Explainable and Adaptive AI For Predicting and Profiling Patients Visiting NHS Emergency Departments: An Early Exploratory Study

Produced by Megan Lind Morgan, 2142745
Supervised by Prof. Jiaxiang Zhang and Dr. Alma Rahat

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Master of Science

Department of Computer Science
Swansea University

October 1st, 2023

Declaration

This work has not been previously accepted in substance for any degree and is not being con- currently submitted in candidature for any degree.

Signed M L Morgan

Date 30/9/23

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed M L Morgan

Date 30/9/23

Statement 2

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed M L Morgan

Date 30/9/23

Abbreviations

ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
ED	Emergency Department
NHS	National Health Service
RF	Random Forest
NN	Neural Network
KNN	K-Nearest Neighbour
SVM	Support Vector Machine
GAN	Generative Adversarial Network

Abstract

This project is an early exploratory study investigating the health, lifestyle and demographic factors that lead to an increased risk of hospitalisation, and whether traditional ML and DL models are able to use these factors to predict hospital length of stay. Insights from professionals within the NHS were gathered using a short survey, and findings from this were used to generate a synthetic dataset containing 8 features, with two labels: short or long stays. The chosen features were Age, Gender, Health Conditions, Mental Health, Smoking, Alcohol Consumption, Socio-economic Status and Exercise. Synthetic data was chosen as a basis for experimentation due to availability of sensitive data, and for this early proof-of-concept stage, the flexibility of synthetic data is beneficial.

Data generation was carried out using Python, and four models were tested: Random Forest, KNN, SVM and Neural Network. The Random Forest performed best with a F1 score of 0.81, with SVM performing worst with a F1 score of 0.61. Each model was analysed using F1 score and confusion matrices. None of the models performed poorly, demonstrating that machine learning is a viable approach for predicting length of stay and supporting effective allocation of resources in NHS emergency departments. Limitations of this work include the lack of correlation between features in the synthetic dataset, which does not fully reflect empirical data.

Future work will develop a more sophisticated model to predict and profile patients visiting the ED, using data from Hywel Dda NHS Health Board. This work serves as a starting point, with considerations for which models should be chosen, societal factors leading to increased patient risk, and the potential benefits of AI within a healthcare setting.

1. Introduction

This chapter will introduce the project, including a brief summary of the work, the motivation, aims and objectives for this MSc component, a brief description of the human-centred element of the work, and the hypothesis.

1.1. Problem Statement

This project serves as an introduction into the wider PhD project, which seeks to predict and profile patients likely to attend NHS EDs at particular times, with the aim of reducing unknown variables, and allowing for effective allocation of resources and staff. This MSc component is a short exploratory study considering the potential of various ML models in predicting length of stay of patients. Access to NHS data is currently unavailable, meaning synthetic data must be used as a proof of concept. Due to the ability to create custom data, and as the purpose of the project is to gain some insight into the contributing factors leading to hospitalisation and the role of ML in supporting decision making, length of stay was chosen as a prediction goal due to simplicity and its relevance in literature. Future work will use the findings from this project to identify suitable models, and the features and interesting patterns identified will be a starting point into investigating the profiles of patients.

1.2 Motivation

NHS emergency departments are facing significant demand, with many departments overworked and understaffed[1]. Current targets state that patients should be treated, discharged or admitted to hospital for further treatment within 4 hours of arrival at the department, a target that has not yet been reached at a national level[2]. As of 2022, this target has not been met in any month since July 2015[3]. Despite this, overall patient sentiments are positive, and medical staff work hard to provide good care, indicating that this issue can only be resolved by reconsidering how we allocate resources and predict demand.

Waiting times are consistently increasing, which is no surprise considering the number of attendees at emergency departments is also consistently increasing, with the annual number of attendees increasing by over 4 million from 2011 to 2019[4]. Certain times of the year are particularly challenging, such as winter[5], although many of the other factors that influence demand and increase the likelihood of emergency department attendance may not be obvious.

Artificial intelligence could become a crucial tool in detecting patterns and influencing factors that can be used to predict when patients will attend the emergency department[6], which would allow for effective allocation of suitable staff, reducing waiting times and supporting the human practitioners. By relying on the flexibility and adaptiveness of the staff alone, the issue will not be resolved, therefore a tool that can work alongside medical staff to improve awareness and reduce unexpected strain on the service would be valuable.

This project serves as a starting point for further work, which will develop an adaptive and explainable AI capable of predicting the profiles of patients attending the ED. The first step is to assess the suitability of various machine learning models as predictive tools in a health context, which will be the focus of this project. A survey will be used to understand influencing factors that increase the risk of hospitalisation, such as lifestyle and demographic, which will be completed by professionals in the health service. This will provide insight into the profiles of patients and the meaning behind the data, which will aid in creating a fair and reliable model. The goal of this project is to create a human-centred system that provides information and does not endanger anyone or take away any human freedom.

1.3 Project Aims and Objectives

This project serves as a starting point into the investigation regarding the potential of machine learning in predicting trends in hospital data, which in the case of this project, will be the length of stay of patients within different demographics.

As of the time of this project's beginning, the process of gaining access to official health data from Hywel Dda is underway but not finalised, therefore it is necessary to create synthetic data for the purpose of testing out machine learning approaches to predicting length of stay. This data must be realistic and fit-for-purpose, therefore we will conduct a short survey targeted at those working in the health profession to discover influencing factors regarding hospitalisation, such as demographics, lifestyle factors and health conditions. The results of this survey will then be used to identify features to be included in the dataset.

This synthetic data will then be used to test out a variety of machine learning models, such as Random Forests, SVM and Neural Networks. The hope is that these models will be able to predict length of stay based on the patient's background information. The strengths and weaknesses of each model will be evaluated, and the results of this work will be used as a starting point for the wider project, which will entail creating an AI model to predict and profile patients entering emergency departments.

The aims and objectives of the project will be summarised below.

1.3.1 Aims

- To consider and discover factors, through a survey and data generation, that increase the risk of hospitalisation and influence length of stay.
- To test the efficacy of machine learning models in predicting length of stay. These include regression, neural networks, SVM and decision trees.
- To determine which approaches are most suitable for future work and how this work can be expanded upon to predict patient profiles entering emergency departments.

1.3.2 Objectives

- Conduct research into hospital patient profiles and the role of AI in current approaches to optimising hospital processes.
- Conduct a thorough literature review to assess current work and successful approaches, as well as any benchmarks.
- Create and distribute a survey targeted at those working in the health profession to discover factors that increase the risk of hospitalisation.
- Generate synthetic data using survey results and background research.
- Develop, train and test a range of machine learning models.
- Compare the accuracy and performance of the models and determine which ones are most suitable for the task.
- Discuss the results and how these will be used to start work on the next, wider issue of predicting the profiles of patients attending the ED.

1.4 Human-Centred Design

Human-centred design places humans and their needs at the heart of the design and development process, with considerations for ethics, requirements and the potential impact of the final product[7]. By following this design process, we can ensure that developments do not harm people or society, and follow responsible innovation procedures. This includes reducing environmental impact and preventing harm as much as possible[8]. Methods of human-centred design include surveys, interviews, user studies, iterative design, and evaluation of project outcomes in regards to intended impact. This project is focused on the use of artificial

intelligence for healthcare purposes, which is a sensitive topic to many due to data privacy and concerns surrounding the reliability of AI in critical settings[9].
The intended impact of this project is to explore the potential of machine learning models in predicting key statistics within the NHS, focusing on the length of stay of patients in hospital EDs. This project is not designed to replace any human jobs, and will not independently diagnose or treat patients, meaning the risk of harm is low. A survey will be used to gather responses from human professionals, which will assist in prioritising certain features. There is no potential for bias currently, although this will be evaluated and addressed during future work.

1.5 Hypothesis

The working hypothesis is that machine learning will be a viable solution for predicting length of stay. It is expected that some models will outperform others, such as Random Forest and Neural Network. It is also expected that the survey responses will demonstrate that there are many different factors leading to longer hospital stays, which will require a model capable of handling complex relationships, with this project proving the potential for machine learning's predictive power.

The following chapters will discuss the background research and literature review, which will shape the design of the survey, models and data.

2. Background Research

This chapter will discuss the relevant background research carried out prior to work commencing, focusing on key topics included within the project, such as synthetic data generation, an overview of the important AI concepts required, the role of emergency departments and current issues with the length of hospital stays.

2.1 Synthetic Data

Synthetic data is generated by developers for the purpose of data intensive tasks such as deep learning, and this technique is becoming more prominent as artificial intelligence becomes ingrained within applications. The process involves generating realistic and diverse data points that closely mimic genuine datasets created from observed data[10]. This can be done in multiple ways: one popular method is to utilise a pre-existing dataset and extend it to include more data by following the trends and distribution of the existing data. Alternatively the synthetic data can be generated on its own using probability distributions, which could be found using literature or analysing existing datasets that are used for similar tasks. Novel methods such as Generative Adversarial Networks have also gained prominence, which generate data by learning patterns within a real dataset, but as with the first method, this relies on having access to an existing dataset[11, 12]. The crucial requirement for each method is to replicate the important statistical properties of real data, so that the models' performance is an accurate representation of its ability to classify and handle real data.

The primary use for synthetic data is to increase the quantity of available data, as machine learning tasks require large datasets to effectively train models, and it is necessary to have unseen data for testing the model's performance[13]. It can also be used to improve the fairness and robustness of the dataset, as real datasets often have skewed data that favours dominant groups within the world, which can lead to biased models. Artificially generated data can also provide insight into rare events, such as rare diseases or natural occurrences with limited observations. There are also other benefits, such as improved privacy, as sensitive data does not need to be used for development[14]. This is particularly useful in domains such as health data science. Synthetic generation also allows for the flexible expansion of the dataset, and noise can be included to increase the difficulty of the task, which allows for testing of the model.

This project will employ synthetic data generation to create a dataset of anonymous patients that have visited an emergency department. This will include features focused on lifestyle and demographic, with the patient's length of stay as the target label. Patient data is highly sensitive and requires extra levels of access, therefore real data was not available for this stage of the project. The artificial data will be used as proof of concept when training and testing deep machine learning models, with the aim of using these findings as a basis for creating models for use with empirical NHS data.

There are limitations to this approach, particularly the possible difficulty in translation between real and artificial data. A model trained using one may not perform well with the other, and synthetic data may not always be representative of real data, especially in the case of rare observations. Care must also be taken to ensure that the data is diverse and an accurate representation of ground truth trends[15, 16]. Also, despite the potential for artificial data to eliminate bias, the opposite could also be true, as developers must ensure that their own bias does not influence the data. Furthermore, real data will contain naturally observed outliers, which will likely be excluded from the synthetic dataset.

2.2 Machine Learning

2.2.1 Overview of Machine Learning and Artificial Intelligence

Machine learning and Artificial Intelligence are very similar, with ML serving as an element of AI. ML is a specific application of AI that is used for pattern recognition, classification and general learning using data, with specific models/algorithms utilised. These adaptable models can be trained and used within industry or research to find relationships within data and act as the 'learning' aspect of AI, producing effective predictions and simulating a human's way of thinking. Applications of this also include text and speech recognition, with Alexa being an example of a ML-based device that interacts in a human manner, and in this example the ML is responsible for learning and producing output, whereas the wider field of AI produces a complex, human-like device that is capable of interacting with humans in an effective way[17, 18, 19]. Turing, considered a founding father of AI, defines the discipline: *"AI is the science and engineering of making intelligent machines, especially intelligent computer programs.*[20]*"* To produce machines and software capable of interacting with and reacting like humans requires a wider consideration of psychology, sociology, biology, product design, linguistic, legal issues and more[21], therefore AI is not solely based on Computer Science. ML on the other hand falls under CS, with a mathematical and programming based approach that underpins the technical aspect of AI. This project focuses on ML, experimenting with different models to assess their suitability for use within the NHS. To extend this work and begin developing an AI system, more work will be done to understand users' needs, develop an interactive interface, and consider the communication style between the users and the AI system, which will hopefully be capable of independent learning on ever-changing data.

2.2.2 Adaptive AI

Adaptive AI is capable of adapting and learning over time, with its performance consistently evaluated using feedback. This system should become more autonomous, capable of handling new, unseen data that changes frequently, with a larger element of independent learning[22]. An adaptive model is required for a task that changes frequently, with NHS emergency departments being a prime example of this. Many health issues are unexpected and the global health scene may unexpectedly change, with COVID being a prime example of an unexpected and major health concern[23]. AI that is not adaptive will rely on past data and experience to perform effectively, which is too restrictive in a tumultuous environment, therefore work must be done to ensure the AI is capable of changing and discovering new trends. This will not be covered within the scope of this MSc project, but it is an important aspect of work going forwards.

2.2.3 Explainable AI

Trust is an important consideration when developing AI systems, with many people feeling hesitant to rely on unknown decision making processes, particularly in critical areas such as healthcare. Explainable AI allows humans to understand the underlying processes and important factors used in decision making, which will hopefully improve trust, and holds the model accountable[24,25]. This is also important in addressing issues such as bias, as the reasons behind decisions will allow humans to identify unfair practices, a popular example being the COMPAS dataset and its unfair risk classification of Black inmates[26]. Black-box algorithms have been used frequently in the past, which means the workings and processes of the algorithm cannot be accessed or understood. We are now working towards incorporating white box algorithms, or XAI, into all AI systems, so that trust and reliability can be improved[27].

2.3 Emergency Departments

NHS emergency departments are a crucial element of healthcare, with consistent staffing and no requirement for appointments, meaning patients arrive daily, at any hour, for the treatment of serious accidents or unexpected health issues such as heart attacks. Some units specialise in minor injuries such as breaks or cuts, whereas many departments offer general emergency care[28, 29]. For the majority of non-minor injury departments, patients present with life-threatening conditions, requiring specialist care, which is a large resource for the NHS. In the ED, patients are triaged, usually by a nurse, and their recorded notes are then passed on to a doctor. The patient will then see the assigned doctor, and depending on the severity of the condition, will either wait in the waiting room or a small, separate room. The doctor can then advise whether any scans or tests, such as X-Ray or blood tests, are required, which will normally incur an additional wait. Sometimes, the issue can be identified easily and medication or other treatment can be administered, leading to discharge. Other times, the patient will need to be monitored or the treatment will prove to be complex, meaning they will have to stay on the ward, which will require an available bed and around-the-clock care. Although the ED is defined as treating urgent and life threatening conditions, during 2016-17 in England, up to 16% of attendances were considered non-urgent, and another study by the NIHR found that 20% of non-urgent attendees arrived by ambulances[30,31], which will place strain on a service that is stretched thin. This could indicate that messaging around the purpose of the ED is unclear, or that a more effective method of predicting demand and allocating the proper treatment or interventions early could lead to better resource allocation within the ED itself.

2.4 Hospital Stays

Length of stay is often used as a quality indicator for emergency departments, with a 4 hour turnover time goal imposed by the government[32]. Long stays are associated with an inefficient department that is unable to treat patients quickly, but length of stay can also be impacted by the patient's profile, as will be explored within this project.

The NHS has launched a scheme 'Reducing Length of Stay' which aims to discharge patients quickly without unnecessary delays, particularly as elderly or frail patients may find waiting uncomfortable. Despite these efforts, in February 2023, 125,505 patients waited 12 hours or more to receive treatment, and the percentage of departments hitting the four hour goal has decreased[33,34]. Many departments report 3-4 hours as the average waiting time, so this could be seen as a short to intermediate length of stay. Longer waiting time in the ED leads to an increase in mortality, and some patients even report foregoing treatment due to an inability to wait to be seen[35]. It is evident that health professionals are committed to providing excellent care and will work hard to ensure everyone is seen, but due to demand, understaffing or lack of resources such as available beds, people are unable to be handled effectively[36]. It is our hope

that an AI tool capable of predicting demand will support clinicians in their time and resource management.
The following chapter will be a literature review of related work.

3. Literature Review

This chapter will review related work, with a focus on machine learning within healthcare. Whereas the previous chapter has summarised background research in relation to the key topics addressed as part of the project, this will compare and consider results from academic pieces of work with similar objectives to this project. Search results have been filtered to only include papers published since 2019 to ensure current relevance.

3.1 Summary

Computing and technology has played a vital role in healthcare since the 1950s, when the first systems were trialled[37]. Since then, their role has adapted from data storage and payroll management to include recommendations, electronic health records, and now the prediction of health issues and drug discovery[38].
The efficacy of machine learning in predicting patient flow and other emergency department statistics has been evaluated by multiple researchers, with some prominent examples related to this project. Grant et al[39] investigate the classification of complex patient profiles using latent feature analysis and k-means clustering. They note that complex patients account for large costs in healthcare, and they conducted their study in Northern California, US. The healthcare system in the US differs from the UK, but the findings are still valid. Previous hospitalizations and health records were utilised, with a sample of 104,869 patients. Clustering and statistical analysis were able to find the top 3% most clinically complex patient profiles, which were then analysed by a panel of clinicians. The goal of this was to assess patient mortality and risk of hospitalisation, with key features leading to this complexity identified, such as lack of engagement, frailty, psychiatric illness and cardiovascular disease. A care plan was then devised for each patient.
Also focusing on patient outcomes, Raita et al[40] employed the use of multiple ML models to predict outcomes of patients visiting the ED. Lasso Regression, Neural Network, Gradient Boosted Decision Trees and Regression were trialled as tools for triage, and it was found that the

benefit of successful triage outweighed any over-triaging that occurred. Each model performed well and proved viable for assisting clinicians in triaging patients, and the model performed particularly well when predicting risk of critical care. This study differs from Grant's in method, as no clustering methods are used, and the model was designed for use within the ED itself rather than suggesting interventions prior to hospitalisation.

Shamout et al[41] worked on a ML model that could predict the deterioration of COVID-19 patients, a valuable tool for future pandemics. As with the aforementioned studies, this work focuses on patient outcomes, although the problem area is specific to COVID, meaning it may not be generalisable in its current state. It is also designed for use with existing, diagnosed patients, and does not identify those at risk prior to infection. Their deep neural network used data from over 3000 patient X-Rays and a gradient boosting model was trained using routine clinical variables. The model was successful in predicting many patient deteriorations within a 96 hour window, and when compared against human radiologists evaluating X-Rays, maintained its good performance. The model was deployed and tested in a New York hospital and was able to predict cases in real-time, highlighting that this approach could be developed and used on a more permanent basis.

Alvares-Chavez et al[42] discuss the importance of effective resource management in the ED, and the role of prediction tools in supporting this. They conducted a study using data from a Spanish civil and military hospital, with the aim of predicting ED attendance within the next 7 days, and the following 4 months. Data was aggregated on a daily basis, and two types of models were used: time series, and feature matrix, with ensembles of models tested to evaluate performance. They found that the prediction capability was beneficial, and worth exploring further, and that ensembles are of particular interest.

Jilani et al[43] investigate the prediction of long and short stays in hospitals, a project with similar aims to our own. They also discuss the potential for length of stay prediction for resource management, using historical data from 2011-2015 to develop a forecasting model. Weekdays and weekends were handled separately, and a fuzzy time series model was developed, which was tested in four different EDs. They found that hospital attendances are not random and do follow a pattern, meaning prediction of attendance is feasible. They also found that prediction accuracy increased when larger time intervals were used, such as using monthly rather than daily. Their heuristic model achieved higher accuracy than compared models, and serves as a basis for future work in implementing such a system.

Also investigating length of stay predictions, Kadri et al[44] compared a GAN model to traditional models such as neural networks and SVM, and found the GAN to be a reliable tool in predicting long or short stays. Experiments were conducted on data obtained from Lille Pediatric Department, and all models were found to have high accuracy, but the GAN did perform best when predicting which patients were likely to stay for a longer time. Kadri's model is more advanced and novel than those developed by Jilani, but both demonstrate the application of deep learning and statistical methods in LOS prediction.

3.2 Strengths and Weaknesses

Every study has strengths and weaknesses, regardless of its impact on the research area. By analysing these, we can learn and employ useful techniques within our own work.

An overall weakness of the included work is that many of the studies are not carried out within the NHS or the UK, which is not a direct weakness in terms of the quality of the study, but greater relevance to the NHS's process would be beneficial for our own project. A strength of each study is that they were all tested in healthcare settings, meaning the results are more reliable and are proven to be suitable for use alongside clinicians, which is a vital aspect of human-centred development.

A strength of Grant et al's work is that high dimensionality data was used, with many samples, improving reliability. They combined quantitative and qualitative methods, which will increase trust and possibly improve accuracy. They also identified interesting findings, such as specialised care plans being required for patients with multiple comorbid conditions, and that lack of patient engagement with health services leads to complexity in their health profile. A limitation of the study is that relying on clinical judgement could slow down the process as human clinicians are often busy, and a fully autonomous system may find interesting features quickly. Clustering methods can also be difficult to validate[45] when compared against other models, such as neural networks.

Alvarez-Chavez et al test time series and feature matrices, which differs from many other publications, which typically focus on common methods such as neural networks. This is a strength as it is vital to consider different approaches, and they also evaluated their models' performance using two different time intervals, which provides a more accurate depiction of results. Many of the algorithms used may not be chosen by many researchers, therefore the findings may not be utilised fully. Furthermore, only three years of data was used, which may limit the model's ability to adapt to new situations.

A strength of both Raita and Kadri et al is that multiple models are compared and evaluated, which is useful in supporting future work and highlights the importance of their own model.

Kadri found that deep learning models significantly outperformed traditional ML models, which is an expected finding. A weakness of Kadri's work, which is shared by Jilani et al, is that the specific LOS predictions could be refined into clearer time intervals, which would further support the allocation of resources. Jilani et al also note that they did not consider external variables in their predictions, such as climate, pollution or natural events, which may alter results. Raita et al state that they excluded data that contained some missing values, meaning some patients will be excluded from the model's predictions, and this should be addressed for use in a real hospital. They also discuss the differences between departments, and that their model may not generalise well to other EDs.

Shamout's work on COVID deterioration has many strengths, particularly its ability to successfully predict at-risk patients in a hospital during the pandemic, which goes beyond proof of concept and demonstrates its capability in a real, useful setting. The output of the model will save lives, and could be adapted to classify at-risk patients with other health conditions. Their dataset was limited, with only 3000 samples, which is a weakness - it is also a potential weakness if the model does not generalise well and cannot be used outside of COVID-19 predictions within this specific hospital. The model also relies on X-Ray images, and many conditions, COVID included, present differently across different patients and may not be reliably predicted using chest symptoms alone[46].

Every aforementioned study discusses the importance of improving allocation of resources, which demonstrates the importance of this work.

3.3 Implications of Research

These examples highlight the power of ML and AI within a healthcare setting, and demonstrate that it is worthwhile to develop and test new models that are capable of predicting patient deterioration, length of stay and patient flow. Each study notes the overcrowding present in EDs globally, and that prediction tools could assist in managing resources, leading to reduced waiting times, better care and less mortality. A range of methods are employed, with traditional models such as k-means, neural networks and tree learners yielding good results, and more novel approaches such as GANs performing very well and potentially paving the way to even more advanced predictions. Grant's work to identify complex patient profiles highlights the value of patterns within data, and how these can be utilised to improve care and identify weak areas within the healthcare system. Many patterns are difficult for humans to notice,

particularly with so much other work required, therefore using AI as a tool for identification and management of complex patients would be beneficial.
COVID-19 was an unexpected global health disaster, and symbolises unknowns that impact lives and force us to rethink our methods. AI that is adaptive could predict such events occurring, or at the very least support our management of these disasters, which is proven by Shamout's model. By predicting which patients are likely to deteriorate, lives can be saved and medical staff can anticipate issues and demand, reducing their workload and mental distress caused by unnecessary deaths. Healthcare is constantly evolving, and as the global population grows, so too do the strains placed upon hospitals. Kadri, Jilani and Alvares-Chavez each discuss the growing demand within the ED and the difficulty in allocating sufficient resources. Their work is valuable to consider within the scope of our own, as each successfully develops a model to predict either patient flow or length of stay to high accuracy. Their methods are complex, and were able to be tested in real hospitals, but their comments on the efficacy of traditional models indicate that highly complex solutions are not always needed. Raita's research also employs traditional models and yields good results when triaging patients, which is a valuable task that is often overlooked when considering processes within EDs. As our own work develops, GANs will be developed and tested, therefore Kadri's work is a valuable starting point, although the end goals will be different. Length of stay has been chosen as a focus for this early exploratory study due to its simplicity, and its relevance is demonstrated by the aforementioned studies. Each study highlights the benefits of AI in healthcare, and promises a more efficient, less unknown future.

This chapter has provided useful insight into related work and the implications of existing research. The following chapter will outline the project's methodology and timeline, and discuss the risks.

4. Project Plan

This chapter will discuss the methodology chosen for the project, alongside a loose plan and milestones that will be used to evaluate the project's progression. A risk assessment will also be conducted.

4.1 Methodology

There are three key parts to the project, all of which must be completed to ensure that the outcome is reliable and complete. These parts are: background research and literature review, the survey and its results, and the generation of data with the subsequent machine learning models. To maintain consistency and productivity, it was prudent to select a work methodology to support planning and delivery of project outcomes. For this, Agile was chosen, specifically Agile Scrum.

One of the key aspects of Agile is the breaking down of work into small, incremental chunks that can be worked on iteratively[47]. This is particularly useful for coding tasks, as code may require many iterations to find an optimal solution, which is especially true for machine learning development, which requires the tweaking of many parameters. Scrum is characterised by short 'sprints' as part of the work process, during which team members work on a set task for a specified amount of time, and then report on their progress to the wider team. This project does not require coordination from multiple team members, but work will be completed in fortnightly sprints, with progress reported during supervisor meetings. This allows for consistent feedback and the ability to improve chunks of work. All parts of the work will be revisited multiple times to ensure quality, although some elements of the work will inform future parts, such as the results of the survey shaping the data generation.

The Waterfall methodology is a popular alternative to Agile, but this is not appropriate for this project due to its linear approach to work, with no emphasis on iteratively improving components[48]. This approach would work well for a task that does not heavily rely on experimentation, but it is not sufficiently flexible for a machine learning project. The Waterfall methodology also has a defined end goal, which is better suited to development projects such as application development, rather than a project with variable outcomes depending on experimentation. Kanban was considered as an approach[49], but due to the limited team members, the use of the Kanban board was not necessary.

4.2 Project Timeline and Milestones

As stated, the project has three primary areas of work, although there are multiple milestones within each part. By considering the expected milestones and the overall planned timeline, we can ensure the project is developing as expected.

Part 1 Milestones:

- Problem space and related material fully researched.
- Literature review completed with sufficient number of papers.

Part 2 Milestones:

- Survey developed and distributed.
- Results from survey displayed and analysed.

Part 3 Milestones:

- Dataset created.
- Labels assigned to the dataset.
- Machine learning models trained and tested.
- Results from each model collated and analysed.

There are also smaller tasks as part of these milestones, and the dissertation will follow on from the results with a discussion about their implications, and suggestions for future work.

The project plan is difficult to define, especially given the iterative approach of Agile, but a loose plan has been created so that progress can be tracked to a reasonable degree. The plan is detailed below.

Project Start: June 1st

June - July: (Part 1)

- Identify the specific problem to be addressed.
- Set up meetings.
- Find papers related to ED waiting times, the use of ML in predicting LoS and patient flow.
- Save papers using Zotero.
- Research synthetic data generation.

July - August: (Part 2)

- Create and distribute survey.
- Start writing the dissertation, focusing on background work, project plan and literature review.

- Try different approaches to generating data.
- Analyse survey results.

August - September: (Part 3)

- Choose a set approach to generating data.
- Create synthetic dataset.
- Choose machine learning models to train.
- Train and test models.
- Write sections relating to data generation and the chosen models.

September - October: (Part 3 and final write-up)

- Tidy up code and check through results.
- Finalise experimentation and conclusions.
- Continue write-up to finish the dissertation.
- Check through the whole project.

Project End: October 2nd - Submit before this.

4.3 Risk Assessment

Prior to any project commencing, it is important to consider the potential risks involved, and how these may impact the project's development. It is also essential to consider any risk of harm to the developer or others[50].

The project is low in risk of harm, as the machine learning will be using synthetic data that is not connected to any individual, there is no risk of a data breach or mishandling of data. The code will be developed independently without a team, therefore there is no risk of conflict or miscommunication, although there will be regular supervisor meetings to ensure the project is developing as expected. There will also be no human participation other than the survey, which has been approved by Swansea University's ethics committee due to its anonymity and low risk of harm to any participants. It is also being completed remotely with no method of identification.

Any other potential risks have been outlined below, with likelihood and impact scored on a scale of 0 (lowest) to 5 (highest).

Risk	**Likelihood**	**Impact**	**Mitigation**
Computer breakdown.	1	3	The primary device is regularly checked and has no known issues. In the case of a break down, the

			university's devices could be used.
Data corruption	1	2	The data is synthetic, therefore it can be adapted and fixed with ease in the case of corruption or loss. Back-up data will be saved continuously.
Illness	3	4	Illness could impact the development of the project, therefore healthy practices should be followed, and alternative plans could be followed, such as prioritising key tasks.
Lack of survey responses	4	3	Ensure the survey is short and simple. In the case of insufficient responses, use literature review findings to bolster discoveries.
Poor results	4	5	The results may not be as expected, which would usually indicate an issue with the dataset or the problem. Ensure data is of high quality, and alter model parameters to find optimal performance. If needed, adapt problem or data if good results are not achievable.
Unreliable results (overfitting, difficult to replicate)	2	4	Consistently test model's performance, alter approach if overfitting is present. Ensure code is well organised and documented so that it can be replicated and tested.
Poor time management	3	5	Keep a diary of tasks and a document of weekly goals that can be reported during meetings. Discuss issues early. Follow Agile methodology to handle tasks in small chunks.
Poor problem definition	2	4	Read related work and discuss the problem with supervisor. Consider whether this is a viable approach early on by consulting with experts or literature. Utilise agile approach to iteratively work on parts of the project so that problem's definition can be evaluated frequently.
Misuse of findings	1	4	No impact on the project, but potential impact on wider society. Consider the goal of the project and the potential for societal harm. Consider potential biases and how this will be used.

Fig 1. Table of risks and mitigations.

All risks will be considered and even those with low likelihood could become prominent issues without sufficient mitigation and planning, therefore the progress of the project will be well documented. The risks will be evaluated frequently, and any issues will be reported quickly to avoid significant disruption.

This chapter has discussed the project plan and the methodology that will be used to ensure consistent development and organisation. These are vital aspects to the project that will support the desired outcomes being met on time. Potential risks have also been highlighted with appropriate mitigations suggested, so that obstacles encountered during the project can be effectively navigated with minimal impact. This is also important in ensuring no harm is caused.

The following chapter will discuss the project's method and design, including data, the survey and the machine learning models used.

5. Method and Design

This chapter will highlight the key methods and design processes employed during the project, including the survey development, data generation and the chosen machine learning models.

5.1 Survey

A human-centred approach is at the heart of the project, therefore it was prudent to include human practitioners within the project's design. To achieve this, we created and distributed a survey to gain insight into the opinions of those working within the NHS, namely their thoughts on factors that increase a patient's risk of hospitalisation. This data was then used as a vital aspect of the synthetic data generation, as it informed the feature columns, leading to a more realistic dataset built upon the expertise of the humans working within the NHS.

5.1.1 Participants

The goal of the survey was to gain insight into factors that can increase the risk of hospitalisation, therefore we required the expertise and knowledge of those working within the NHS to get an accurate picture. For this reason, the survey was exclusively distributed to individuals working within the health service, although a multitude of roles were covered, including data analysts, nurses and dentists. Subjects were found by distributing the survey into the NHS via Hywel Dda's data team.

Ethics were obtained through Swansea University's Research Ethics Committee. All participants were over the age of 18 and none were considered vulnerable participants. The survey was also completely anonymous, therefore no personal or sensitive data was handled. This also allowed participants to express opinions freely without concerns over their views being misconstrued or openly publicised. The survey consisted of 3 multiple choice questions with no harmful content, therefore the risk of harm was low.

The initial aim was to gather 10-15 responses but we obtained 24 unique respondents by the survey's closure. This provided useful results, which will be discussed in the Results chapter.

5.1.2 Survey Design

The primary goal of the survey was to uncover important features to include within the synthetic dataset, therefore three themes were chosen: Health Conditions, Demographics, and Lifestyle Factors. These themes are frequently mentioned in literature, and participants were given multiple choices, alongside an open option for additional thoughts. Health conditions are widely known to increase risk of illness, but the survey sought to find out which conditions were most commonly associated with hospitalisation, including Cardiovascular Diseases, Respiratory Diseases, Neurological Diseases, Mental Illness, Neurodevelopmental Differences and Cancer.

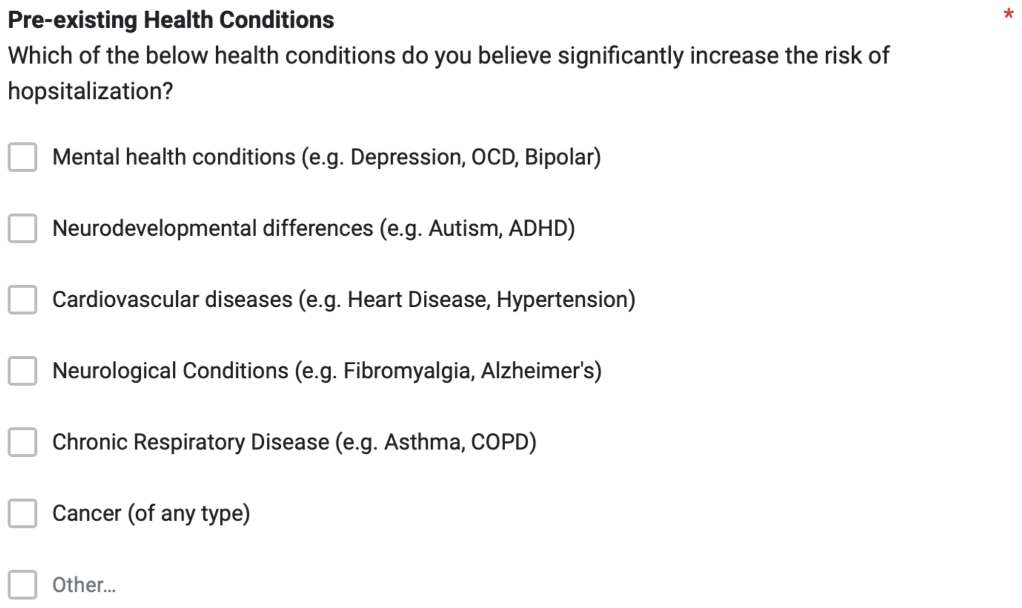
Demographics are frequently mentioned in literature, particularly the impact of poverty on health and access to healthcare. This part of the survey sought to discover the most influencing demographic factors, such as low/high socioeconomic status, race, gender, age and low/high levels of education. It was not expected that higher income and higher educational levels would lead to hospitalisation, but these were included for fairness. Age is commonly associated with hospitalisation, and gender has been mentioned within literature, particularly females being at increased risk[51,52].

The final section was concerned with Lifestyle Factors, which are alterable traits but are often seen as contributing factors to illness. The options here included levels of physical activity, special diets, poor diet (low nutrition), stress, smoking, recreational drug use and problem solving activities, the latter being included as a debiasing option, as the other options are largely considered negative lifestyle factors.

The survey was designed to be short so that working professionals would be more likely to fill it in. Participants could choose as many options as required, and consent was included at the start for ease. The results of the survey were intended for use when generating the synthetic dataset that, as the dataset had to be realistic and mirror real data, therefore the survey results could be utilised to choose features for the data, such as low levels of exercise and gender. These results could also be used in future work to inform decision making, and will serve as a basis of our understanding of the complexity and range of the task of predicting the length of stay of patients, moving forwards into predicting patient profiles in future work.

5.1.3 Survey Questions

The three questions included in the survey are pictured below, with each one focusing on a different aspect of risk elevation: health conditions, demographics and lifestyle factors. The information and consent form are not pictured, but these can be found in the accompanying materials.



Pre-existing Health Conditions *

Which of the below health conditions do you believe significantly increase the risk of hopsitalization?

□ Mental health conditions (e.g. Depression, OCD, Bipolar)

□ Neurodevelopmental differences (e.g. Autism, ADHD)

□ Cardiovascular diseases (e.g. Heart Disease, Hypertension)

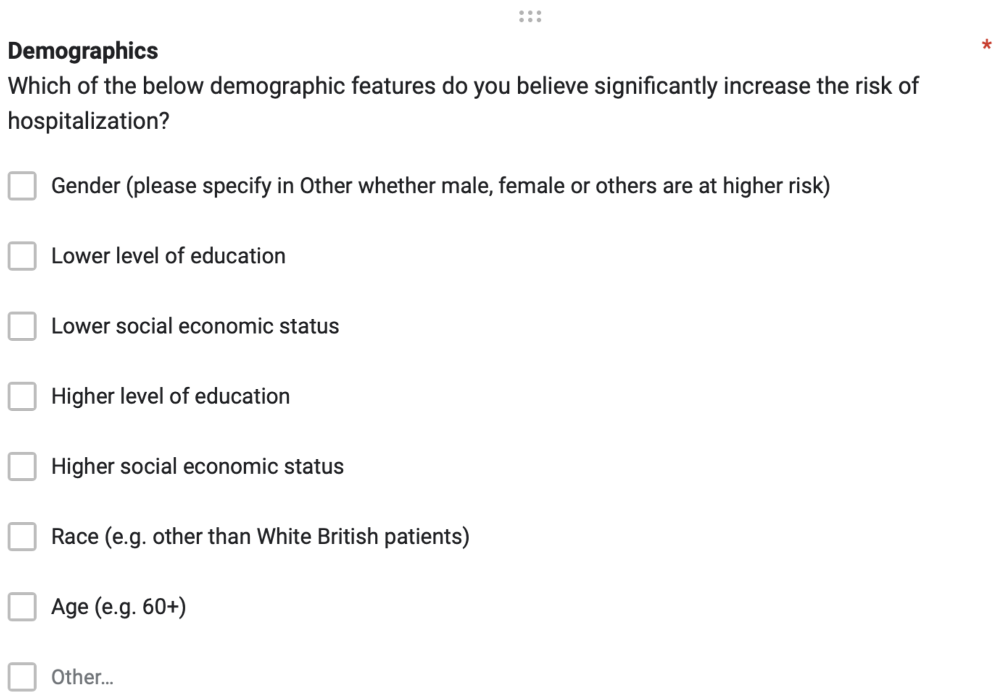
□ Neurological Conditions (e.g. Fibromyalgia, Alzheimer's)

□ Chronic Respiratory Disease (e.g. Asthma, COPD)

□ Cancer (of any type)

□ Other...

Fig 2. Question 1 of survey "Which of the below conditions do you believe significantly increase the risk of hospitalisation?"



Demographics *

Which of the below demographic features do you believe significantly increase the risk of hospitalization?

□ Gender (please specify in Other whether male, female or others are at higher risk)

□ Lower level of education

□ Lower social economic status

□ Higher level of education

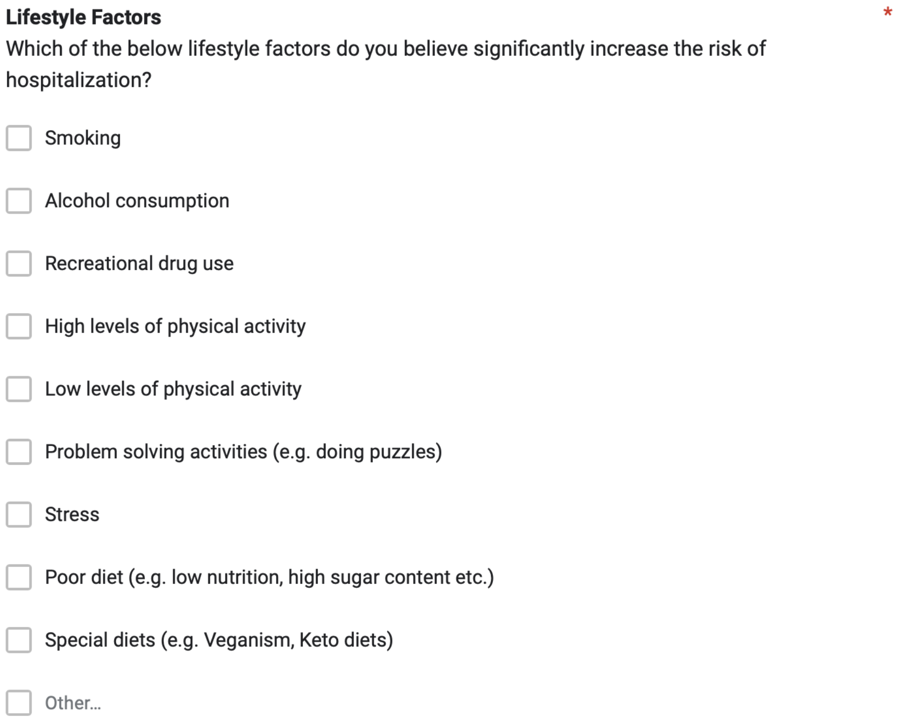
□ Higher social economic status

□ Race (e.g. other than White British patients)

□ Age (e.g. 60+)

□ Other…

Fig 3. Question 2 of survey "Which of the below demographic features do you believe significantly increase the risk of hospitalisation?"



Lifestyle Factors *

Which of the below lifestyle factors do you believe significantly increase the risk of hospitalization?

□ Smoking

□ Alcohol consumption

□ Recreational drug use

□ High levels of physical activity

□ Low levels of physical activity

□ Problem solving activities (e.g. doing puzzles)

□ Stress

□ Poor diet (e.g. low nutrition, high sugar content etc.)

□ Special diets (e.g. Veganism, Keto diets)

□ Other…

Fig 4. Question 3 of survey "Which of the below lifestyle factors do you believe significantly increase the risk of hospitalisation?"

5.2 Synthetic Data Generation

This section will briefly discuss the format of the synthetic dataset, alongside the data generation process.

5.2.1 Data Description

The synthetic dataset contains demographic and lifestyle data for a varying number of synthetic patients. The number of rows has been changed frequently for testing purposes, with values ranging from 5000 to 25000, although the number of feature columns have remained consistent at 8 features. The 8 features are: Age, Gender, Health Condition, Mental Health, Alcohol Consumption, Smoking, Exercise and Socio-Economic Status.

Many of the features are a binary true or false, such as the presence of smoking, or the presence of a mental health condition. Others have more nuance, such as levels of exercise activity, and different numeric values to represent different health conditions: 1 represents respiratory disease, 2 represents cardiovascular disease, 3 represents neurological disease, 4 represents cancer, and 0 represents the absence of a health issue. In some tests, a binary value has been used for the presence of a health condition, but separating this into distinct categories is a more accurate representation of real world data. Three categories have been used for socio-economic status, representing, low, middle and upper classes in terms of financial status. Age is ordinal, with values ranging from 0 to 100, and a mean value of 41 to mirror the average age in the UK[53].

The labels for the database are simply 0 and 1, corresponding with short and long stays in the hospital. These labels are influenced by the aforementioned features, with the presence of multiple factors correlating with longer hospital stays. Whilst the features influence the outcome, the final label generation is random so that the outcome is not deterministic, which makes the task more challenging and more representative of empirical data.

5.2.2 Data Generation

The data was generated using Python[54], with Google Colaboratory[55] as the development environment. Synthetic data can be generated from scratch, or a reference dataset can be utilised as a base, which is then expanded upon. The latter option opens up Generative Adversarial Networks as an option for data generation, but our data is created without a real dataset to reference, therefore simpler, more traditional techniques were used, namely generating random points across probability distributions.

A normal distribution was used to generate the Age data points, which mimics the real distribution of ages within the population. This was achieved using Numpy's random number generator[56] and specifying a normal distribution, with a mean of 41. To generate the Gender data, Numpy's random module was once again to select a number between 0 and 1, with the outcomes then rounded to the closest integer to create a series of 0 and 1 classes. The number of each gender was counted, but it was essential that this remain a random selection, as gender is not split evenly in the real population. A similar method was utilised for generating the other features, with Numpy's random choice used with the specified potential outcomes and a probability for each, with the resulting choices saved to a list. The probabilities for each outcome were chosen using related literature and researching known probability distributions. For example, approximately 13% of the population in Wales smoke[57], so this number was increased to 15% for the purpose of data generation. In some areas, particularly those that are deprived, this number is higher, therefore it is difficult to get a true representation for the hypothetical population involved in the creation of this dataset.

Once each list of values was created, these were used as DataFrame column values. This can be achieved by creating a dictionary of values and then using this when defining the DataFrame. Once checked, the DataFrame was saved to pickle[58] so that the values did not randomise each time the script was run, improving consistency.

The saved DataFrame was then re-loaded, and labels were generated. This was done by creating two functions to calculate the odds of each patient requiring a longer stay. The patient's odds increase linearly with age, at a rate of 0.3 per year. The other features increase odds by varying amounts, such as smoking increasing odds by 10%, and the presence of a mental health condition causing an increase of 15%. Two approaches were tested in regards to the odds related to health conditions, with a blanket increase of 20% if there is a health condition present, or varying increases depending on the condition, such as cancer increasing by 25%, and neurological disease increasing by 15%. The values for these were gathered using the survey distributed to health professionals.

The values for odds increases were easily changed, meaning different outcomes could be tested without influencing the other columns in the dataset. The odds calculations were then used as a threshold for each patient, with random number generation used to ensure the labels were not too deterministic. The odds for each patient were normalised by dividing by the highest odds value, resulting in each value falling between 0 and 1. Then, a random number between 0 and 1 was generated, and if the number was below the specified odds threshold for that patient, a label of 1 was assigned. This means that a higher odds value leads to an increased probability of a longer stay label being assigned.

5.3 Machine Learning

This section will discuss the training and testing process that we utilise when carrying out machine learning experiments, and the chosen models will be summarised.

5.3.1 Training and Testing

Training and testing are vital parts of developing a machine learning model, and data must be split so that both training and testing data are available. When we develop machine learning models, it is necessary to test their performance on data that is unseen. This is because the model will utilise a large amount of data in its training process, which is when the model will learn relationships within the data that influence the data's class or label. The model will become accustomed to the training data as it will iterate over it many times, which can lead to 'overfitting', which is when a model becomes so accustomed to the training data, it cannot generalise well to other data, making the model difficult to use for any other tasks or with any other data. By testing the model using unseen data, we can evaluate its performance on data that has not been used for training, which will give a better indication of how the model will perform for new tasks[59,60,61].

Training and testing data is obtained by splitting one dataset, with common splits for the testing subset being 33% and 20%. The training data is the largest, as large quantities of data are required to train a model effectively. The testing data is simply used as an indicator of the model's performance, and is not used as part of the training process. When splitting the data, many developers prefer to have random data points assigned to each subset rather than defining a set threshold, as this may cause the training dataset to consist solely or mainly of one specific class, which will not be an accurate representation of the model's performance. We can use shuffling to achieve this randomisation, or sklearn's train_test_split module can be used, which will select rows/data points at random. A random seed can be specified to ensure the selections are kept consistent across runs of the script.

Validation data can also be created as a subset of the training data, which will be used as testing data during iterations of training. This is commonly used in neural network training, as the model's performance can be tested after each epoch using the validation data[62]. This gives an indication of the model's learning progress, and is used to indicate moments of overfitting, as the validation performance will either stagnate or reduce. It is also possible to allocate a ratio of validation data rather than pre-separating the training data, but this will allocate the final data points in a batch, which is not random and is not always the best representation of performance.

5.3.2 Chosen Models

This section will summarise the chosen models, including how they work, their strengths and limitations. The results for each of these will be discussed in the following chapter, Results.

5.3.2.1 Supervised Models

Supervised machine learning models are trained using ground-truth labels, with the name supervised implying a level of oversight by consistently checking the model's performance using the labels[63]. This method is very common, and is well suited to classification tasks. The following models are all examples of supervised learning.

Random Forest

Random Forest is a popular and simple model, commonly used for classification and regression. The model consists of multiple Decision Trees, making it an ensemble classifier[64,65]. Each Tree will make a series of decisions regarding the characteristics of the data's features, leading to the predicted label. In the case of classification, the model chooses the majority vote to select the overall predicted label. Increasing the number of trees can improve accuracy. Random Forests use 'bagging' or bootstrap aggregating to select features at random at each candidate split, which can reduce variance in the model.

This model is simple, and does not require a large number of parameters to work effectively. It is also versatile in that it can be used for both classification and regression. The model's primary disadvantage is that it does not describe the relationships between data, only predicts, and an increase in the number of trees can vastly slow down its performance.

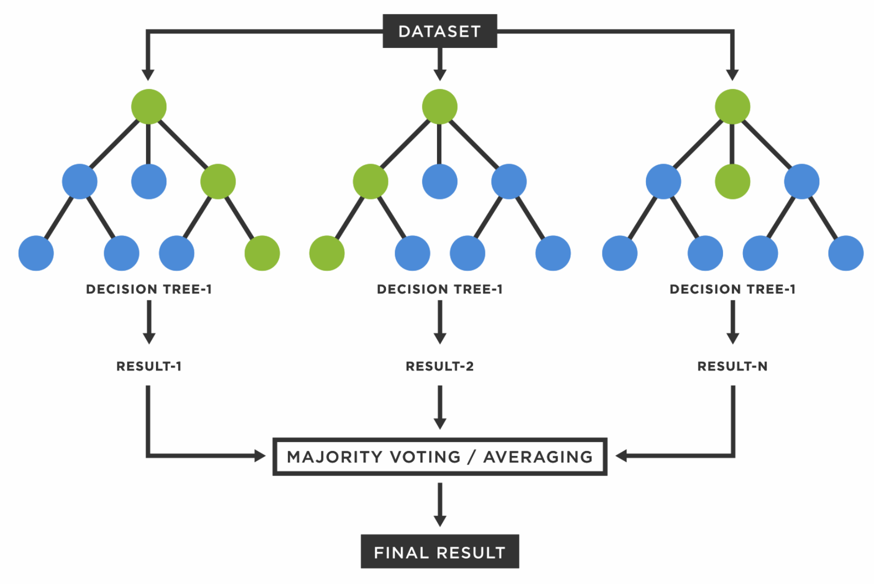


Fig 5. The Random Forest classification process by TIBCO[66]

Support Vector Machine

Support Vector Machines, also known as Support Vector Classifiers, are also used for classification and regression problems. It is a linear classifier, as the model categorises data points as belonging to one of two categories, and a hyperplane is used to determine the class of new data points. This hyperplane is a crucial part of the model's behaviour, as the underlying goal is to find the best margin of separation. The maximum margin can be used to find the maximum distance between two classes, with data points represented in space, and a large margin is considered a good outcome. The hyperplane/margin is used as a decision boundary, therefore a large decision boundary ensures that there are significant differences between features, as data points close together are ambiguous in their classification. For non-linear classification tasks, a kernel is used to alter the data so that it can be used with the SVM model, as real world data is often non-linear. SVM is not probabilistic in nature, but statistical, and performs best with small, complex datasets[67, 68, 69].

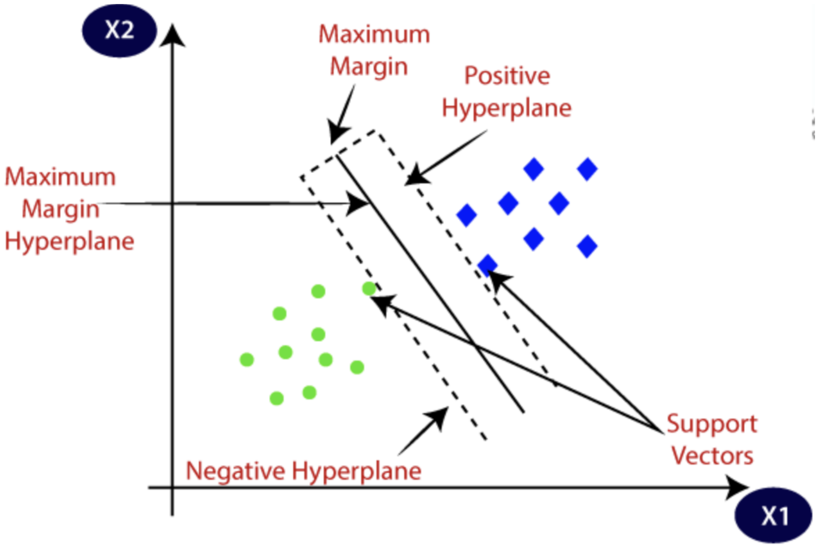


Fig 6. SVM Classifier, with the margin of separation between two classes [70]

K-Nearest Neighbour

K-nearest Neighbour is a supervised clustering method that assumes similar data points will be close to each other in proximity. As with other clustering techniques, the distance between the query data point and other example data points is calculated, with the nearest neighbours to the point identified. Then, we examine the classes of the nearest neighbours, and find the mode class. This is then chosen as the class for the query data point, or in the case of regression, the mean of the data points are used. K refers to the number of neighbours that will be used for voting. For classification tasks, K should normally be an odd number, to ensure that there is a tie-breaking vote. Choosing a low value for K would lead to unstable results, as only one close neighbour will be chosen which does not provide a comprehensive results in cases such as 5 points of red classification and 1 point of green classification, but the green is the closest point, as only the closest will be taken into account regardless of its minority status. This is a simple classifier that is versatile, but its performance can be slow depending on complexity and dataset size[71,72]. Our classifier employs the use of 9 neighbours as the value of K, which yields good results. These results are also consistently good at a lower K value of 3, but lower numbers can generalise poorly.

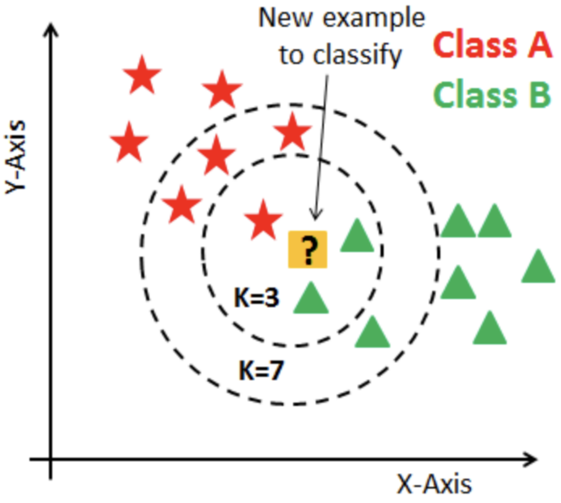


Fig 7. KNN with number of neighbours parameter set to 3 [73]

Neural Network

Neural Networks are the most complex of models, with this type of classifier moving into the field of Deep Learning. They are modelled after the human brain, with the first iterations of these models called Artificial Neural Networks. The model consists of many layers, each of which has interconnected neurons. Each model has an input layer and output layer, the latter of which will consist of one node in the case of binary classification, or multiple nodes to match the number of potential classes if more than two outcomes are possible. There are also hidden layers, each of which processes the data and passes the output on to the next layer. As the number of hidden layers increases, so too does the complexity and processing power of the model. Weights are assigned to the neurons, which dictates the influence of each neuron on the others, with a positive value increasing the power, and a negative value suppressing power. Neural Networks are computationally intensive and can take a long time to train, with the number of loops over the data called 'epochs' which determines the training time of the model. There are also many other parameters that affect training time and performance, including learning rate, activation functions, L2 regularizers, kernel size in the case of convolution layers and layers such as pooling and dropouts. There are different types of networks that can be employed depending on the task, such as Long-Short Term Memory, 1D Convolutional Neural Network, 2D CNN, Recurrent Neural Network and Dense Neural Network. The chosen type for this task is the Dense Network. Neural Networks perform well on a variety of complex tasks, such as text processing and image classification[74,74].

The complexity of a model is dictated by the number of hidden layers and the number of units per layer, with a higher number of one or both leading to a powerful model that may be prone to overfitting if the task is too simple. On the other hand, a less powerful model may not be sufficient for solving a challenging task. There are also methods of reducing overfitting, such as using dropout layers which will randomly set input units to 0. Reducing the learning rate can also prevent overfitting, as the model will learn slowly and will not converge as quickly, meaning the model will be less sensitive to noise[76].

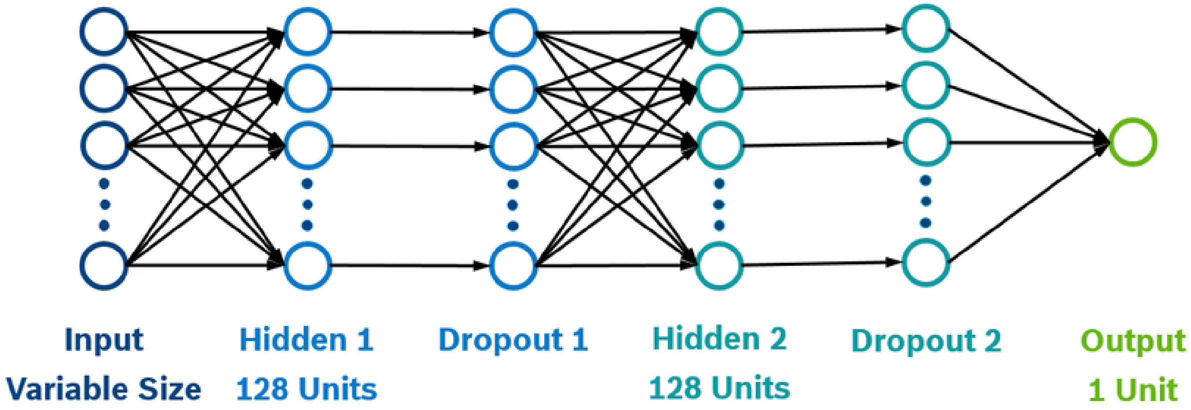


Fig 8. An example of a neural network with Dense layers and dropout, with a single output [77].

Sometimes, complexity is not required, as the model may not see a performance increase. Two versions of the neural network will be tested, one being architecturally similar to the model pictured above, with multiple hidden layers and a high number of neurons, while another version will only include one hidden layer and will utilise a small number of neurons, as depicted in the diagram below.

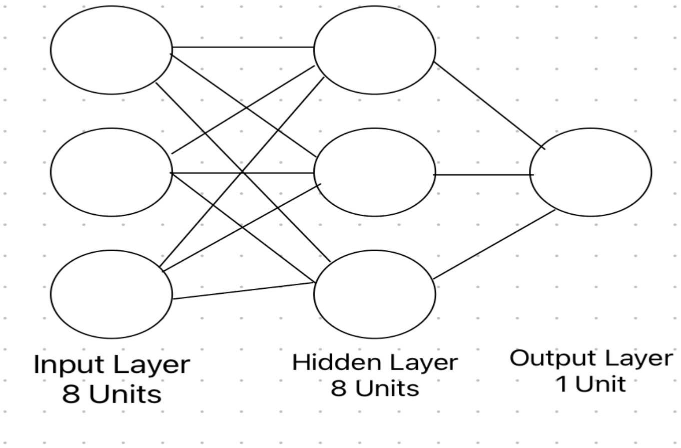


Fig 9. Diagram of a simple dense neural network with only one hidden layer, representing our network.

5.3.2.2 Unsupervised

Unsupervised learning is done without ground-truth labels, therefore the model establishes patterns and relationships within the data through its own pattern recognition abilities. The data is trained using only the X values, and can be useful for identifying unknown relationships in data without human intervention[78]. Only one unsupervised model has been chosen for this project, as the labels are known and it is expected that supervised models such as Neural Networks and Random Forest will perform better.

K-Means Clustering

K-means Clustering is one of the most popular clustering techniques, and aims to partition N observations into K clusters. The number of clusters (k) is chosen by the developer during initialization, and the model will then initialise a corresponding number of centroids which will serve as initial central points of each cluster. Each data point's distance to the centroid is calculated, and it is assigned to the cluster of the closest centroid. Once each data point has been assigned, the centroid's position will be updated by calculating the mean of the data points within the cluster. This method has an adjustable number of iterations, meaning this process will repeat a set amount of times until the clusters are finalised. This model excels at finding relationships within the data, and is reasonably simple to implement as it can be adapted and used for different tasks with ease. One downside is the model's dependance on finding the ideal number of iterations and clusters, and it may not perform well for every task, such as cases where features are not dependent on each other[79,80].

The following chapter will summarise the results from the survey, data generation and the machine learning experimentation.

6. Results

This chapter will summarise the results from the survey responses, the data generation process, and the machine learning experimentation. The results will be discussed briefly, with greater detail and conclusions provided in the following chapter.

6.1 Survey Results

This section will discuss the data generated by survey responses, with the insights and findings provided.

The survey received 24 responses from professionals working in health, which is a small sample size but provides sufficient data for analysis. Each response was gathered from a reputable source, as the survey was only distributed to those verified as employees of the NHS. The majority of responses were contained to the provided options, but some respondents included additional responses.

The first question enquired into health issues most commonly associated with hospitalisation. The responses are summarised below.

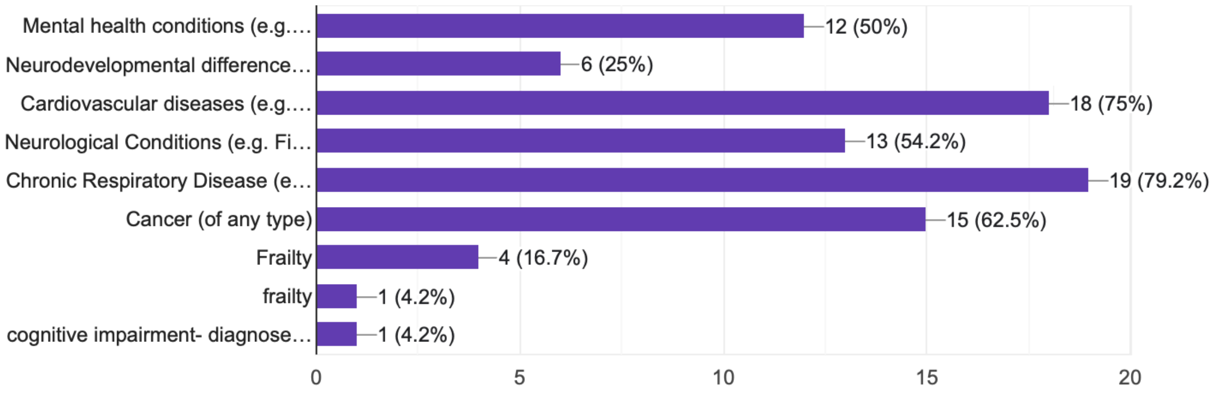


Fig 10. Breakdown of responses for Q1.

Respiratory Diseases and *Cardiovascular Diseases* were most commonly chosen, both of which are common health conditions, which may contribute to their prominence in ED observations. *Neurodevelopmental Differences* was not chosen by many, coinciding with recent schools of thinking that diagnoses such as Autism should not be considered a health condition, but a difference in neurotype, although it is often comorbid with other health issues, such as IBS and POTS[81]. *Cancer, Neurological Conditions and Mental Illness* were all frequently chosen, therefore these have been included in the dataset alongside *Respiratory* and *Cardiovascular Disease.* Respondents included their own suggestions, with 5 people including *Frailty* as a factor, and 1 suggesting *Cognitive Impairment*, although this could include a myriad of conditions. Frailty typically occurs with age but can be present at any point in a lifetime, though it has been excluded from the dataset due to the prominence of other factors.

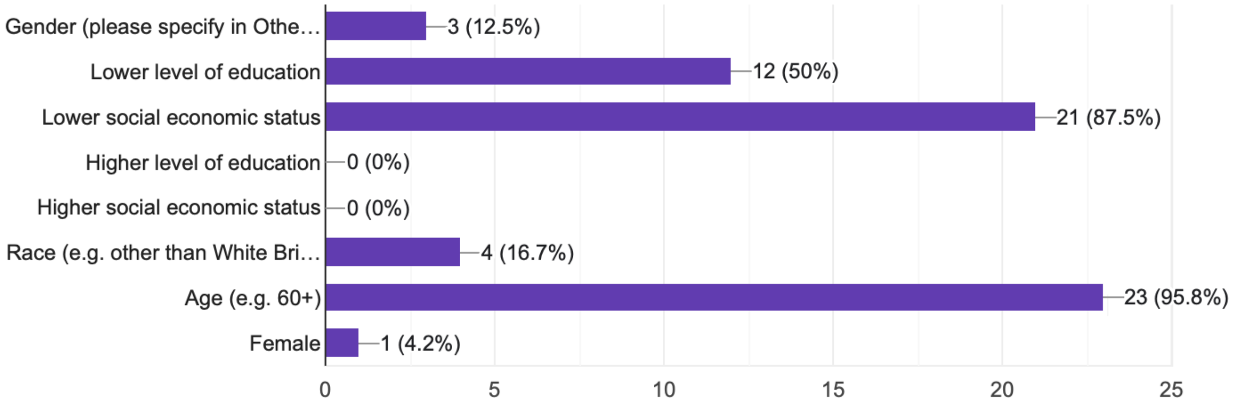


Fig 11. Breakdown of responses for Q2.

Demographic factors were investigated next, with the results captured above. *Age* and *Lower Socioeconomic Status* were chosen by the majority of respondents. *Age* is correlated with the emergence of health issues, and is arguably one of the most important factors when considering risk of hospitalisation. *Socioeconomic status* is more complex, and is associated with societal issues and lack of access to suitable care, but this is an interesting aspect of a patient's risk factor. *Gender* and *Race* were less commonly chosen, but *gender* is included within the synthetic dataset due to its simplicity to capture and the difference in potential health issues, such as reproductive disorders. *Lower levels of education* were excluded from the synthetic dataset despite being chosen by 12 respondents, as this would be difficult to capture and could lead to ethical issues. The two options that were not selected by respondents, higher level of education/socioeconomic status, were also excluded from the data generation.

The final question considers lifestyle factors that increase the risk of hospitalisation, such as smoking or alcohol consumption.

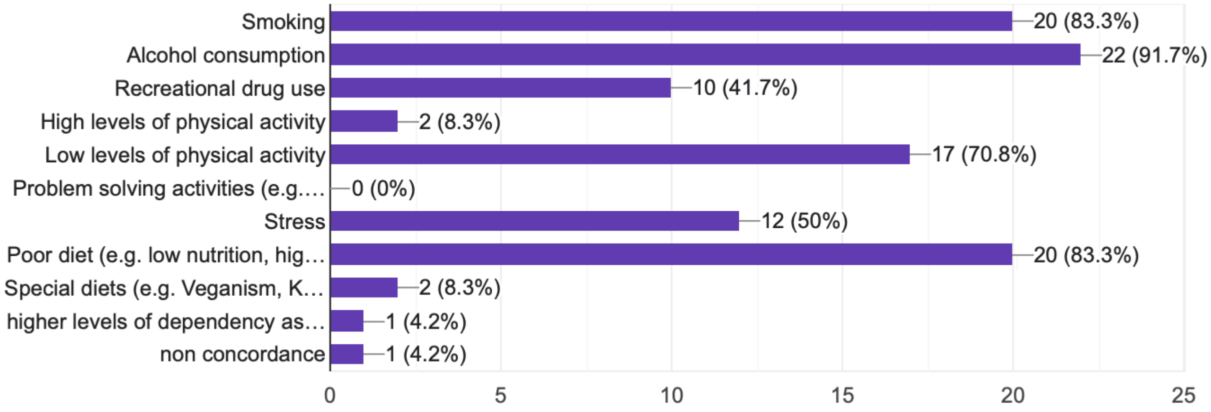


Fig 12. Breakdown of responses for Q3.

There were a greater number of options for this question, although *Smoking, Alcohol Consumption and Poor Diet* were chosen most frequently. Both *Low and High Levels of Activity* were selected, although low levels of activity is considered the more influencing factor with 17 responses versus 2. For the synthetic dataset, levels of activity have been included on a scale of low to high, but stress and poor diet have been excluded. This question saw a wide spread of answers, with all options other than *Problem Solving Activities* being selected by at least one person. *Non-concordance* and *dependency on others* due to frailty were suggested by respondents.

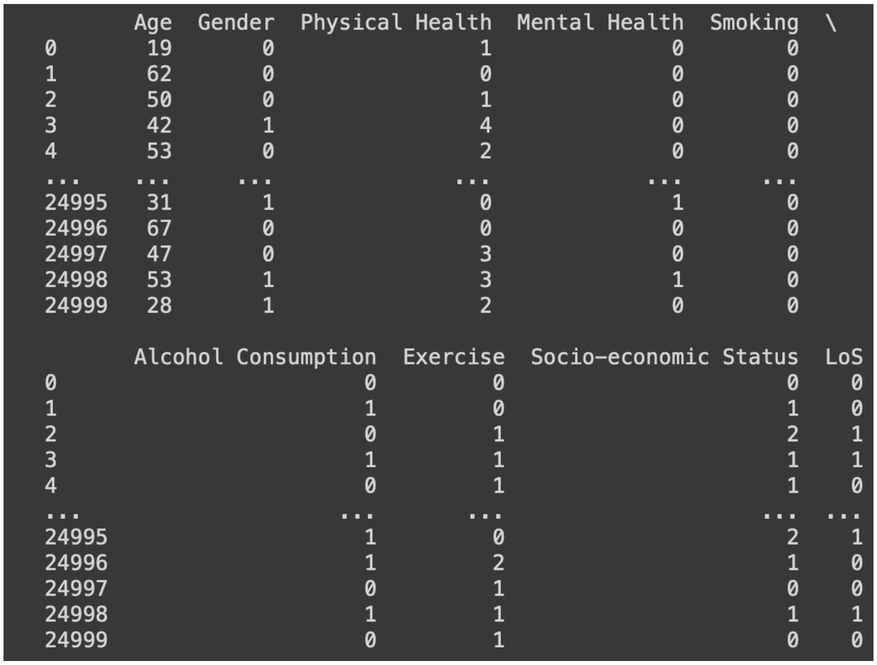
6.1.2 Survey Conclusions

The results from the survey demonstrate the wide range of influencing factors that contribute to a patient's risk level, with many more likely not considered. Health conditions are commonly thought of when we discuss risk of hospitalisation, and these are closely tied to both demographic and lifestyle factors, with age and socioeconomic status being examples of strong correlating factors with the development of health conditions[82]. This survey did not consider rare diseases, and does not discuss the relationship between the three different factors. The range and complexity involved in the prediction of who will require a longer hospital stay strengthens the need for machine learning in creating efficient processes. The factors chosen by respondents have been included within the synthetic data set in a simplistic form, though some features such as Stress and Educational Level have been excluded.

6.2 Data Generation Results

This section will discuss the features and distribution of the generated data, including a breakdown of ages, gender and the resulting odds for each patient.

The DataFrame consists of 8 feature columns and a single label column, and a maximum of 25000 rows. The dataset is pictured below.



```
       Age  Gender  Physical Health  Mental Health  Smoking  \
0       19       0                1              0        0   
1       62       0                0              0        0   
2       50       0                1              0        0   
3       42       1                4              0        0   
4       53       0                2              0        0   
...    ...     ...              ...            ...      ...   
24995   31       1                0              1        0   
24996   67       0                0              0        0   
24997   47       0                3              0        0   
24998   53       1                3              1        0   
24999   28       1                2              0        0   

       Alcohol Consumption  Exercise  Socio-economic Status  LoS  
0                        0         0                      0    0  
1                        1         0                      1    0  
2                        0         1                      2    1  
3                        1         1                      1    1  
4                        0         1                      1    0  
...                    ...       ...                    ...  ...  
24995                    1         0                      2    1  
24996                    1         2                      1    0  
24997                    0         1                      0    0  
24998                    1         1                      1    1  
24999                    0         1                      0    0  
```

Fig 13. Summary of the generated dataset, presented using Pandas.

Each row was generated individually using probabilities, and some features are balanced whilst others are imbalanced, reflecting real data. Age follows a normal distribution with a peak at 41 years old, and the resulting distribution is presented as a histogram below.

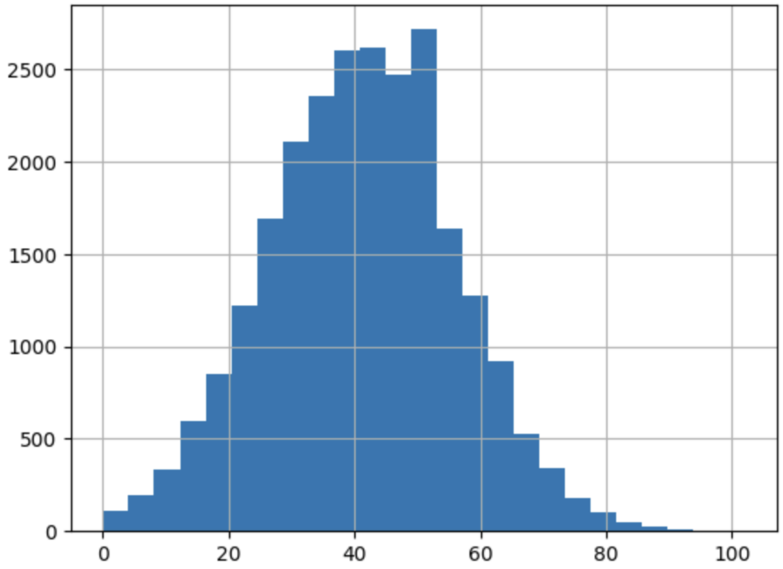


Fig 14. Distribution of Age, plotted as a histogram.

The 'tail' of the distribution is longer as there are a greater number of people alive at younger ages, although if this dataset was only concerned with ED admissions rather than a general summary of the population, it is likely that there would be a higher percentage of older individuals, with a higher peak.

Gender is considered to be approximately equal in its distribution, but this is random and may fluctuate. Random choice was used to generate these values, but the probability of each occurring is equal. The breakdown of each category (male, female) proves the nearly equal distribution.

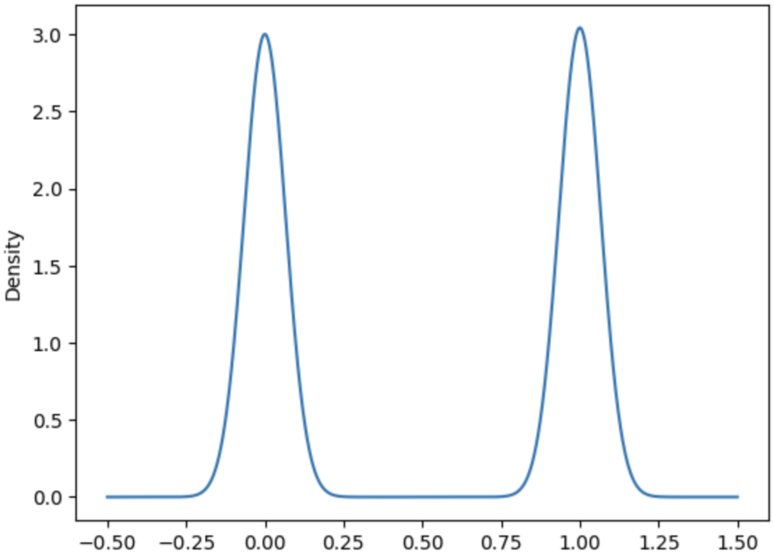


Fig 15. Distribution of Gender.

Smoking is less common within the general population, therefore this column is imbalanced, with the majority of values being 0, or non-smoking. Alcohol consumption is significantly more common, with over 60% of individuals reporting recent alcohol consumption[83], therefore this column has a more equal distribution, with a slight leaning towards a value of 1. Both are pictured below.

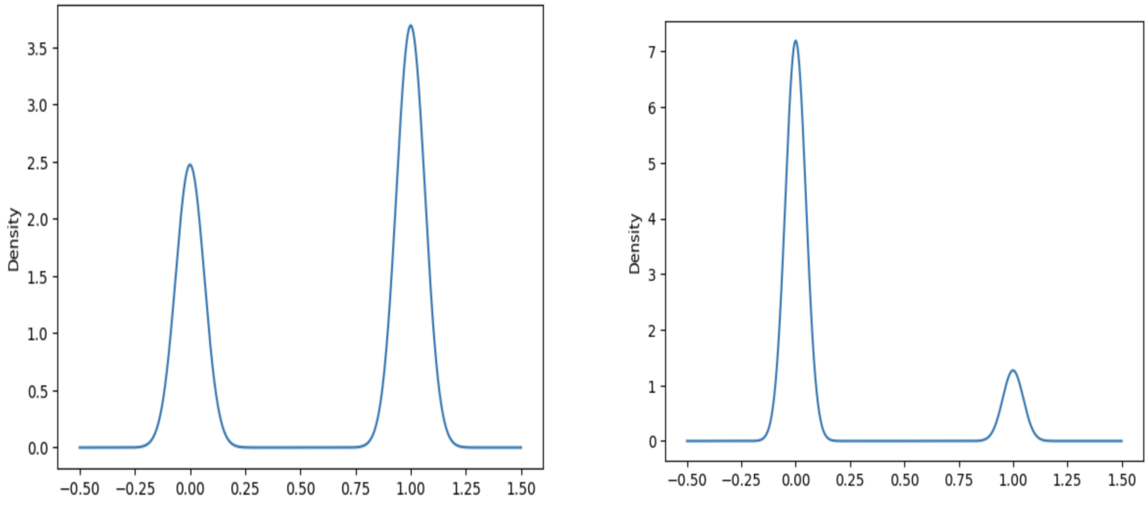


Fig 16. Distribution of Alcohol Consumption (left) and Smoking (right).

Mental health issues and physical health conditions that are significant enough to contribute to hospitalisation risk are not typically found within the general population, with a greater number of 'healthy' individuals. The values for physical health conditions were broken down into 5 categories, including respiratory, cardiovascular and neurological diseases, cancer, and the lack of a health condition. Mental health has a binary value of true or false. In both cases, a value of 0, or false, is most likely, although physical health has a greater number of positive cases due to the range of potential conditions.

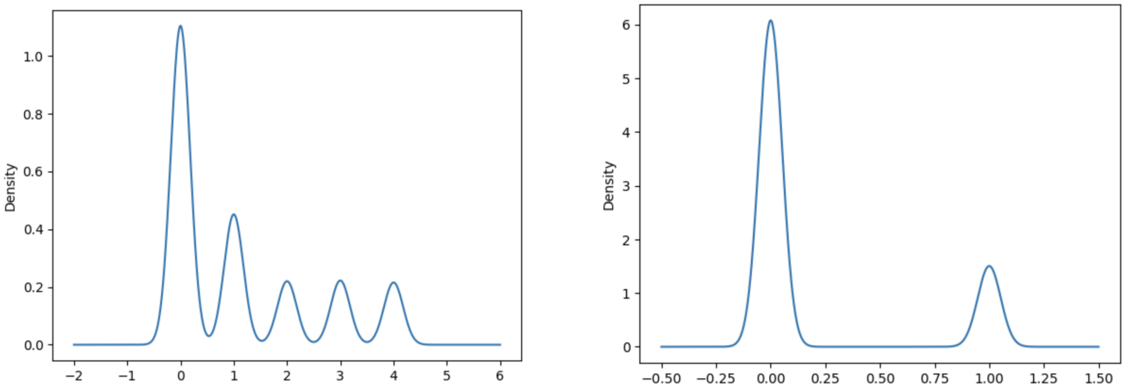


Fig 17. Distribution of physical health conditions (left) and presence of a mental health condition (right).

Socio-economic Status and Exercise are split into three categories, with 0 having the least impact on the patient's odds of hospitalisation, and 2 the most. 1 is the most common value, which is also the average level of impact, and represents most people in the population.

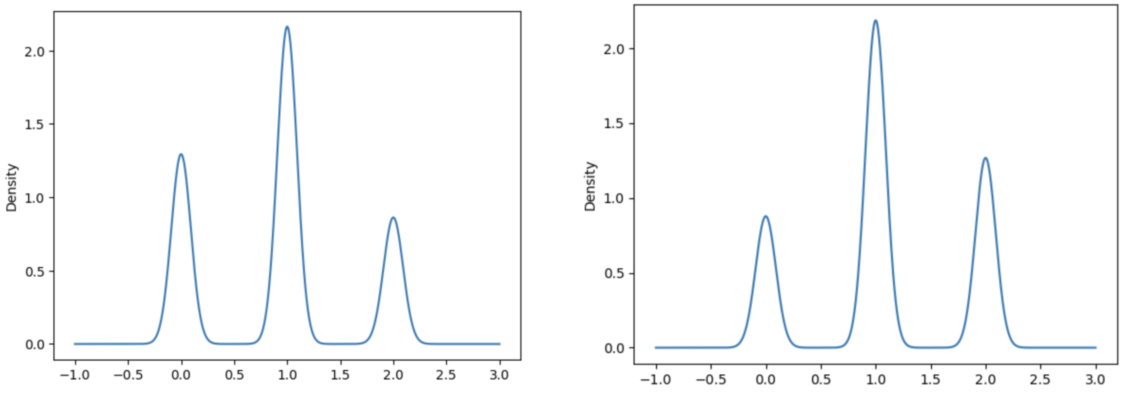


Fig 18. Distribution of SES (left) and Exercise (right).

The generated labels have potential values of 0 or 1, with the number of 1 labels influenced by the number of influencing features. As demonstrated above, many of the factors that increase risk have a low probability and are less common than values that have no impact on the patient's odds. As with a real dataset, the number of short stays is greater than the number of long stays, although this is not significantly imbalanced. The number of 1 labels could be increased by increasing the probability of values of 1 or 2 being chosen for each column, although this may become less representative of real world data, as probabilities have been chosen to reflect current statistics. The breakdown of short and long stays is pictured below.

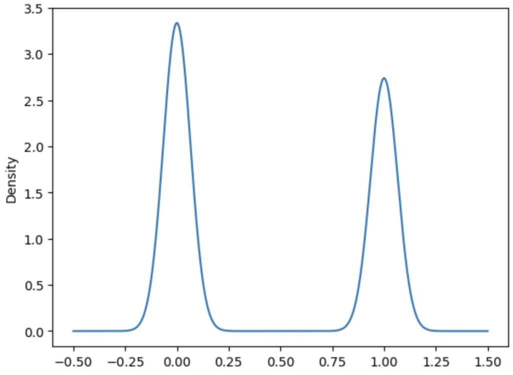


Fig 19. Distribution of 0 and 1 labels.

The following section will discuss the results of the machine learning experimentation carried out on this synthetic dataset.

6.3 Machine Learning Results

This section will discuss the findings from the machine learning experimentation, with some smaller investigations including the impact of dataset size and noise.

The results from the machine learning experimentation are largely positive, and indicate that machine learning can be used to predict length of stay, which could ultimately allow for better resource allocation and intervention. When analysing the results of the machine learning, raw accuracy was used, alongside the F1 score and confusion matrices. It is important to utilise multiple techniques, as raw accuracy alone does not provide a breakdown of where performance is better or worse, and does not include precision or recall. Precision denotes the ratio between true positives and total positives, which is calculated by dividing the number of true positive predictions by the total number of positive predictions, whereas recall is the true positive rate, which is the number of true positives divided by true positives and false negatives. It can be difficult for the model to perform well using both metrics, which is why the F1 score is used as a balancing metric, and a more reliable indicator of the model's performance in identifying positive samples. Confusion matrices provide a breakdown of predictions versus true labels for each class, which is useful in identifying where the model is performing badly, and can provide insight into why this may be the case[84,85]. We also consider the performance in relation to the chance level, which is the probability of randomly selecting the correct class, which in this case would be 50%. A model's performance should be higher than chance level[86].

The five tested models are Random Forest (RF), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Neural Networks (NN) and K-Means (KM), the latter being the only unsupervised model. The F1 score of each model is included in the table below.

Model	F1 Score
Random Forest	0.81
SVM	0.61
KNN	0.74
K-means	0.46
Neural Network	0.63

Fig 20. Table of F1 scores for each model.

6.3.1 Random Forest

These results demonstrate that RF is the superior classifier in terms of F1, with a strong performance which is also evidenced by the associated confusion matrix, which is pictured below.

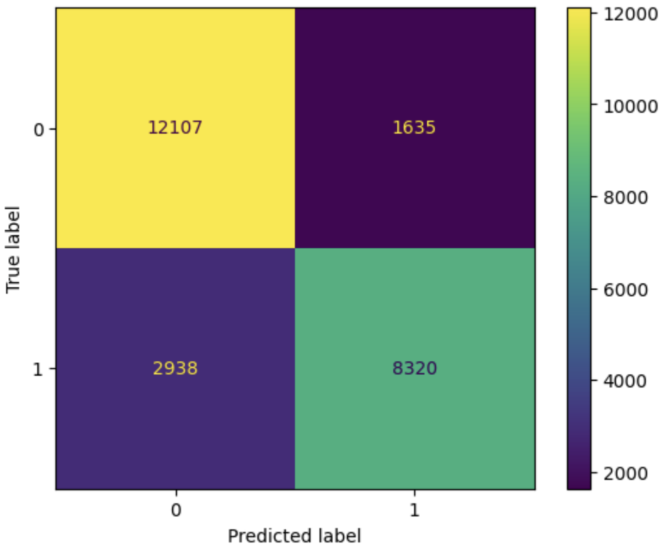


Fig 21. Confusion matrix for random forest.

The two ground truth labels are 0 and 1, and it is clear that the model does well with correctly classifying a label of 0. However, we are interested in deducing the model's capability to correctly classify longer hospital stays. The dataset is slightly imbalanced, with a larger number of short stays, but we can see from the confusion matrix that approximately 74% of longer stays are correctly classified. This is a good score, particularly as there is an element of randomness in the label generation, and this score sets the model nearly 25% above chance level. The highest number of incorrect predictions come from the misclassification of 1 as 0, with only approximately 12% of incorrect classifications of 0 as 1. This indicates that the model leans towards predictions of 0, which is not too concerning given the imbalanced ratio of 0:1. These results suggest that the RF classifier is a viable solution, with a low runtime and a respectable performance on the synthetic data. This is likely due to decision trees performing well with boolean data, such as the presence of smoking or not, which allows the model to classify with a higher degree of accuracy. The performance drop comes from the anomalous data created by random number generation, as their features will indicate a longer or shorter stay, but they may be given a different label.

6.3.2 KNN

The KNN classifier also performs well, with a F1 score of 0.74. The accuracy of label 1 predictions is 68%, placing it below the RF, but still above chance level, showing that the model has learned relationships within the data to a reasonable extent. There are also more incorrect classification of 0 as 1, with a 20% misclassification rate. This indicates that the model is not biased in its classification, and instead loses accuracy due to the random label generation placing data points of differing labels in close proximity in terms of features.

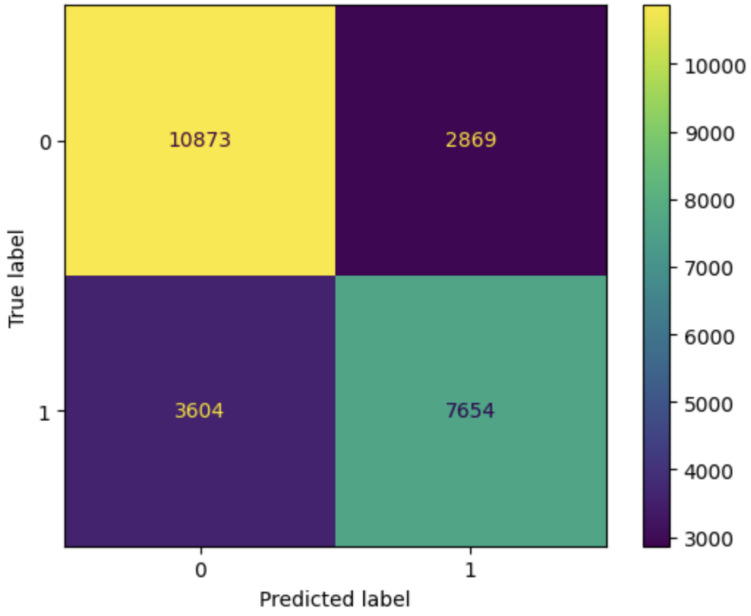


Fig 22. Confusion matrix for KNN.

The KNN classifier would likely be improved by reducing the randomness of the label generation, as clustering methods typically rely on correlation between features and the proximity of data points to one another in a feature space, which is difficult given that some of the data points will have a combination of features that indicate a label of 1, but due to the random number generation, will have been given a label of 0. KNN is a potential model for future work due to its good performance, with the expectation that it will improve with real data.

6.3.3 SVM

The SVM did not perform as well as expected, given that the presence of only two labels allows for a linear classification, and achieved a F1 score of 0.61. This is above chance level and is a moderate performance, but would likely not be chosen as a competitive model based on the performance using the synthetic data.

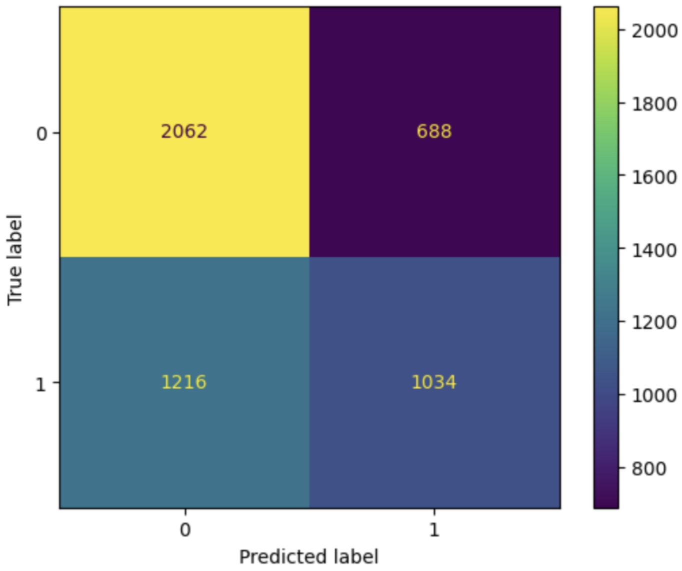


Fig 23. Confusion matrix for SVM

We can see from the confusion matrix for SVM that the classifier does not do well with classifying the label of 1, with a below chance level performance. The classifier's score comes from the model's ability to classify 0, but it leans too far towards this label in both cases. This indicates that this model is not suitable for the task, as the primary goal of the project is to predict when a patient will stay for longer, which it does not achieve. It is possible the model may be able to identify a more suitable separation margin given real data.

6.3.4 Neural Network

The Neural Network did slightly better than the SVM, but did not achieve a very high score, with a F1 score of only 0.63. The model's learning rate and dropout layers were tweaked to find the optimal performance, but this stagnated in the 60-70% accuracy range. From the accuracy and loss curves pictured below, we can see that training and validation data do not significantly differ in performance, but learning becomes less efficient around epoch 20. Both the simple and complex NN performed very similarly with scores of 0.63 each, therefore complexity is not the issue.

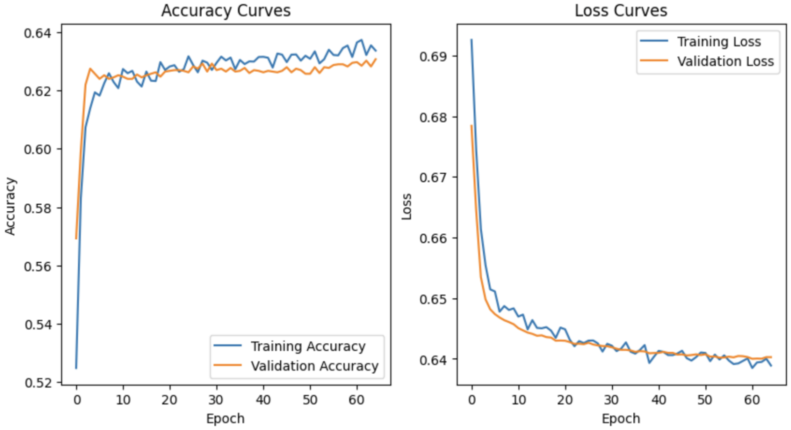


Fig 24. Accuracy and loss curves for neural network.

This performance implies an issue with the synthetic data itself, or the design of the network, as this model is typically flexible and capable of learning complex data effectively. Given that the classification task only involves two labels, it was expected that this performance would be higher, but with more time this model could be adapted to a suitable structure. This model is still relevant for the project going forward, but it will need more work.

6.3.5 Model Stability

The performance of the models across different dataset sizes was measured, with sample sizes of 5000, 15000 and the full 25000 chosen. Each model was rerun on the smaller dataset, with its F1 measured. These results have been plotted on the graph pictured below.

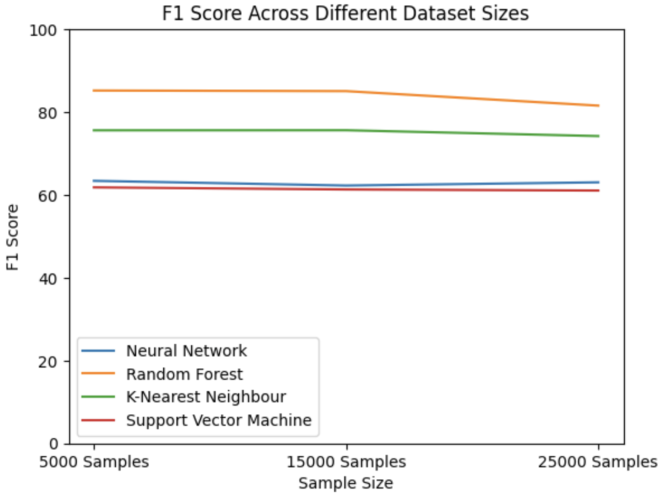


Fig 25. F1 scores across dataset sizes.

From this, we can see that the models' performances have remained consistent across the dataset sizes, indicating that the models are stable. There is a slight dip in performance, especially with the overall best performing model (RF) but this is expected due to the presence of increased noise in larger subsets. The SVM yields similar accuracies across each test, but due to its weak performance, it is reasonable to assume it was influenced by noise in the original dataset. Smaller datasets can often lead to overfitting due to limited data, meaning the model will learn the training data well, but it will not generalise to unseen data as the model's experience with different data is limited.

6.3.6 Performance with Noise

Noise is present in the majority of datasets, as data is not clean and will feature irrelevant or incorrect features. While usually we would want a model to perform well regardless of noise, including significant noise can test whether the model's performance drops to chance level, which would be 50% accuracy in this case. The dataset's labels and the odds of each patient were generated with significant noise and randomness and this was saved as a different dataset. When tested, each model performed worse, as pictured below.

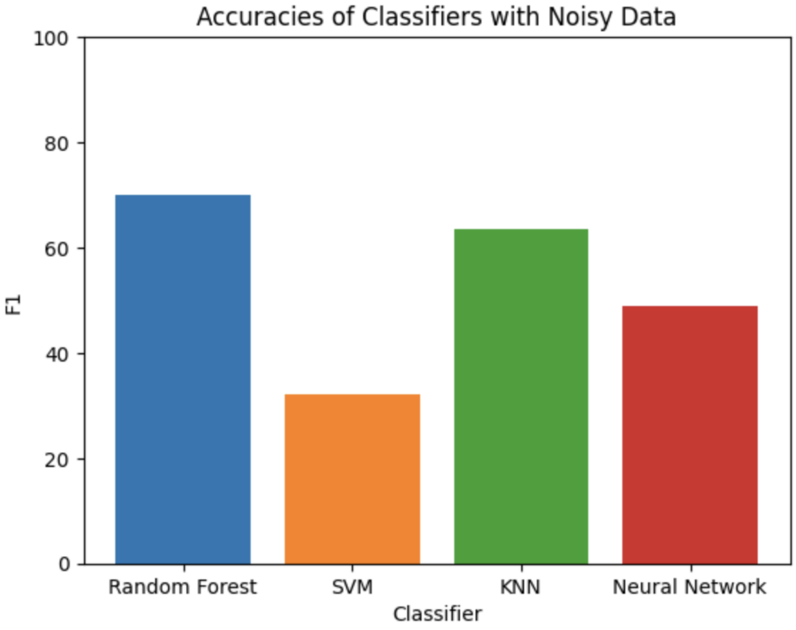


Fig 26. Graph showing performance of classifiers on noisy data.

We can see that the SVM drops below chance level, which is not unexpected given its poor performance on the standard data. The other models sit around or above chance level, but each drops in performance. The drops below chance level are due to the model fitting to the noise and predicting incorrect labels more frequently, which demonstrates the importance of formatting data and ensuring the model does not overfit to the training data and make incorrect assumptions.

Model	**F1**
Random Forest	0.70
KNN	0.63
Neural Network	0.48
SVM	0.32

Fig 27. Table of F1 scores from noisy data.

This method also included testing with minor noise, which did not see a significant drop in performance, indicating that the models should generalise well and should not be overly impacted by moderate noise.

This chapter has discussed the project's results, with this section detailing the F1 results and experimental findings from testing the machine learning models. The following chapter will discuss the impact of these findings in relation to the wider project, and the required next steps to transition into creating an advanced model capable of profiling those visiting the ED.

7. Discussion

This chapter will discuss the meaning of the results and their impact on the wider project, with a consideration for future work and next steps.

7.1 Survey Implications

The survey results demonstrate that there are many possible influencing factors when considering a patient's risk of hospitalisation, with many aspects of lifestyle, demographic and health chosen as prominent observations in a healthcare setting. The emergence of this pattern indicates that there is a pattern to the flow of patients visiting the ED, which has implications for the development of the wider project. By identifying key features and trends within the data and presentations of features, we can begin to identify suitable models and a potential algorithm for predicting the ED attendance of specific patients.

The presence of lifestyle factors such as smoking and alcohol consumption indicate that regions within Wales are at higher risk of hospitalisation, with ties in with demographic features such as lower socioeconomic status being an influencing feature. There are counties in Wales with higher rates of poverty, which often see higher rates of alcohol consumption, recreational drug use and smoking[87, 88], indicating that societal inequalities and issues such as these are a risk factor, which is important to consider when attempting to identify patients requiring interventions. Other factors such as age and frailty are often thought of when considering health risks, but the rate of responses demonstrates that these are persistently accurate indicators of impending hospitalisation, therefore the age, medication use and rate of check-ups could be a vital indication of the likelihood of a patient arriving at the ED[89].

7.2 Model Performance

The results from testing various machine learning models are positive and demonstrate the potential of this approach in improving ED efficiency by predicting the profiles of incoming patients. This is a complicated task due to the large variation in potential features, including many that are unknown. Many features may be related, such as socio-economic status and lifestyle factors, or health issues caused by lifestyle, and factors such as number of medications taken and frequency of medication reviews can impact risk level. This element of the project has utilised synthetic data as proof of concept, which is a simpler task due to the limited features and controlled environment, therefore results will likely change with future work. Currently, the Random Forest classifier performs best, with a high F1 score and a good accuracy score across both potential classes, much higher than chance level. The nature of the synthetic data lends itself well to the decision thresholds utilised by decision trees, particularly as many of the features are simple true or false. This linear method of working through the included features may not work as well with very complex data, such as that obtained from a real ED. Therefore, it is prudent to consider whether the performance of this model will be reliable, or whether a model with many adjustable parameters, such as a neural network, will perform best with new, unseen data.

The Neural Network did not perform as well as expected, but the randomness of label generation will influence the model's ability to learn, as the combination of features leading to an assumed label of 1 (longer stay) may not lead to the expected outcome. Neural Networks have many parameters, including learning rate, different layers and numbers of neutrons, and optimisers. Related work indicates that these models are very effective at identifying longer hospital stays, therefore the use of the Neural Network should be investigated further. Finding the correct combination for the task can be challenging, and with more time and data that has greater correlation between features, the Neural Network would likely perform very well. Neural Networks are prone to overfitting due to complexity, which is one potential issue with the model's performance, as its complexity allows the model to learn relationships within the training data, which will make the task of classifying unseen test data more difficult. Often, a poor performance will be the result of the model overfitting on noise, leading to incorrect classifications. This can be rectified by reducing learning rate or adapting the hidden layers, although sometimes the data itself is incorrectly formatted or requires preprocessing. As stated, the future ED data will contain many variables, and there will be many unknown relationships and patterns, which is when a Neural Networks works at its best to highlight these patterns. Noise will continue to be an issue, but feature analysis and preprocessing will be used to counter this.

The KNN model performed well, and serves as an example of machine learning's ability to classify complex tasks with minimal workload, as KNN is simple and does not require significant computing power in its basic form, which was used to generate these results. Despite this, it is unlikely that the model will continue to be a good choice for the very large scale and challenging data, as the resource requirement will increase significantly, as its simplified method of classifying data based on proximity will not account for complex relationships and causations between data. K-Means, the second clustering method tested, did not perform well.

This is due to the lack of correlation between features within the synthetic data set, as each column was created independent of others. Clustering methods rely on relationships within the data to work well, with the KNN model being the exception to this due to its voting system, whereas k-means assigns centroids and clusters based on distance to other data points based on features, but similarity amongst some features does not necessarily mean that there are similarities among others. This performance is likely to improve with real data, as there are many hidden relationships within features when the data is not generated in a controlled, restrictive setting.

SVM also did not perform well, although its linear classification process should favour binary classification tasks. Once again, this is likely due to the random number generation involved in the label generation, as previous iterations of the dataset that produced deterministic labels resulted in very high scores. Due to SVM's reliance on identifying a suitable separation margin, it is sensitive to noise and its performance will have been impacted by the presence of random labels. This is likely to continue to be an issue as future data will contain noise.

7.3 Societal and Ethical Concerns

Many people are wary of artificial intelligence and its use in sensitive areas such as healthcare. This is becoming a wider debate as AI becomes ingrained within society and developments within the sector lead to better performance, and applications such as the generation of art and music create issues surrounding fairness and legality. It is important to consider the intended impact of the project, and whether it could lead to any ethical issues or unwanted harm. One potential issue is bias, which can be conscious or unconscious. The latter is most common within health due to increased efforts to improve fairness, but lack of representation and bias within medical research have led to misdiagnoses. Areas within Wales may not represent all ethnic groups or socioeconomic classes, therefore the data from these regions will be biased. This can lead to false predictions and a model that consistently disregards certain groups of people, which could directly harm the affected groups[90].

Other ethical concerns include inaccuracies of the AI, data privacy, and the potential for replacing humans within the industry[91]. To counter this, data should be handled securely and as anonymously as possible. Also, development of human-centred AI ensures that the tool only supports and enhances operations, and does not take away from human creativity or ingenuity. The role of AI in diagnosing and treating patients without human supervision is unclear, and it is evident that many will take issue with the potential for misdiagnosis or breakdown, therefore it is unlikely that this will become a consideration for development currently. This project and its associated future work will only serve as a tool to support resource allocation, and will not attempt to diagnose, treat or manage any conditions.

7.4 Limitations

There are limitations to this work, which will be discussed in this section.

A key limitation of this project is the lack of access to ED data, meaning synthetic data had to be generated. While synthetic data can be useful and leads to lower costs, it is not necessarily an accurate representation of real data. There is a lack of relationship between features, leading to decreased learning performance, and the complexity of the real-world task may not be fully captured. Furthermore, the method of generating labels for each row is a subjective method that may not be an ideal option, with the value for the odds increase for each present factor being decided by the researcher, which does not fully reflect data concerning real people. The random generation of labels using odds as a guide was necessary to ensure the data and learning were not deterministic, as this would not be representative or sufficiently complex, but the randomness resulted in an overall reduction in performance, as many of the labels could be considered noise. Preprocessing of real data would handle noise in a way that would lead to good results that are still generalisable.

As stated, there is no correlation between features, which is not truly representative of real world data. Typically, the likelihood of health conditions will increase with age, and the likelihood of lifestyle factors such as smoking may increase with lower levels of socioeconomic status, but this is not represented within the synthetic dataset. This will be addressed once the NHS ED dataset can be accessed and used for experimentation.

A further limitation is that a small number of models were tested, which could be expanded upon to gather more results. The neural network could also have been improved, as attempts to improve its performance did not succeed, meaning greater work should be done to understand the reasons for its weak performance.

The survey was a success, but a greater number of responses would improve overall findings as more opinions could broaden the scope of considerations. To improve this, the survey could be distributed on a wider scale, or could be open for responses for longer, but this was reduced to ensure there was ample time for data generation and experimentation.

The final limitation is that the script is written in Python, whereas R has powerful data analysis capabilities and is used by Hywel Dda Health Board, therefore this work cannot be seamlessly integrated. As this is a short proof of concept project, this is not a major issue, but R will be used going forwards, to ensure that the work can be used alongside Hywel Dda's own system.

8. Conclusion

This project has explored the potential factors leading to an increased risk of hospitalisation, including health conditions, lifestyle choices and demographic profiles. Synthetic data has been generated using Python, NumPy, Google Colab and Pandas. The breakdown of the synthetic data has been summarised, with factors such as smoking, alcohol, low levels of exercise and age influencing length of stay in hospitals. Machine learning models have been tested in their ability to predict LoS, with a strong performance from the Random Forest, but a weaker than expected performance from the Neural Network. The limitations of this work lie in the synthetic dataset, as it is difficult to fully capture data that is complex as health profiles, and relationships among features such as links between demographic and lifestyle have not been included.

8.1 Future Work

Future work should employ real ED data, which will be truly representative of the real world and will yield interesting findings. A wider range of data would also be beneficial, including specific visits to the ED, primary care and secondary care data, and all personal information that may be relevant. Ideally, data should be sourced from diverse health boards to ensure there is no bias within the models, as this could harm minority groups. Further work should also consider the potential of Neural Networks, including the novel GAN, as this is likely to perform well. This project will also continue, with the goal evolving to focus on profiling of patients visiting the ED, although other tasks could also be considered, such as predicting higher risk regions within Wales or recommended patient medication or condition progression, as machine learning can be utilised for many health-based tasks. Future work should also consider the ethical implications of using AI for healthcare, with research into the attitudes and concerns of the wider public.

Words: 14,922

References

[1] Oliver, D. (2015) David Oliver: Stop blaming patients for emergency visits. BMJ, 351

[2] Weber, E. J., Mason, S., Carter, A., & Hew, R. L. (2011). Emptying the corridors of shame: organizational lessons from England's 4-hour emergency throughput target. *Annals of emergency medicine*, *57*(2), 79-88.

[3] Sullivan, C., Staib, A., Khanna, S., Good, N. M., Boyle, J., Cattell, R., ... & Scott, I. A. (2016). The National Emergency Access Target (NEAT) and the 4-hour rule: time to review the target. *Medical Journal of Australia*, *204*(9), 354-354.

[4] Baker, C. NHS pressures in England: waiting times, demand and capacity. In House of Commons Library, accessed on 1/9/23 at:
https://commonslibrary.parliament.uk/nhs-pressures-in-england-waiting-times-demand-and-capacity/

[5] Scobie, S. (2018). Snowed under: understanding the effects of winter on the NHS. *The Nuffield Trust*, 2018-12.

[6] Dawoodbhoy, F. M., Delaney, J., Cecula, P., Yu, J., Peacock, I., Tan, J., & Cox, B. (2021). AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. *Heliyon*, *7*(5).

[7] Giacomin, J. (2014). What is human centred design?. *The design journal*, *17*(4), 606-623.

[8] Lubberink, R., Blok, V., Van Ophem, J., & Omta, O. (2017). Lessons for responsible innovation in the business context: A systematic literature review of responsible, social and sustainable innovation practices. *Sustainability*, *9*(5), 721.

[9] Bilal Unver, M., & Asan, O. (2022, September). Role of Trust in AI-Driven Healthcare Systems: Discussion from the Perspective of Patient Safety. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (Vol. 11, No. 1, pp. 129-134). Sage CA: Los Angeles, CA: SAGE Publications.

[10] El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.

[11] Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733.

[12] Abufadda, M., & Mansour, K. (2021, December). A survey of synthetic data generation for machine learning. In *2021 22nd international arab conference on information technology (ACIT)* (pp. 1-7). IEEE.

[13] James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: exploring use cases to optimise data utility. *Discover Artificial Intelligence*, *1*(1), 15.

[14] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493-497.

[15] Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., ... & Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, *12*(4), 616-647.

[16] Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: a narrative review. *PLOS Digital Health*, *2*(1), e0000082.

[17] Reese, H. (2017). Understanding the differences between AI, machine learning, and deep learning. *URL: https://www. techrepublic. com/article/understandingthedifferencesbetweenaimachine learninganddeeplearning*.

[18] Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: definitions and differences. *Clinical and experimental dermatology*, *45*(1), 131-132.

[19] Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D., & Tzovaras, D. (2021). Machine learning and deep learning in smart manufacturing: The smart grid paradigm. *Computer Science Review*, *40*, 100341.

[20] Millican, P., & Clark, A. (Eds.). (1996). *Machines and Thought: The Legacy of Alan Turing, Volume I*. Oxford University Press.

[21] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

[22] Grossberg, S. (2020). A path toward explainable AI and autonomous adaptive intelligence: deep learning, adaptive resonance, and models of perception, emotion, and action. *Frontiers in neurorobotics*, *14*, 36.

[23] Battaglia, D., Paolucci, E., & Ughetto, E. (2021). The fast response of academic spinoffs to unexpected societal and economic challenges. Lessons from the COVID-19 pandemic crisis. *R&D Management*, *51*(2), 169-182.

[24] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8* (pp. 563-574). Springer International Publishing.

[25] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., ... & Sharma, R. (2022). Explainable AI for healthcare 5.0: opportunities and challenges. *IEEE Access*.

[26] Xue, S., Yurochkin, M., & Sun, Y. (2020, June). Auditing ml models for individual bias and unfairness. In *International Conference on Artificial Intelligence and Statistics* (pp. 4552-4562). PMLR.

[27] Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, *7*, 154096-154113.

[28] Vezyridis, P., & Timmons, S. (2014). National targets, process transformation and local consequences in an NHS emergency department (ED): a qualitative study. *BMC emergency medicine*, *14*, 1-11.

[29] Cowling, T. E., Soljak, M. A., Bell, D., & Majeed, A. (2014). Emergency hospital admissions via accident and emergency departments in England: time trend, conceptual framework and policy implications. *Journal of the Royal Society of Medicine*, *107*(11), 432-438.

[30] NIHR (2019). Non-urgent attendances to the emergency department are more common among younger adults. doi: 10.3310/signal-000657, accessed on 2/9/23 at: https://evidence.nihr.ac.uk/alert/non-urgent-attendances-to-emergency-departments-are-more-common-among-younger-adults/

[31] Iacobucci, G. (2014). All emergency departments should include GPs, say experts. *bmj*, 1.

[32] Jones, P., & Schimanski, K. (2010). The four hour target to reduce emergency department 'waiting time': a systematic review of clinical outcomes. *Emergency Medicine Australasia*, *22*(5), 391-398.

[33] Martin, S., & Smith, P. (1996). Explaining variations in inpatient length of stay in the National Health Service. *Journal of Health Economics*, *15*(3), 279-304.

[34] Miani, C., Ball, S., Pitchforth, E., Exley, J., King, S., Roland, M., ... & Nolte, E. (2014). Organisational interventions to reduce length of stay in hospital: a rapid evidence assessment.

[35] Jhanji, S., Thomas, B., Ely, A., Watson, D., Hinds, C. J., & Pearse, R. M. (2008). Mortality and utilisation of critical care resources amongst high-risk surgical patients in a large NHS trust. *Anaesthesia*, *63*(7), 695-700.

[36] Goulding, L., Adamson, J., Watt, I., & Wright, J. (2012). Patient safety in patients who occupy beds on clinically inappropriate wards: a qualitative interview study with NHS staff. *BMJ quality & safety*, *21*(3), 218-224.

[37] Hannah, K. J., Ball, M. J., Edwards, M. J., Hannah, K. J., Ball, M. J., & Edwards, M. J. (2006). History of Healthcare Computing. *Introduction to nursing informatics*, 27-40.

[38] David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, *12*(1), 1-22.

[39] Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., & Kennedy, C. J. (2020). Use of latent class analysis and k-means clustering to identify complex patient profiles. *JAMA network open*, *3*(12), e2029068-e2029068.
[40] Raita, Y., Goto, T., Faridi, M. K., Brown, D. F., Camargo, C. A., & Hasegawa, K. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. *Critical care*, *23*(1), 1-13.

[41]Shamout, F.E., Shen, Y., Wu, N. *et al.* An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *npj Digit. Med.* 4, 80 (2021).
https://doi.org/10.1038/s41746-021-00453-0

[42] Álvarez-Chaves, H., Muñoz, P., & R-Moreno, M. D. (2023). Machine learning methods for predicting the admissions and hospitalisations in the emergency department of a civil and military hospital. *Journal of Intelligent Information Systems*, 1-20.

[43]Tahseen Jilani, Gemma Housley, Grazziela Figueredo, Pui-Shan Tang, Jim Hatton, Dominick Shaw,
Short and Long term predictions of Hospital emergency department attendances,
International Journal of Medical Informatics, Volume 129,2019,Pages 167-174,ISSN 1386-5056,
https://doi.org/10.1016/j.ijmedinf.2019.05.011.
(https://www.sciencedirect.com/science/article/pii/S1386505618302429)

[44] Kadri, F., Dairi, A., Harrou, F., & Sun, Y. (2023). Towards accurate prediction of patient length of stay at emergency department: A GAN-driven deep learning framework. *Journal of Ambient Intelligence and Humanized Computing*, *14*(9), 11481-11495.

[45] Chen, K., & Liu, L. (2004). VISTA: Validating and refining clusters via visualization. *Information Visualization*, *3*(4), 257-270.

[46] Di Micco, P., Russo, V., Carannante, N., Imparato, M., Rodolfi, S., Cardillo, G., & Lodigiani, C. (2020). Clotting factors in COVID-19: epidemiological association and prognostic values in different clinical presentations in an Italian cohort. *Journal of clinical medicine*, *9*(5), 1371.

[47] Cervone, H. F. (2011). Understanding agile project management methods using Scrum. *OCLC Systems & Services: International digital library perspectives*, *27*(1), 18-22.

[48] Amalia, P. P., Hendrawan, A. H., & Riana, F. (2022). Application Of The Waterfall Method In The Final Project Guidance Realization Information System. *Jurnal Mantik*, *6*(2), 1449-1458.

[49] Huang, C. C., & Kusiak, A. (1996). Overview of Kanban systems.

[50] Mobey, A., & Parker, D. (2002). Risk evaluation and its importance to project implementation. *Work study*, *51*(4), 202-208.

[51] Wong, E., Ballew, S. H., Daya, N., Ishigami, J., Rebholz, C. M., Matsushita, K., ... & Coresh, J. (2019). Hospitalization risk among older adults with chronic kidney disease. *American journal of nephrology*, *50*(3), 212-220.

[52] Hawkins, R. C. (2003). Age and gender as risk factors for hyponatremia and hypernatremia. *Clinica chimica acta*, *337*(1-2), 169-172.

[53] Statista, UK median age in UK, by region, accessed on 24/7/23, accessed at: https://www.statista.com/statistics/367796/uk-median-age-by-region/
[54] Python, W. (2021). Python. *Python Releases for Windows*, *24*.

[55] Bisong, E., & Bisong, E. (2019). Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, 59-64.

[56] Oliphant, T. E. (2006). *Guide to numpy* (Vol. 1, p. 85). USA: Trelgol Publishing.

[57] National Government for Wales (2022), Addressing the harms from tobacco, in National Survey for Wales, accessed on 20/7/23 at https://www.gov.wales/tobacco-control-strategy-wales-html#:~:text=Around%2013%25%20of%20adults%20in,2022.

[58] Lebanon, G., El-Geish, M., Lebanon, G., & El-Geish, M. (2018). Essential Knowledge: Data Stores. *Computing with Data: An Introduction to the Data Industry*, 471-493.

[59] Braiek, H. B., & Khomh, F. (2020). On testing machine learning programs. *Journal of Systems and Software*, *164*, 110542.

[60] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.

[61] Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering*, *2020*.

[62] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

[63] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.

[64] Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.

[65] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

[66] TIBCO (2022), What is a Random Forest Classifier? Accessed at https://www.tibco.com/reference-center/what-is-a-random-forest

[67] Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

[68] Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.

[69] Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 283-289.

[70] JavatPoint (2023) Support Vector Machine Algorithm, accessed at: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[71] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986-996). Springer Berlin Heidelberg.

[72] Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems*, *29*(5), 1774-1785.

[73] Nour Al-Rahman Al-Serw (2021) K-nearest Neighbor: The maths behind it, how it works and an example, in Medium.

[74] Müller, B., Reinhardt, J., & Strickland, M. T. (1995). *Neural networks: an introduction*. Springer Science & Business Media.

[75] Gurney, K. (2018). *An introduction to neural networks*. CRC press.

[76] Uzair, M., & Jamil, N. (2020, November). Effects of hidden layers on the efficiency of neural networks. In *2020 IEEE 23rd international multitopic conference (INMIC)* (pp. 1-6). IEEE.

[77] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart and J. Nieto, "A data-driven approach for pedestrian intention estimation," *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, 2016, pp. 2607-2612, doi: 10.1109/ITSC.2016.7795975.

[78] Dayan, P., Sahani, M., & Deback, G. (1999). Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, 857-859.

[79] Wu, J. (2012). *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.

[80] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

[81] Casanova, E. L., Baeza-Velasco, C., Buchanan, C. B., & Casanova, M. F. (2020). The relationship between autism and ehlers-danlos syndromes/hypermobility spectrum disorders. *Journal of personalized medicine*, *10*(4), 260.

[82] McEwen, B. S., & Gianaros, P. J. (2010). Central role of the brain in stress and adaptation: links to socioeconomic status, health, and disease. *Annals of the New York Academy of Sciences*, *1186*(1), 190-222.

[83] NHS Digital (2022) Frequency of drinking in the last 12 months, by age and sex, accessed at https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/part-3-drinking-alcohol#:~:text=In%202021%2C%2079%25%20of%20participants,%25%20and%2076%25%20respectively).

[84] Beauxis-Aussalet, E., & Hardman, L. (2014). Visualization of confusion matrix for non-expert users. In *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings* (pp. 1-2).

[85] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).

[86] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

[87] Garrett, B. E., Martell, B. N., Caraballo, R. S., & King, B. A. (2019). Peer reviewed: socioeconomic differences in cigarette smoking among sociodemographic groups. *Preventing Chronic Disease*, *16*.

[88] Rosengren, A., Smyth, A., Rangarajan, S., Ramasundarahettige, C., Bangdiwala, S. I., AlHabib, K. F., ... & Yusuf, S. (2019). Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the Prospective Urban Rural Epidemiologic (PURE) study. *The Lancet Global Health*, *7*(6), e748-e760.

[89] Covino, M., Russo, A., Salini, S., De Matteis, G., Simeoni, B., Della Polla, D., ... & Franceschi, F. (2021). Frailty assessment in the emergency department for risk stratification of COVID-19 patients aged≥ 80 years. *Journal of the American Medical Directors Association*, *22*(9), 1845-1852.

[90] Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health*, *9*(2).

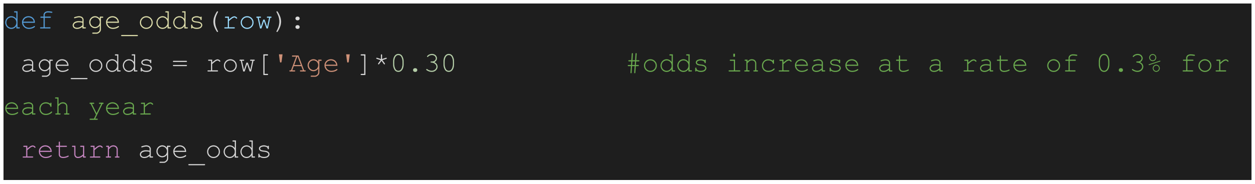
[91] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M. (2017). Big data security and privacy in healthcare: A Review. *Procedia Computer Science*, *113*, 73-80.

[92] Chambers, J. M. (2008). *Software for data analysis: programming with R* (Vol. 2, No. 1). New York: Springer.

Appendix

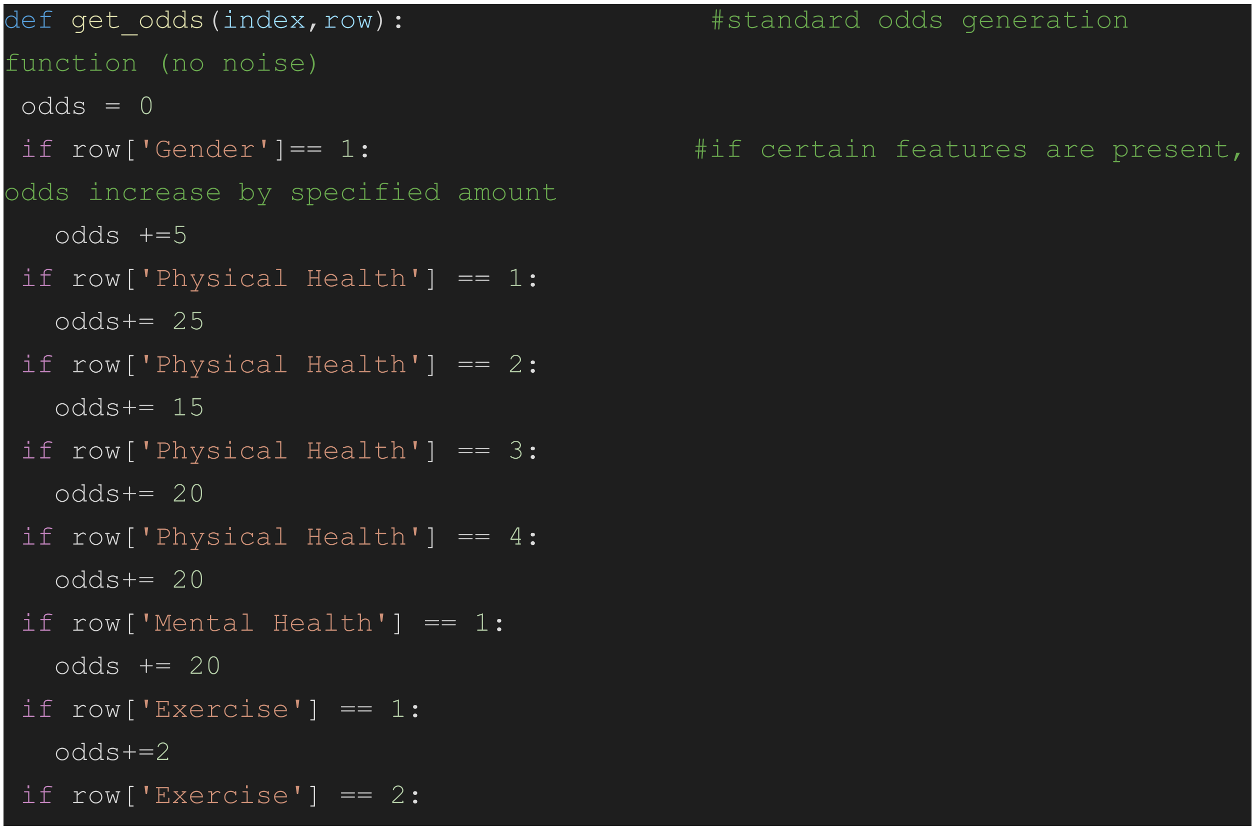
This appendix contains some screenshots of code used within the project. This is only to capture some of the interesting custom elements to the project, and is not a full representation of the code, which has been uploaded separately.

Function to generate odds by age.



```
def age_odds(row):
 age_odds = row['Age']*0.30          #odds increase at a rate of 0.3% for
each year
 return age_odds
```

Function to generate odds (limited noise)



```
def get_odds(index,row):                          #standard odds generation
function (no noise)
 odds = 0
 if row['Gender']== 1:                           #if certain features are present,
odds increase by specified amount
   odds +=5
 if row['Physical Health'] == 1:
   odds+= 25
 if row['Physical Health'] == 2:
   odds+= 15
 if row['Physical Health'] == 3:
   odds+= 20
 if row['Physical Health'] == 4:
   odds+= 20
 if row['Mental Health'] == 1:
   odds += 20
 if row['Exercise'] == 1:
   odds+=2
 if row['Exercise'] == 2:
   odds+=5
 if row['Smoking']== 1:
   odds+=10
 if row['Alcohol Consumption'] == 1:
   odds+= 12
 if row['Socio-economic Status'] == 1:
   odds+= 5
 if row['Socio-economic Status'] == 2:
   odds+=10
 age_odd = age_odds(row)
 return odds + age_odd
```

Generating noisy odds

```
def get_odds_noise(index,row):              #odds function that adds noise
by randomly generating odds increases with large ranges
 odds = 0                                  #I did try smaller ranges but
it didn't add enough noise and models did fine
 if row['Gender']== 1:
   odds +=np.random.randint(8,15)
 if row['Physical Health'] == 1:
   odds+= np.random.randint(12,60)
 if row['Physical Health'] == 2:
   odds+= np.random.randint(12,60)
 if row['Physical Health'] == 3:
   odds+= np.random.randint(12,60)
 if row['Physical Health'] == 4:
   odds+= np.random.randint(15,65)
 if row['Mental Health'] == 1:
   odds += np.random.randint(20,60)
 if row['Exercise'] == 1:
   odds+=np.random.randint(10,20)
 if row['Exercise'] == 2:
   odds+=np.random.randint(10,25)
 if row['Smoking']== 1:
   odds+=np.random.randint(10,45)
 if row['Alcohol Consumption'] == 1:
   odds+= np.random.randint(5,40)
 if row['Socio-economic Status'] == 1:
   odds+= np.random.randint(2,20)
 if row['Socio-economic Status'] == 2:
   odds+= np.random.randint(2,25)
 age_odd = age_odds(row)
 return odds + age_odd
```

Generating labels.

```
def label_generation(odds_list):
   random = np.random.rand(1)
   if random <= odds_list:
     cat = 1
   else:
     cat = 0
   return cat
```

Complex NN example.

```
model = tf.keras.Sequential(layers=[
   tf.keras.layers.Dense(128, activation=tf.nn.relu), #dense layer
   tf.keras.layers.Dense(128, activation=tf.nn.relu),
   tf.keras.layers.Dropout(.1),#dropout to prevent overfitting
   tf.keras.layers.Dense(128, activation=tf.nn.relu),
   tf.keras.layers.Dense((1), activation=tf.nn.sigmoid)])      #number of
neurons = number of potential classes
```

```
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.1e-6),
#low learning rate to prevent overfitting
             loss=tf.keras.losses.BinaryCrossentropy(),
             metrics=tf.keras.metrics.BinaryAccuracy())
```

```
istory = model.fit(scaled_train, y_train, epochs=50, validation_split=0.2,
verbose=1)
```

Simple NN example.

```
model2 = tf.keras.Sequential(layers=[
   tf.keras.layers.Dense(9, activation=tf.nn.relu),
   tf.keras.layers.Dense(8, activation=tf.nn.relu),
   tf.keras.layers.Dense((1), activation=tf.nn.sigmoid)])
```

Generating a graph of all f1 scores across different data sizes.

```
data_NN = {'5000 Samples': F1_NN_5k[1]*100, '15000 Samples':
F1_NN_15k[1]*100, '25000 Samples': NN_F1*100} #saving values and names as
keys
```

```
names_NN = list(data_NN.keys())
values_NN = list(data_NN.values())
data_RF = {'5000 Samples': F1_tree_5k*100, '15000 Samples':
F1_tree_15k*100, '25000 Samples': tree_f1*100}     #each one multiplied by
100 to get percentage
names_RF = list(data_RF.keys())
values_RF = list(data_RF.values())
data_KNN = {'5000 Samples': F1_KNN_5k*100, '15000 Samples':
F1_KNN_15k*100, '25000 Samples': KNN_f1*100}
names_KNN = list(data_KNN.keys())
values_KNN = list(data_KNN.values())
data_SVM = {'5000 Samples': F1_svm_5k*100, '15000 Samples':
F1_svm_15k*100, '25000 Samples': svm_f1*100}
names_SVM = list(data_SVM.keys())
values_SVM = list(data_SVM.values())
plt.plot(names_NN, values_NN, label='Neural Network')       #line plots on
one graph
plt.plot(names_RF, values_RF, label='Random Forest')
plt.plot(names_KNN, values_KNN, label='K-Nearest Neighbour')
plt.plot(names_SVM, values_SVM, label='Support Vector Machine')
plt.ylim(0,100)      #accuaracy is from 0 to 100
plt.legend()
plt.ylabel('F1 Score')
plt.xlabel('Sample Size')
plt.title('F1 Score Across Different Dataset Sizes')
plt.show()
```

Graph for noisy data performance

```
plt.bar("Random Forest",F1_tree_noise*100)        #bar graph to show
accuracies with noisy dataset
plt.bar("SVM",f1_svm_noise*100)
plt.bar("KNN",F1_KNN_noise*100)
plt.bar("Neural Network",F1_NN_noise[1]*100)
plt.title("Accuracies of Classifiers with Noisy Data")
plt.ylim(0,100)
plt.ylabel("F1")
plt.xlabel("Classifier")
plt.show()
```

Link to survey:
https://docs.google.com/forms/d/1AMLNIzjZkj9ccCU8H4I2qWBAQXBZEzxKHXVABY7jeJQ/edit