal conditions and abdominal pain share the longest ED stays, while chest pain diagnoses consistently exhibit the shortest stays.

In conclusion, while the duration variability is similar across all three diagnoses, gastrointestinal conditions display the greatest variability, whereas chest pain diagnoses show the least.

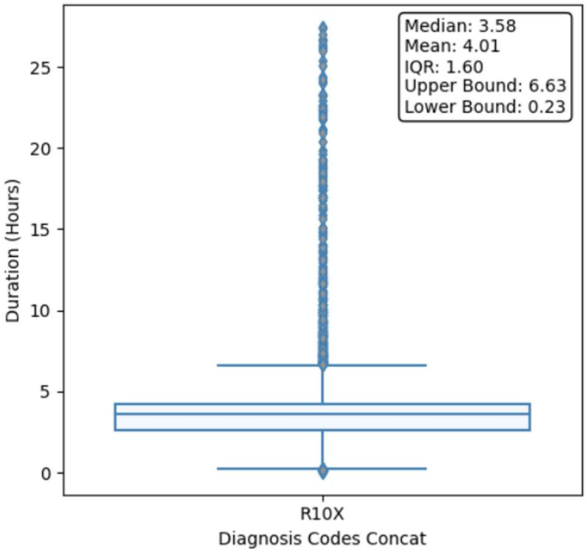


Figure 13: Duration of ED stays when abdominal pain is diagnosed

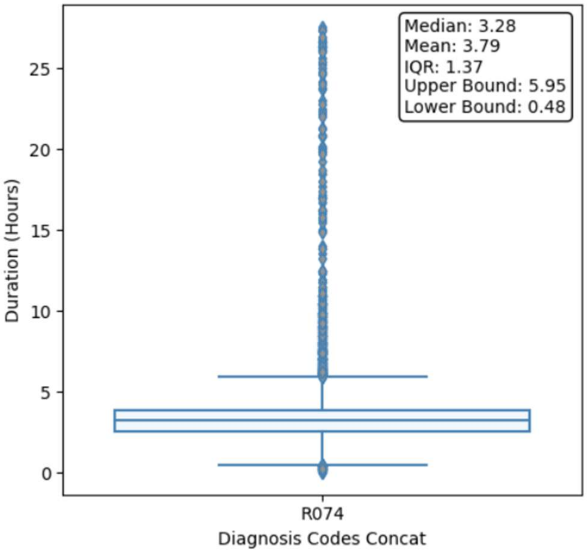


Figure 14: Duration of ED stays when chest pain is diagnosed.

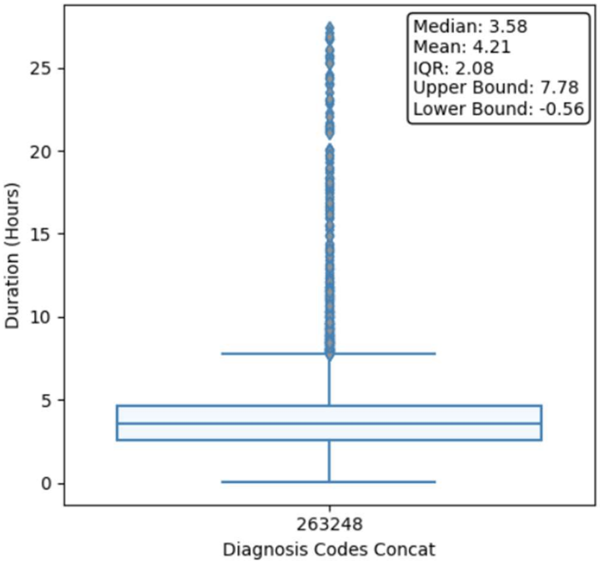


Figure 15: Duration of ED stays when gastrointestinal condition is diagnosed.

As suggested by *Zou and Schiebinger* (2018), and widely treated in the contemporary academic landscape, such as by *Dixon-Román, Nichols and Nyame-Mensah* (2019) or, more specifically in the healthcare sector by *Owens and Walker* (2020); in this last section of the data exploration, is embarked an examination of the dataset's ethnic representation and its consequential implications. As previously discussed, were have been scrutinised various aspects of the dataset, such as age-related diagnoses and patient dispositions.

However, it has been considered crucial to acknowledge that the dataset's ethnic makeup plays a pivotal role in shaping the subsequent discussion on healthcare outcomes and accessibility. In the forthcoming analysis, it has been so investigated the distribution of ethnicities within the dataset to shed light on potential imbalances and their significance in the broader context of healthcare disparities.

In particular, the analysis, illustrated in *Figure 16*, reveals a considerable diversity of ethnic groups seeking access to the Royal Berkshire Hospital's ED. Nevertheless, one ethnic group stands out as the most prevalent in the data: the *'British'* ethnic group.

This contrast becomes even more pronounced when, as illustrated in *Figure 17*, are examined the combined data for the *'British'* ethnic group and those akin to it (*'Irish'* and *'other white'*), comparing it to the combined data for all other ethnic groups.

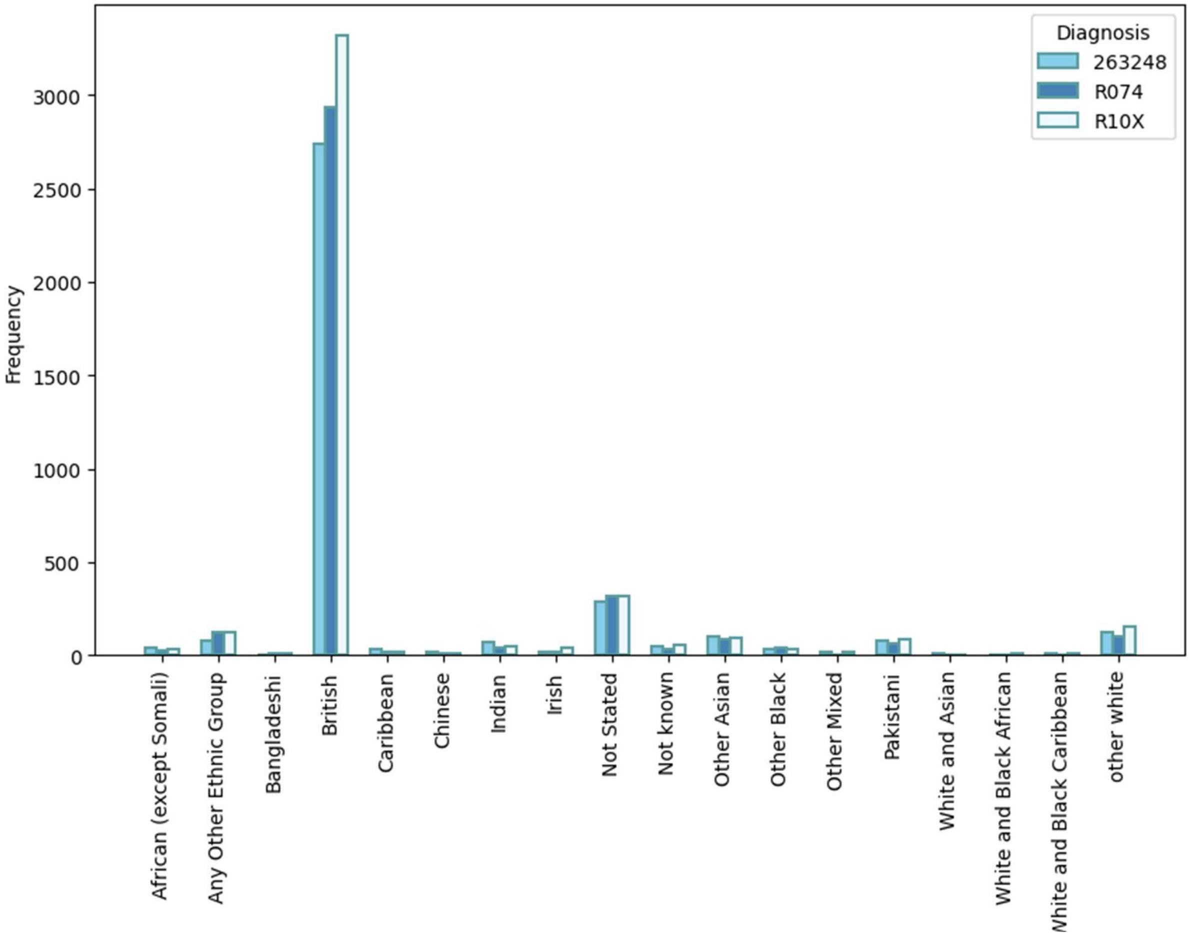


Figure 16: Frequency of diagnoses by ethnicity group

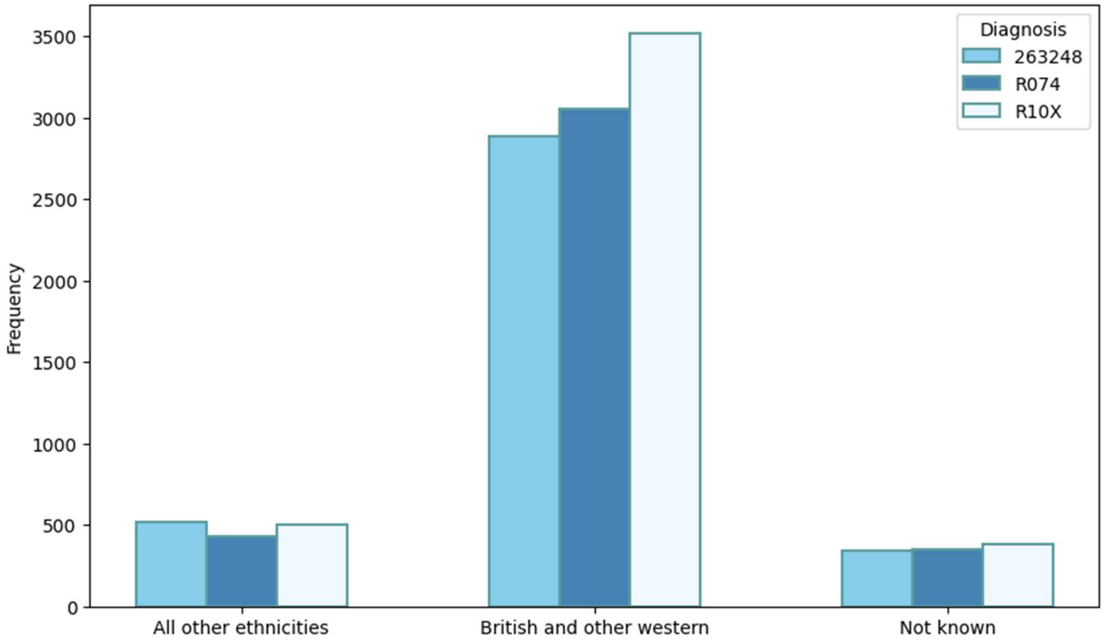


Figure 17: Frequency of diagnoses by ethnicity group (aggregated)

Model building and evaluation.

Binary classification

As outlined in the description of the methodologies employed for model construction, the project initially focused on predicting the occurrence of the most prevalent diagnosis within the dataset, namely R10X, associated with abdominal pain.

As depicted in *Table 14*, the original data exhibited a substantial imbalance, necessitating the application of under sampling techniques.

Table 14: Classes' distribution of binary target variable

Bianry Target variable		
Class	*Diagnosis*	*Number of observations*
0		259524
1	R10X	4152

Upon mitigating the data imbalance issue, we proceeded with the model construction, as previously detailed. The outcomes are summarily presented in *Table 15*.

For models subject to parameter tuning, the optimal parameter configurations are as follows:

- Grid search Decision Tree:
  - Best Parameters:
    - Criterion: Gini index
    - Max Depth: None
    - Max Leaf Nodes: 50
- Bayes optimized Decision Tree:
  - Best Parameters:
    - Criterion: Gini index
    - Max Depth: 149
    - Max Leaf Nodes: 48
- Grid search Random Forest:
  - Best Parameters:
    - Criterion: Entropy reduction
    - Max Depth: 30
    - Number of Estimators: 40
- Bayes optimized Random Forest:
  - Best Parameters:
    - Criterion: Entropy reduction
    - Max Depth: 59
    - Max Features: 0.1
    - Min samples leaf: 1
    - Min samples split: 20
    - Number of Estimators: 100

Table 15: Binary classification models performances

Model	Accuracy	F1_Score	Recall_Score
BayesOpt RF	0.86514	0.86325	0.85181
Bagging DT	0.86273	0.86315	0.86627
BayesOpt Bagging DT	0.86213	0.8623	0.86386
Random Forest	0.8543	0.85439	0.85542
BayesOpt DT	0.8519	0.85529	0.8759
Gridsearch DT	0.85069	0.85429	0.8759
Bayesopt AdaBoost DT	0.82601	0.8293	0.84578
AdaBoost DT	0.81878	0.82242	0.83976
Naive Bayes	0.80253	0.81214	0.85422
Logistic Regression	0.7929	0.78607	0.76145

Upon reviewing the table, it becomes evident that the model achieving the highest F1 score performance is the Bayes optimized random forest the detailed performance of which is elucidated in *Table 16*.

Table 16: Bayes optimized random forest binary classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.86	0.88	0.87
1	0.88	0.85	0.86
Accuracy	0.87		
weighted avg	0.87	0.87	0.87

As the table shows for class 0, which is the class related to the observation which do not present the diagnosis, the model exhibits a precision of 0.86, implying that when it predicts instances as class 0, it is correct about 86% of the time. Additionally, the recall for class 0 is 0.88, indicating that the model correctly identifies 88% of the actual class 0 instances. The F1-score, which balances precision and recall, is 0.87 for class 0.

For class 1, the model displays a slightly higher precision of 0.88, suggesting that when it predicts instances as class 1, it is correct about 88% of the time. However, the recall for class 1 is 0.85, meaning the model captures 85% of the actual class 1 instances and its associated F1-score is 0.86.

In summary, the model demonstrates a balanced performance in classifying both classes, with class 0 having a slightly higher recall and class 1 showing slightly higher precision. The F1-scores for both classes are quite close, indicating a well-rounded ability to make accurate predictions for both categories.

Multilevel classification – Three levels

Following this, the focus of the project shifted towards a classification task, extending its scope to predict the two most common diagnoses at the Royal Berkshire Hospital. This expansion involved the inclusion of observations related to diagnosis R074, associated with chest pain.

As depicted in *Table 17*, the initial dataset, as for the binary classification, exhibited significant imbalance, prompting the utilization of under-sampling techniques.

Table 17: Classes' distribution of three levels target variable

Three levels Target variable		
Class	*Diagnosis*	*Number of observations*
0		259524
1	R10X	4152
2	R074	3717

After addressing the data imbalance problem, models were built as previously described. *Table 18* provides a summary of the findings.

The ideal parameter combinations for models subject to parameter adjustment are as follows:

- Grid search Decision Tree:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: None
        - Max Leaf Nodes: 25
- Bayes optimized Decision Tree:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: 126
        - Max Leaf Nodes: 25
- Grid search Random Forest:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: 30
        - Number of Estimators: 40
- Bayes optimized Random Forest:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: 99
        - Max Features: 0.486
        - Min samples leaf: 6
        - Min samples split: 34
        - Number of Estimators: 100

Table 18: Three levels classification models performances

Model	Accuracy	F1_Score	Recall_Score
BayesOpt RF	0.83069	0.86325	0.85181
Random Forest	0.82656	0.85439	0.85542
Gridsearch DT	0.82243	0.85429	0.87590
Bagging DT	0.82243	0.86315	0.86627
BayesOpt DT	0.82243	0.85529	0.87590
BayesOpt Bagging DT	0.82243	0.86230	0.86386
AdaBoost DT	0.77351	0.82242	0.83976
Bayesopt AdaBoost DT	0.77351	0.82930	0.84578
Naive Bayes	0.72078	0.81214	0.85422

It is clear from the table that the model with the best F1 score performance is the Bayes optimised random forest, whose specific performance is explained in *Table 19*.

Table 19: Bayes optimized random forest three levels classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.84	0.86	0.85
1	0.80	0.78	0.79
2	0.85	0.81	0.83
Accuracy	0.83		
weighted avg	0.83	0.83	0.83

As the table shows class 0 has the highest precision of 0.84, indicating that when the model predicts an instance as class 0, it is correct 84% of the time. Class 2 also has a high precision of 0.85. However, class 1 has the lowest precision of 0.80, meaning it has a slightly higher rate of false positives.

Simultaneously class 0 has a recall of 0.86, indicating that the model captures 86% of all actual class 0 instances. Class 1 has a recall of 0.78, which means it is less effective at identifying all class 1 instances and class 2 falls in between with a recall of 0.81.

Additionally, class 0 has the highest F1-score of 0.85, indicating a good balance between precision and recall and class 2 also has a respectable F1-score of 0.83. Class 1, however, has the lowest F1-score of 0.79, suggesting that it might need further optimization.

In summary, this Bayes-optimized Random Forest model shows decent performance overall. It excels in classifying class 0, with high precision and recall, and maintains a good balance between precision and recall for class 2.

Multilevel classification – Four levels

As a final effort, the development of the project led to the attempt to develop a classifier considering the top three diagnoses by patient count at the referral hospital, including in the analysis, consequently, patient observations related to gastrointestinal issues.

The initial dataset, as for the other two classification tasks, showed severe imbalance, as seen in *Table 20*, which, as before, led to the use of under-sampling approaches.

Table 20: Classes' distribution of four levels target variable

Four levels Target variable		
Class	*Diagnosis*	*Number of observations*
0		259524
1	R10X	4152
2	R074	3717
3	263248	3573

After addressing the issue of the data imbalance, models were constructed as described in methodology section and, as before, the results are summarised in *Table 21* and the best parameters combinations are the following:

- Grid search Decision Tree:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: None
        - Max Leaf Nodes: 25
- Bayes optimized Decision Tree:
    - Best Parameters:
        - Criterion: Entropy reduction
        - Max Depth: 6
        - Max Leaf Nodes: 200
- Grid search Random Forest:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: 30
        - Number of Estimators: 40
- Bayes optimized Random Forest:
    - Best Parameters:
        - Criterion: Gini index
        - Max Depth: 29
        - Max Features: 0.456
        - Min samples leaf: 1
        - Min samples split: 2
        - Number of Estimators: 100

Table 21: Four levels classification models performances

Model	Accuracy	F1_Score	Recall_Score
Gridsearch DT	0.74197	0.73339	0.74197
Bagging DT	0.73935	0.72543	0.73935
BayesOpt RF	0.73476	0.72189	0.73301
Random Forest	0.73301	0.62113	0.63142
BayesOpt Bagging DT	0.73236	0.71518	0.73127
BayesOpt DT	0.73127	0.68011	0.67227
AdaBoost DT	0.67227	0.68011	0.67227
Bayesopt AdaBoost DT	0.63142	0.71540	0.73236
Naive Bayes	0.61372	0.62111	0.61372

The grid search decision tree whose precise performance is described in *Table 22*, has the best F1 score performance, as is evident from the table.

Table 22: Grid search decision tree four levels classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.81	0.87	0.84
1	0.58	0.64	0.61
2	0.79	0.81	0.80
3	0.61	0.38	0.47
Accuracy	0.74		
weighted avg	0.73	0.74	0.73

As the table shows, class 1 stands out with a precision of 0.58, suggesting that when the model predicts this class, it is often incorrect. This could mean that the model tends to produce a high number of false positives for class 1 being a substantial limitation. Furthermore, the recall for class 3 is only 0.38, indicating that the model struggles to capture true positive instances in this category. This implies that class 3 is often misclassified as another class or not identified correctly.

Considering the F1-scores, which strike a balance between precision and recall, class 3 still lags with an F1-score of 0.47 highlighting the challenge of the model in achieving both high precision and recall simultaneously for class 3. In contrast, classes 0 and 2 appear to be better predicted by the model, as evidenced by their higher precision, recall, and F1-scores.

The overall accuracy of 0.74 might seem decent. However, it's crucial to recognize that this accuracy is heavily influenced by the model's performance on classes 0 and 2 while neglecting the difficulties it faces with classes 1 and 3.

In summary, this classification report underscores the limitations of the model in predicting classes 1 and 3. It tends to produce false positives for class 1 and struggles to capture true positive instances for class 3. Finally, performances of all the model build to perform the classification tasks are reported in *Appendix III*.

Discussion and Evaluation

Introduction

In this chapter, critical aspects surrounding the validity, evaluation, and limitations of research are analysed. The cornerstone of study's credibility lies in the meticulous examination of its validity, encompassing dimensions such as data and methodological validity, as well as the management of biases inherently present in healthcare data. It is explored how research dealt intricacies of data challenges, harnessed machine learning techniques, and addressed potential biases. Moreover, it is assessed the real-world applicability and transformative potential of research findings.

Validity of research

The research draws from a dataset meticulously gathered from reputable healthcare sources, reflecting extensive diversity in patient demographics, conditions, and treatments. However, were encountered certain data challenges such as inconsistent data formats. These inconsistencies, if left unaddressed, could have jeopardized the accuracy and reliability of analysis. The complexity of dealing with inconsistent data formats underscores the significance of data validity in research. These efforts were essential to ensure that data accurately represented healthcare landscape the project sought to investigate.

Additionally, the application of various machine learning techniques served as the foundation of methodological framework. These techniques were thoughtfully chosen and rigorously implemented to ensure the reliability of findings.

Yet, it is imperative to acknowledge the inherent potential for bias within healthcare data. Healthcare data, by its nature, can be influenced by numerous factors, including variations in healthcare access, socio-economic disparities, and regional healthcare practices. These factors can introduce subtle but impactful biases into the dataset.

Additionally, it's essential to acknowledge the "black box" nature of some machine learning models. While undeniably powerful, these models can be challenging to interpret, potentially hindering their clinical adoption. As such, is recognized the importance of future research endeavours focusing on developing more interpretable models or robust methods for explicating the rationale behind model predictions. This transparency is crucial for mitigating the opacity often associated with machine learning models enhancing the chances of being applied in real world scenarios.

In conclusion, our research scrutinized the terrain of data validity and methodological considerations. Despite the challenges posed by missing data and potential bias, were adopted rigorous approaches to enhance the trustworthiness of findings. These measures underscore commitment into producing reliable insights into healthcare outcomes, ultimately contributing to informed healthcare practice and policy decisions.

Evaluation of the Research

This research is aligned with the overarching objective of enhancing patient care, optimizing resource allocation, and shaping healthcare policies for the betterment of society.

By harnessing the predictive power of machine learning models, the study has the aim of highlighting potential direction to develop healthcare practices. Early identification of disease risks, indeed, could empower healthcare practitioners to allocate their resources judiciously and tailor interventions to suit the needs of individual patients potentially leading to improved patient outcomes and, eventually, reductions in the overall cost burden of healthcare systems.

Nonetheless to truly evaluate the research's impact, it must be considered its real-world applicability. The research, indeed, while demonstrating strong predictive capabilities within our dataset, needs validation in actual healthcare settings that should involve collaborative efforts with healthcare institutions to implement and assess the practicality and effectiveness of our predictive model, being real-world testing the only possible provider of invaluable insights into the adaptability of the research within healthcare systems and the extent to which it can enhance patient care.

Additionally, another critical aspect of the research evaluation could be considered its potential to promote patient-cantered care. By tailoring interventions based on individualized risk assessments, it can aim to shift the focus from a one-size-fits-all approach to a patient-centric model of care improving patient satisfaction and, also, contributing to better healthcare outcomes.

In conclusion, the research demonstrates to have potential to influence healthcare practices positively, particularly in the realms of disease outcome prediction and resource allocation. However, it is acknowledged that the true impact of the research will be realized through collaborative efforts with healthcare providers and a commitment to making machine learning models more transparent and clinically viable.

Limitations of the research

As any other research also, this study is subject to several significant limitations that are imperative to address, spanning both data-related constraints and methodological considerations.

Firstly, one notable limitation lies in the representativeness of the training data. Despite its extensive nature, the dataset used in this study does not comprehensively capture the intricate tapestry of ethnic diversity encountered in real-world healthcare settings. To mitigate this limitation, it could be considered employing oversampling techniques or generating synthetic data for underrepresented ethnic groups or, additionally, it could be explored the effect of the incorporation of external data sources providing a more diverse representation of ethnic backgrounds in healthcare.

Another significant limitation is the absence of comparative data for human intelligence in predicting disease outcomes during the initial stages of the triage process. Without such data, it is challenging to assess whether the machine learning models outperform or complement human intelligence in disease outcome prediction. To address this limitation, it could be considered a further collaborating with healthcare institutions to conduct studies that directly compare the performance of the predictive models built with human clinicians or alternatively, develop simulated human intelligence models based on historical clinical decision-making data for comparison.

Furthermore, the inherent subjectivity associated with variable creation poses a potential limitation in this study, impacting model performance and the reliability of predictions. One aspect that requires specific attention is the selection of the subset of comorbidities most strongly correlated with target diagnoses for inclusion in the predictive model. Indeed, the selection process for comorbidities relied on the identification of those with the highest correlations to the target diagnoses entails inevitably some degree of subjectivity and led to the loss of a portion of the information load present in the dataset.

To address this limitation and potentially enhance the utilization of the full dataset, an alternative approach could be considered. One such approach involves the implementation of embedding techniques for comorbidities and a vectorization method, akin to how language models are constructed. By embedding comorbidities into a continuous vector space, in fact, the relationships between different comorbidities and their associations with target diagnoses can be more comprehensively captured allowing the model to learn intricate patterns and dependencies among comorbidities and could offer a more data-driven and less subjective approach to feature creation. By implementing such advanced techniques, the study could potentially enhance the richness of the feature space, capture latent patterns, and contribute to more accurate and robust predictive models. Moreover, this approach aligns with the current advancements in machine learning, particularly in the realm of natural language processing, where embedding and vectorization methods have demonstrated remarkable success in handling complex and high-dimensional data.

Recognizing and addressing these limitations remains pivotal for advancing the field, ensuring equitable healthcare AI systems, and fostering transparency and reliability in machine learning applications within healthcare contexts.

Summary

In summary, this chapter provides a comprehensive overview of research validity, evaluation, and the inherent limitations. Data and methodological validity have been scrutinized while recognizing the complexities and potential biases in healthcare data. Evaluation underscores the real-world applications of research, particularly in enhancing healthcare practices and patient care. Moreover, addressed the study's limitations have been openly addressed, offering practical strategies for mitigation. Throughout, commitment to transparency and the advancement of healthcare knowledge remains evident.

Conclusion

Summary of Research

This research has delved into the intricate landscape of healthcare data analysis, employing machine learning techniques to predict disease outcomes within ED settings. The primary objective was to enhance patient care, optimize resource allocation, and inform healthcare policies for the betterment of society. Through a meticulous examination of data quality, methodological rigor, and model performance, this study has strived to contribute valuable insights to the field of healthcare.

The research commenced with a thorough evaluation of data validity. A diverse and comprehensive dataset was meticulously collected from reputable healthcare sources, comprehending a wide array of patient demographics, conditions, and treatments. Nevertheless, the journey through this dataset was not without its challenges, notably the presence of inconsistent data formats. Addressing these inconsistencies was pivotal to ensuring the reliability and accuracy of analyses.

Furthermore, this research leaned on the robust foundation of various machine learning techniques. These techniques were thoughtfully selected and rigorously implemented to bolster the reliability of findings. However, it's crucial to acknowledge that "black box" nature of these models poses challenges in terms of interpretability, potentially impeding their seamless integration into clinical practice. Furthermore, it is important to recognize that healthcare data, even when meticulously collected, carries the potential for subtle biases rooted in healthcare access, socio-economic disparities, and regional practices. Acknowledging these potential biases is a step toward addressing them effectively.

Summary of Findings

The rich and diverse dataset utilized in this study has provided valuable insights into the intricate dynamics of healthcare. It has unveiled correlations, patterns, and dependencies that enhance our understanding of disease outcomes, resource allocation, and patient care. Furthermore, the application of machine learning techniques has demonstrated its effectiveness in predicting disease outcomes and built models have exhibited robust predictive capabilities, showcasing the potential for data-driven decision-making in healthcare.

Nonetheless, the interpretability challenge posed by some machine learning models has emerged as a significant hurdle, indeed, while these models have good predictive performance, their limited interpretability could hinder their adoption in clinical settings. This underscores the need for future research to focus on developing more interpretable models.

Importantly, the limitations acknowledged in this research have been reframed as opportunities for advancement. Addressing issues related to data representativeness, the absence of comparative human intelligence data, and subjectivity in variable creation holds promise for the field and fosters transparency in healthcare AI systems.

Conclusion and Recommendations

In summary, this research has traversed the landscape of healthcare data analysis and predictive modelling, with a focus on enhancing patient care, optimizing resource allocation, and shaping healthcare policies. Wrapping up this study several key takeaways and recommendations can be identified.

- The potency of machine learning models in healthcare is unmistakable, offering a pathway to more informed decision-making. However, the practical implementation of these models in real-world healthcare settings necessitates thoughtful consideration and validation.
- To bridge the divide between research findings and practical impact, collaborative initiatives with healthcare institutions are vital. Validation studies conducted in clinical settings can furnish invaluable insights into the adaptability and effectiveness of predictive models.
- Efforts to enhance the transparency and clinical applicability of machine learning models should remain a top priority in future research engendering trust among healthcare practitioners.

This research could be seen as a step toward a future where data-driven decision-making empowers healthcare providers to deliver tailored care, minimize resource allocation inefficiencies, and enhance patient outcomes.

Future Work

In concluding this research, it is recognized the continuous nature of the quest to enhance healthcare outcomes through predictive modelling.

One avenue for future exploration lies in the application of deep learning models, in fact, while built models have shown promising performance, delving into deep learning architectures could potentially elevate our predictive capabilities, enhancing the number of diagnoses predicted aiming to the prediction of all the ones included in the dataset.

Addressing the limitations related to data representativeness is another important future consideration. Incorporating external data sources that offer a more comprehensive representation of ethnic backgrounds and healthcare settings could significantly improve our model's ability to make accurate predictions for underrepresented populations, contributing to more equitable healthcare.

Furthermore, future research should involve collaborative efforts with healthcare institutions to validate the practicality and effectiveness of our predictive models. Comparative studies that evaluate the performance of machine learning models alongside human clinicians during the initial stages of triage can provide invaluable insights into how technology and human expertise can complement each other.

To seamlessly integrate predictive models into healthcare workflows, designing an intuitive and transparent user interface is essential. This interface should empower healthcare practitioners to interact with the model effortlessly, understand its predictions, and incorporate them into patient care decisions, all while prioritizing user-friendliness and efficiency.

Ethical considerations are also of paramount importance. As healthcare AI systems become increasingly prevalent, addressing issues related to data privacy, bias mitigation, and informed consent becomes critical. Future research should delve deeper into these ethical dimensions, ensuring responsible and equitable AI deployment in healthcare.

By exploring these possibilities, it can be possible to work toward a healthcare system that is not only data-driven but also equitable, transparent, and focused on patient well-being.

With these prospects for future work in mind, this dissertation is concluded with confidence that its findings can contribute to the ongoing transformation of healthcare practices and policies.

References

Arnaud, E. *et al.* (2020) *Deep Learning to Predict Hospitalization at Triage: Integration of Structured Data and Unstructured Text*. Available at: https://doi.org/10.1109/bigdata50022.2020.9378073.

Azari, A., Janeja, V.P. and Levin, S. (2015) *Imbalanced learning to predict long stay Emergency Department patients*. Available at: https://doi.org/10.1109/bibm.2015.7359790.

Boggan, J.C. *et al.* (2020) "Effectiveness of Acute Care Remote Triage Systems: A Systematic Review," *Journal of General Internal Medicine*, 35(7), pp. 2136–2145. Available at: https://doi.org/10.1007/s11606-019-05585-4.

Bunaciu, M. (2016) *Public Healthcare Services - Component of Tertiary Economy*. Available at: https://www.semanticscholar.org/paper/Public-Healthcare-Services-Component-of-Tertiary-Bunaciu/911561c610183c2bb8c6004574ebfefd962caae1.

Dahouda, M.K. and Joe, I. (2021) "A Deep-Learned embedding technique for categorical features encoding," *IEEE Access*, 9, pp. 114381–114391. Available at: https://doi.org/10.1109/access.2021.3104357.

Department of Health and Social Care (2013) "Guide to the Healthcare System in England," *GOV.UK* [Preprint]. Available at: https://www.gov.uk/government/publications/guide-to-the-healthcare-system-in-england.

Dixon-Román, E.J., Nichols, T.P. and Nyame-Mensah, A. (2019) "The racializing forces of/in AI educational technologies," *Learning, Media, and Technology*, 45(3), pp. 236–250. Available at: https://doi.org/10.1080/17439884.2020.1667825.

Domingues, I. *et al.* (2018) "Evaluation of Oversampling Data Balancing Techniques in the Context of Ordinal Classification," *2018 International Joint Conference on Neural Networks (IJCNN)* [Preprint]. Available at: https://doi.org/10.1109/ijcnn.2018.8489599.

Eccles, A. *et al.* (2019) "Patient use of an online triage platform: a mixed-methods retrospective exploration in UK primary care," *British Journal of General Practice*, 69(682), pp. e336–e344. Available at: https://doi.org/10.3399/bjgp19x702197.

Eichler, H.-G. *et al.* (2018) "Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?" *American Society for Clinical Pharmacology and Therapeutics*, 105(4), pp. 912–922. Available at: https://doi.org/10.1002/cpt.1226.

FitzGerald, G. *et al.* (2010) "Emergency department triage revisited," *Emergency Medicine Journal*, 27(2), pp. 86–92. Available at: https://doi.org/10.1136/emj.2009.077081.

fuzzywuzzy (2020). Available at: https://pypi.org/project/fuzzywuzzy/.

GIRFT (2021) *Emergency medicine GIRFT programme national specialty report, https://gettingitrightfirsttime.co.uk/girft-reports/*. Available at: https://gettingitrightfirsttime.co.uk/wp-content/uploads/2022/08/Emergency-Medicine-Apr22q.pdf.

Hodge, A. *et al.* (2013) "A review of the quality assurance processes for the Australasian Triage Scale (ATS) and implications for future practice," *Australasian Emergency Nursing Journal*, 16(1), pp. 21–29. Available at: https://doi.org/10.1016/j.aenj.2012.12.003.

Huang, P.S. *et al.* (2019) "A study of using syntactic cues in short-text similarity measure," *Journal of Internet Technology*, 20(3), pp. 839–850. Available at: https://doi.org/10.3966/160792642019052003017.

ICD-10 version:2010 (no date). Available at: https://icd.who.int/browse10/2010/en.

Jernite, Y. *et al.* (2011) *Predicting chief complaints at triage time in the emergency department.* [Workshop on Machine Learning for Clinical Data Analysis and Healthcare]. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fbe0ef9417aff74709607820d6f1d3318 3451bcc.

Joseph, J.J. *et al.* (2020) "Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information," *Journal of the American College of Emergency Physicians Open*, 1(5), pp. 773–781. Available at: https://doi.org/10.1002/emp2.12218.

Kaabi, S.A., Varughese, B. and Singh, R. (2022) "Public and Private Healthcare System in Terms of both Quality and Cost: A Review," *Journal of Clinical and DIagnostic Research* [Preprint]. Available at: https://doi.org/10.7860/jcdr/2022/55387.16742.

Klug, M. *et al.* (2020) "A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score," *Journal of General Internal Medicine*, 35(1), pp. 220–227. Available at: https://doi.org/10.1007/s11606-019-05512-7.

Kwon, J.-M. *et al.* (2021) "Deep Learning Algorithm to Predict Need for Critical Care in Pediatric Emergency Departments," *Pediatric Emergency Care*, 37(12), pp. e988–e994. Available at: https://doi.org/10.1097/pec.0000000000001858.

Leeuwenberg, A. *et al.* (2021) "Comparing methods addressing multi-collinearity when developing prediction models," *arXiv (Cornell University)* [Preprint]. Available at: https://doi.org/10.48550/arxiv.2101.01603.

Luo, H. *et al.* (2021) "Oversampling by a Constraint-Based Causal Network in Medical Imbalanced Data Classification," *Luo, Hao Et Al. "Oversampling by a Constraint-Based Causal Network in Medical Imbalanced Data Classification." 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021): 1-6.* [Preprint]. Available at: https://doi.org/10.1109/icme51207.2021.9428083.

Mackway-Jones, K., Marsden, J. and Windle, J. (2014) *Emergency Triage: Manchester Triage Group*. John Wiley & Sons.

Mantovani, R.G. (2018) *An empirical study on hyperparameter tuning of decision trees*. Available at: https://arxiv.org/abs/1812.02207.

Martinez-Gil, J. (2012) "An overview of textual semantic similarity measures based on web intelligence," *Artificial Intelligence Review*, 42(4), pp. 935–943. Available at: https://doi.org/10.1007/s10462-012-9349-8.

Mithrakumar, M. (2021) “How to tune a Decision Tree? - Towards Data Science,” *Medium*, 12 December. Available at: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680.

Nagarajah, S., Krzyzanowska, M.K. and Murphy, T. (2022) “Early Warning Scores and Their Application in the Inpatient Oncology Settings,” *JCO Oncology Practice*, 18(6), pp. 465–473. Available at: https://doi.org/10.1200/op.21.00532.

National Information Board (2019) *Personalised Health and Care*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/384650/NIB_Report.pdf.

NHS (2019) *NHS long term plan*. Available at: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf.

NHS England (2020) *Transformation of urgent and emergency care: Models of care and measurement*. Available at: https://www.england.nhs.uk/wp-content/uploads/2020/12/transformation-of-urgent-and-emergency-care-models-of-care-and-measurement.pdf.

NHS England (2022) *Guidance for emergency departments: Initial assessment*. Available at: https://www.england.nhs.uk/guidance-for-emergency-departments-initial-assessment/#patient-flow.

NHS England (2023) *Delivery plan for recovering urgent and emergency care services*, *https://www.england.nhs.uk*. Available at: https://www.england.nhs.uk/wp-content/uploads/2023/01/B2034-delivery-plan-for-recovering-urgent-and-emergency-care-services.pdf.

NHS England (no date) *NHS England » Urgent treatment centres*. Available at: https://www.england.nhs.uk/urgent-emergency-care/urgent-treatment-centres/.

NHS UK (no date a) *When to call 999*. Available at: https://www.nhs.uk/nhs-services/urgent-and-emergency-care-services/when-to-call-999/.

NHS UK (no date b) *When to use NHS 111 online or call 111*. Available at: https://www.nhs.uk/nhs-services/urgent-and-emergency-care-services/when-to-use-111/.

Nissim, M. and Markert, K. (2003) “Syntactic features and word similarity for supervised metonymy resolution,” *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* [Preprint]. Available at: https://doi.org/10.3115/1075096.1075104.

O’Neill, S.L. *et al.* (2021) “Why do healthcare professionals fail to escalate as per the early warning system (EWS) protocol? A qualitative evidence synthesis of the barriers and facilitators of escalation,” *BMC Emergency Medicine*, 21(1). Available at: https://doi.org/10.1186/s12873-021-00403-9.

Owens, K. and Walker, A. (2020) “Those designing healthcare algorithms must become actively anti-racist,” *Nature Medicine*, 26(9), pp. 1327–1328. Available at: https://doi.org/10.1038/s41591-020-1020-3.

Raileanu, L.E. and Stoffel, K. (2004) “Theoretical Comparison between the Gini Index and Information Gain Criteria,” *Annals of Mathematics and Artificial Intelligence*, 41(1), pp. 77–93. Available at: https://doi.org/10.1023/b:amai.0000018580.96245.c6.

Rozeva, A. and Zerkova, S.I. (2017) “Assessing semantic similarity of texts – Methods and algorithms,” *Nucleation and Atmospheric Aerosols* [Preprint]. Available at: https://doi.org/10.1063/1.5014006.

Sánchez-Salmerón, R. *et al.* (2022) “Machine learning methods applied to triage in emergency services: A systematic review,” *International Emergency Nursing*, 60, p. 101109. Available at: https://doi.org/10.1016/j.ienj.2021.101109.

Seow, E. (2013) “Leading and managing an emergency department—A personal view,” *Journal of Acute Medicine*, 3(3), pp. 61–66. Available at: https://doi.org/10.1016/j.jacme.2013.06.001.

Shyu, M.-L. *et al.* (2005) “Handling Nominal Features in Anomaly Intrusion Detection Problems,” *15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications* [Preprint]. Available at: https://doi.org/10.1109/ride.2005.10.

Smith, G.A. *et al.* (2013) “The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death,” *Resuscitation*, 84(4), pp. 465–470. Available at: https://doi.org/10.1016/j.resuscitation.2012.12.016.

Soliday, E. and Hoeksel, R. (2001) “Factors related to paediatric patients’ emergency department utilization,” *Psychology Health & Medicine*, 6(1), pp. 5–12. Available at: https://doi.org/10.1080/713690225.

Sterling, N.W. *et al.* (2019) “Prediction of emergency department patient disposition based on natural language processing of triage notes,” *International Journal of Medical Informatics*, 129, pp. 184–188. Available at: https://doi.org/10.1016/j.ijmedinf.2019.06.008.

Sterling, N.W. *et al.* (2020) “Prediction of emergency department resource requirements during triage: An application of current natural language processing techniques,” *Journal of the American College of Emergency Physicians Open*, 1(6), pp. 1676–1683. Available at: https://doi.org/10.1002/emp2.12253.

Subbe, C.P. *et al.* (2001) “Validation of a modified Early Warning Score in medical admissions,” *QJM: An International Journal of Medicine*, 94(10), pp. 521–526. Available at: https://doi.org/10.1093/qjmed/94.10.521.

Tahayori, B., Chini-Foroush, N. and Akhlaghi, H. (2021) “Advanced natural language processing technique to predict patient disposition based on emergency triage notes,” *Emergency Medicine Australasia* [Preprint]. Available at: https://doi.org/10.1111/1742-6723.13656.

Teubner, D. *et al.* (2015) “Model to predict inpatient mortality from information gathered at presentation to an emergency department: The Triage Information Mortality Model (TIMM),” *Emergency Medicine Australasia*, 27(4), pp. 300–306. Available at: https://doi.org/10.1111/1742-6723.12425.

The Royal College of Emergency Medicine [RCEM] (2017) *Initial assessment of emergency department patients*. Available at: https://rcem.ac.uk/wp-content/uploads/2021/10/SDDC_Intial_Assessment_Feb2017.pdf.

Wah, Y.B. *et al.* (2013) “An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets,” in *Lecture notes in electrical engineering*, pp. 13–22. Available at: https://doi.org/10.1007/978-981-4585-18-7_2.

Wang, B. *et al.* (2022) “Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning,” *Decision Support Systems*, 166, p. 113899. Available at: https://doi.org/10.1016/j.dss.2022.113899.

Wang, L. *et al.* (2006) “Combining decision tree and Naive Bayes for classification,” *Knowledge Based Systems*, 19(7), pp. 511–515. Available at: https://doi.org/10.1016/j.knosys.2005.10.013.

Wier, L. (2013) *Overview of Children in the Emergency Department, 2010*. Available at: https://www.semanticscholar.org/paper/Overview-of-Children-in-the-Emergency-Department%2C-Wier-Yu/6ec634cfb2042cc30567f51eb0284d21b6ebbdb7.

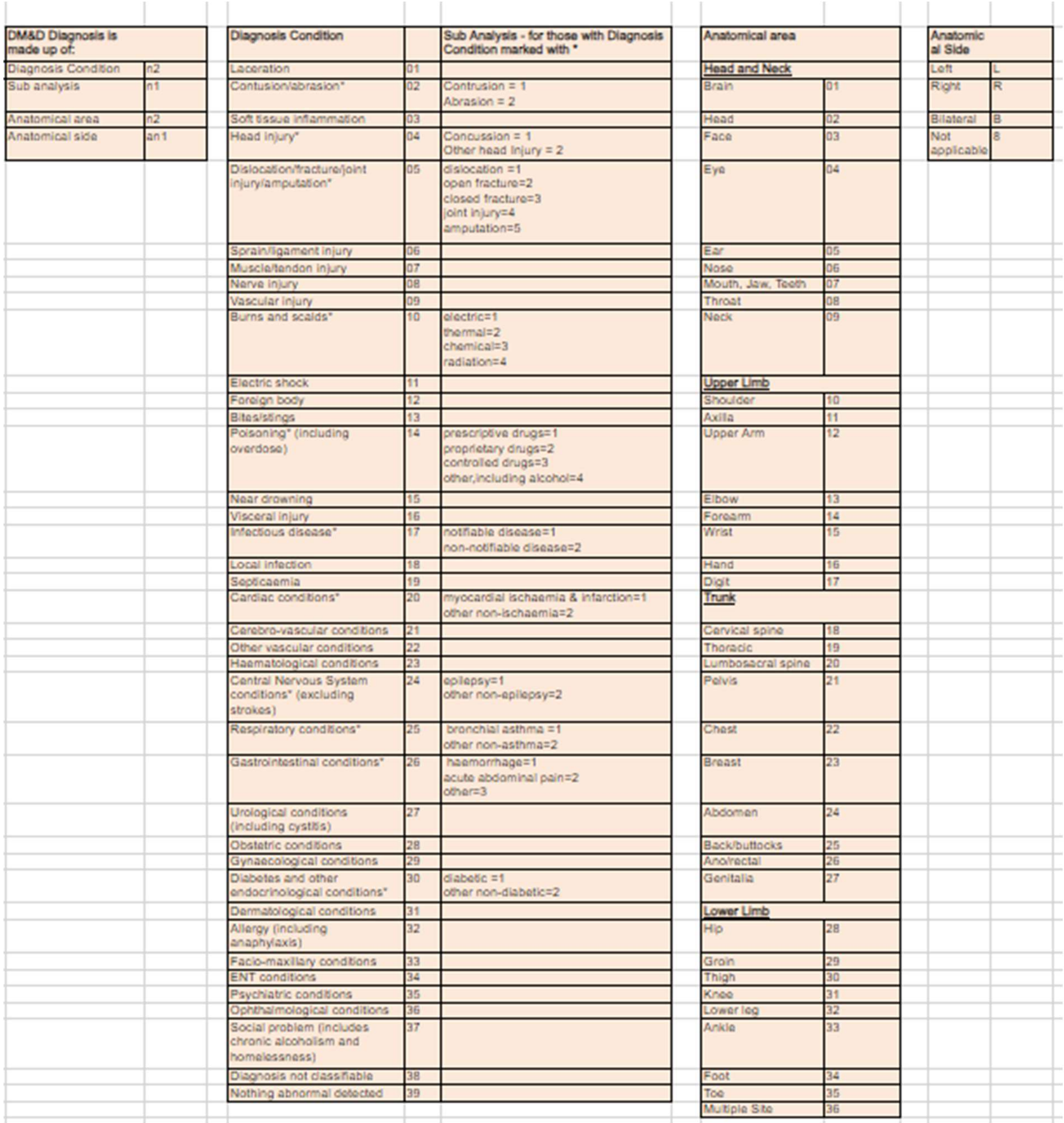
Xu, J. *et al.* (2009) “An Improved Decision Tree Algorithm on Bayesian,” *International Workshop on Intelligent Systems and Applications* [Preprint]. Available at: https://doi.org/10.1109/iwisa.2009.5072715.

Zmiri, D., Shahar, Y. and Taieb-Maimon, M. (2012) “Classification of patients by severity grades during triage in the emergency department using data mining methods,” *Journal of Evaluation in Clinical Practice*, 18(2), pp. 378–388. Available at: https://doi.org/10.1111/j.1365-2753.2010.01592.x.

Zou, J. and Schiebinger, L. (2018) “AI can be sexist and racist — it’s time to make it fair,” *Nature*, 559(7714), pp. 324–326. Available at: https://doi.org/10.1038/d41586-018-05707-8.

Appendices

Appendix I – Diagnosis description dictionary



DM&D Diagnosis is made up of:	
Diagnosis Condition	n2
Sub analysis	n1
Anatomical area	n2
Anatomical side	an1

Diagnosis Condition		**Sub Analysis - for those with Diagnosis Condition marked with ***
Laceration	01	

Contusion/abrasion*	02	Contrusion = 1
		Abrasion = 2

Soft tissue inflammation	03	
Head injury*	04	Concussion = 1 Other head Injury = 2
Dislocation/fracture/joint injury/amputation*	05	dislocation =1 open fracture=2 closed fracture=3 joint injury=4 amputation=5
Sprain/ligament injury	06	
Muscle/tendon injury	07	
Nerve injury	08	
Vascular injury	09	

Burns and scalds*	10	electric=1
		thermal=2
		chemical=3
		radiation=4

Electric shock	11	
Foreign body	12	
Bites/stings	13	
Poisoning* (including overdose)	14	prescriptive drugs=1 proprietary drugs=2 controlled drugs=3 other,including alcohol=4
Near drowning	15	
Visceral injury	16	
Infectious disease*	17	notifiable disease=1 non-notifiable disease=2
Local infection	18	
Septicaemia	19	
Cardiac conditions*	20	myocardial ischaemia & infarction=1 other non-ischaemia=2
Cerebro-vascular conditions	21	
Other vascular conditions	22	
Haematological conditions	23	
Central Nervous System conditions* (excluding strokes)	24	epilepsy=1 other non-epilepsy=2
Respiratory conditions*	25	bronchial asthma =1 other non-asthma=2

Gastrointestinal conditions*	26	haemorrhage=1
		acute abdominal pain=2
		other=3

Urological conditions (including cystitis)	27	
Obstetric conditions	28	
Gynaecological conditions	29	
Diabetes and other endocrinological conditions*	30	diabetic =1 other non-diabetic=2
Dermatological conditions	31	
Allergy (including anaphylaxis)	32	
Facio-maxillary conditions	33	
ENT conditions	34	
Psychiatric conditions	35	
Ophthalmological conditions	36	
Social problem (includes chronic alcoholism and homelessness)	37	
Diagnosis not classifiable	38	
Nothing abnormal detected	39	

Anatomical area	
Head and Neck	
Brain	01
Head	02
Face	03
Eye	04
Ear	05
Nose	06
Mouth, Jaw, Teeth	07
Throat	08
Neck	09
Upper Limb	
Shoulder	10
Axilla	11
Upper Arm	12
Elbow	13
Forearm	14
Wrist	15
Hand	16
Digit	17
Trunk	
Cervical spine	18
Thoracic	19
Lumbosacral spine	20
Pelvis	21
Chest	22
Breast	23
Abdomen	24
Back/buttocks	25
Ano/rectal	26
Genitalia	27
Lower Limb	
Hip	28
Groin	29
Thigh	30
Knee	31
Lower leg	32
Ankle	33
Foot	34
Toe	35
Multiple Site	36

Anatomical Side	
Left	L
Right	R
Bilateral	B
Not applicable	8

Figure I.1: Diagnosis description dictionary

Appendix II – Complaints mapping dictionary

Original complaint	Mapping
abdo loin pain	**abdominal pain**
abdominal	abdominal pain
abdominal pain-	abdominal pain
abdominal pain-+*	abdominal pain
abdominal pain--	abdominal pain
abdominal pain1`0	abdominal pain
abdominal pain3+	abdominal pain
abdominal pain7	abdominal pain
abdominal pain`	abdominal pain
abdominal pain``1```	abdominal pain
abdominal painder	abdominal pain
abdominal painpur	abdominal pain
abdominal painready	abdominal pain
abdominal painwzssze	abdominal pain
abdominal pain	abdominal pain
alleged assault0	alleged assault
alleged assault[	alleged assault
alleged assaultpe	alleged assault
arm injury	head injury
back problem---	back problem
back problem/	back problem
back problem0	back problem
back problem`	back problem
bleeding p/r	pv bleeding
buttock pain	back pain
catheter problempar	catheter problem
chest injury	head injury
chest pain/problem+	chest pain/problem
chest pain/problem..	chest pain/problem
chest pain/problem0	chest pain/problem
chest pain/problem50	chest pain/problem
chest pain/problem5116870	chest pain/problem
chest pain/problem`	chest pain/problem
chest pain/problemed	chest pain/problem
chest pain/problemr	chest pain/problem
chest pain/problemx xcc	chest pain/problem
chest problem	chest pain/problem
collape	collapse
collapse+	collapse
collapse.	collapse
collapse.+	collapse
collapse.0	collapse
collapsed	collapse
collapsej	collapse
convulsion/fitm	convulsion/fit
convulsion/fits	convulsion/fit
diarrhoea &/or vomiting6852	diarrhoea &/or vomiting
diarrhoea &/or vomitingb	diarrhoea &/or vomiting
diarrhoea &/or vomitingj	diarrhoea &/or vomiting
diarrhoea &/or vomitingp --	diarrhoea &/or vomiting
ear injury	head injury
eye problem	eye problems

eye problems---	eye problems
facial problem++	facial problem
facial problem--	facial problem
facial problem676	facial problem
fall	falls
falls+	falls
falls\	falls
fallsch	falls
fallsm	falls
foreign body in foot	foreign body
foreign body.	foreign body
foreign body0	foreign body
foreign body3	foreign body
groin pain	loin pain
hand injury	head injury
head inju	head injury
head injury.	head injury
head injury0	head injury
head injury5048354	head injury
head injurys	head injury
head pain	abdo pain
hip pain -'	hip pain
hip problems	limb problems
knee pain	neck pain
leg injury	limb injury
limb in	limb injury
limb injury*	limb injury
limb injury-	limb injury
limb injury./ ./	limb injury
limb injury0.	limb injury
limb injury6	limb injury
limb injury7	limb injury
limb injury89	limb injury
limb injury`	limb injury
limb injuryman	limb injury
limb injuryr	limb injury
limb injurys	limb injury
limb problem	limb problems
limb problems-	limb problems
limb problems.	limb problems
limb problems0	limb problems
limb problemsd	limb problems
limb problemsh	limb problems
lower abdo pain	abdo pain
lower back pain	back pain
mental health problemg	mental health problem
nail injury	truncal injury
p/v bleeding	pv bleeding
patient recently on chemo.	patient recently on chemo
patient recently on chemo0	patient recently on chemo
patient recently on chemo1q	patient recently on chemo
patient recently on chemo8	patient recently on chemo
patient recently on chemoxf	patient recently on chemo

personal problem````````````````	personal problem
personal problemkawoo	personal problem
poisoning?	poisoning
post op problem+-/	post op problem
post op problemxxxxx	post op problem
pr bleeding	pv bleeding
pregnancy relat	pregnancy related
pregnancy relatedgoing	pregnancy related
pv bleed	pv bleeding
pyrexiaal	pyrexia
rashes/skin problem```	rashes/skin problem
rashes/skin problemne	rashes/skin problem
rib injury	limb injury
self harm w	self harm
self harm/	self harm
self harmjj	self harm
shoulder injury back pain	shoulder injury
shoulder injury+	shoulder injury
shoulder injury0.	shoulder injury
shoulder injury[	shoulder injury
shoulder pain	shoulder injury
sob+	sob
sob, abdominal pain	abdominal pain
sob-	sob
sob.	sob
sob.0	sob
sob32	sob
sob6	sob
sob;	sob
sob`	sob
sob~	sob
sore throat848726	sore throat
syncope/abdominal pain	abdominal pain
testicular lump	testicular pain
testicular pain 3/52	testicular pain
testicular paing	testicular pain
truncal injuryw	truncal injury
unwel	unwell
unwell `````````````````````````````````````` ```` ``	unwell
unwell uti	unwell
unwell m	unwell
unwell nb,	unwell
unwell#	unwell
unwell+	unwell
unwell,	unwell
unwell-	unwell
unwell-*	unwell
unwell.	unwell
unwell.0	unwell
unwell0	unwell
unwell00	unwell
unwell0ws+	unwell
unwell2hrs	unwell
unwell3	unwell
unwell`	unwell

unwelld	unwell
unwellk	unwell
unwellsh	unwell
urinary problems.	urinary problems
urinary problems0	urinary problems
urinary problems`	urinary problems
urinary problemsn	urinary problems
v q1` 00abdominal pain8	abdominal pain
vomiting blood0	vomiting blood
vomitting	vomiting
weakness of one sidea	weakness of one side
worried parentas	worried parent
wounds0	wounds

Table II.1: Complaints mapping dictionary

Appendix III – Performances of every classification algorithm

Binary classification

	Classification Report		
	Precision	*recall*	*f1-score*
0	0.84	0.75	0.79
1	0.77	0.85	0.81
Accuracy	0.80		
weighted avg	0.81	0.80	0.80

Table III.1: Naive bayes classifier binary classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.78	0.82	0.80
1	0.81	0.76	0.79
Accuracy	0.79		
weighted avg	0.79	0.79	0.79

Table III.2: Logistic regression binary classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.87	0.83	0.85
1	0.83	0.88	0.85
Accuracy	0.85		
weighted avg	0.85	0.85	0.85

Table III.3: Grid search decision tree binary classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.87	0.86	0.86
1	0.86	0.87	0.86
Accuracy	0.86		
weighted avg	0.86	0.86	0.86

Table III.4: Bagging decision tree binary classification performances

	Classification Report		
	precision	*recall*	*f1-score*
0	0.83	0.80	0.81
1	0.81	0.84	0.82
Accuracy	0.82		
weighted avg	0.82	0.82	0.82

Table III.5: AdaBoost decision tree binary classification performances

Classification Report			
	precision	*recall*	*f1-score*
0	0.87	0.83	0.85
1	0.84	0.88	0.86
Accuracy	0.85		
weighted avg	0.85	0.85	0.85

Table III.6: Bayes Optimized decision tree binary classification performances

Classification Report			
	precision	*recall*	*f1-score*
0	0.86	0.86	0.86
1	0.86	0.86	0.86
Accuracy	0.86		
weighted avg	0.86	0.86	0.86

Table III.7: Bayes Optimized bagging decision tree binary classification performances

Classification Report			
	precision	*recall*	*f1-score*
0	0.84	0.81	0.82
1	0.81	0.85	0.83
Accuracy	0.83		
weighted avg	0.83	0.83	0.83

Table III.8: Bayes Optimized AdaBoost decision tree binary classification performances

Classification Report			
	precision	*recall*	*f1-score*
0	0.84	0.81	0.82
1	0.81	0.85	0.83
Accuracy	0.83		
weighted avg	0.83	0.83	0.83

Table III.9: Grid search random forest binary classification performances

Multilevel classification task – Three levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.80	0.72	0.76
1	0.75	0.69	0.72
2	0.58	0.75	0.65
Accuracy	0.72		

weighted avg	0.74	0.72	0.72

Table III.10: Naive bayes classifier multilevel classification performances – Three levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.82	0.86	0.84
1	0.82	0.75	0.78
2	0.84	0.82	0.83
Accuracy	0.82		
weighted avg	0.82	0.82	0.82

Table III.11: Grid search decision tree multilevel classification performances - Three levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.81	0.87	0.84
1	0.82	0.75	0.78
2	0.86	0.80	0.83
Accuracy	0.82		
weighted avg	0.82	0.82	0.82

Table III.12: Bagging decision tree multilevel classification performances - Three levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.77	0.84	0.80
1	0.75	0.70	0.72
2	0.80	0.72	0.76
Accuracy	0.77		
weighted avg	0.77	0.77	0.77

Table III.13: AdaBoost decision tree multilevel classification performances - Three levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.82	0.86	0.84
1	0.82	0.75	0.78
2	0.84	0.82	0.83
Accuracy	0.82		
weighted avg	0.82	0.82	0.82

Classification Report			
	precision	*recall*	*f1-score*
0	0.81	0.87	0.84
1	0.82	0.75	0.78
2	0.86	0.80	0.83

Accuracy	0.82		
weighted avg	0.82	0.82	0.82

Table III.14: Bayes optimized decision tree multilevel classification performances - Three levels

Table III.15: Bayes optimized bagging decision tree multilevel classification performances - Three levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.77	0.84	0.80
1	0.75	0.70	0.72
2	0.80	0.72	0.76
Accuracy	0.77		
weighted avg	0.77	0.77	0.77

Table III.16: Bayes optimized AdaBoost decision tree multilevel classification performances - Three levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.83	0.86	0.85
1	0.80	0.77	0.79
2	0.85	0.80	0.83
Accuracy	0.83		
weighted avg	0.83	0.83	0.83

Table III.17: Grid search random forest multilevel classification performances - Three levels

Multilevel classification task – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.81	0.67	0.74
1	0.54	0.49	0.51
2	0.45	0.74	0.56
3	0.45	0.45	0.45
Accuracy	0.61		
weighted avg	0.65	0.61	0.62

Table III.18: Naive Bayes classier multilevel classification performances – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.81	0.87	0.84
1	0.56	0.70	0.62
2	0.79	0.81	0.80
3	0.66	0.29	0.41
Accuracy	0.74		
weighted avg	0.74	0.74	0.73

Table III.19: Grid search bagging decision tree multilevel classification performances – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.83	0.72	0.77
1	0.50	0.51	0.50
2	0.74	0.78	0.76
3	0.44	0.59	0.50
Accuracy	0.67		
weighted avg	0.70	0.67	0.68

Table III.20: Grid search AdaBoost decision tree multilevel classification performances – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.79	0.88	0.83
1	0.55	0.67	0.60
2	0.79	0.79	0.79
3	0.68	0.28	0.39
Accuracy	0.73		
weighted avg	0.73	0.73	0.72

Table III.21: Bayes optimized decision tree multilevel classification performances – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.78	0.89	0.83
1	0.57	0.65	0.60
2	0.79	0.79	0.79
3	0.67	0.28	0.39
Accuracy	0.73		
weighted avg	0.73	0.73	0.72

Table III.22: Bayes optimized bagging decision tree multilevel classification performances – Four levels

	Classification Report		
	precision	*recall*	*f1-score*
0	0.70	0.81	0.75
1	0.49	0.39	0.43
2	0.70	0.60	0.65
3	0.41	0.36	0.38
Accuracy	0.63		
weighted avg	0.62	0.63	0.62

Table III.23: Bayes optimized AdaBoost decision tree multilevel classification performances – Four levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.79	0.88	0.83
1	0.59	0.58	0.58
2	0.79	0.79	0.79
3	0.58	0.37	0.45
Accuracy	0.73		
weighted avg	0.72	0.73	0.72

Table III.24: Grid search random forest multilevel classification performances – Four levels

Classification Report			
	precision	*recall*	*f1-score*
0	0.82	0.86	0.84
1	0.57	0.60	0.58
2	0.79	0.81	0.80
3	0.57	0.43	0.49
Accuracy	0.73		
weighted avg	0.73	0.73	0.73

Table III.25: Bayes optimized random forest multilevel classification performances – Four levels.