A close up of a logo

Description automatically generated

Modelling of Factors associated with Hospital Admissions

Final Year Project

Gerard Holian

Department of Mathematics and Statistics

University of Limerick

Supervisor

Cathal Walsh

Abstract

This project will attempt to determine the most important factors that have an effect on hospital admissions. Predictive models will also be built that consist of the most influential factors. These influential factors are plotted to see the varying effects they have on admissions. A preliminary data set which was synthesised and contains a small number of variables is used to trial the models. A final data set of de-indentified patient records with substantially more variables is used to build final models. These models are assessed to see how accurate they are, and if and how they could be used in a real life hospital to improve efficiency.

Acknowledgement

First of all, I would like to extend my gratitude to my supervisor for this project, Prof. Cathal Walsh for his guidance and help throughout the course of this project. It was a pleasure to work with him and he always made himself available to answer any queries I had regarding the project.

Secondly, I would like to thank my second reader Valentina Balbi for her advice and availability throughout the course of the project.

Finally I would like to thank Fionn McCarthy of the planning and performance department of the UL hospitals group for meeting with me and advising me on the topic and also assisting me with my application for data.

Contents

Abstract....................................................................................................................................................2

Acknowledgement...................................................................................................................................3

List of Figures..........................................................................................................................................6

Chapter 1: Introduction ...........................................................................................................................8

1.1 Background .....................................................................................................................................8

1.2 Aims and Objectives........................................................................................................................8

1.3 Problems obtaining data and ensuring data protection ...................................................................9

1.3 Bayesian versus Frequentist Approach..........................................................................................10

Chapter 2: Literature Review ................................................................................................................12

Chapter 3: Methodology....... ................................................................................................................14

3.1 Data................................................................................................................................................14

3.12 Analysis of data.......................................................................................................................14

3.2 Binary Classification .....................................................................................................................15

3.3 Decision Tree.................................................................................................................................16

3.4 Logistic Regression .....................................................................................................................17

3.41 Testing for multicollinearity ..................................................................................................18

3.42 ROC curve and AUC..............................................................................................................18

3.5 Support Vector Machine ..............................................................................................................19

3.5 Naive Bayes Classifier... ..............................................................................................................20

Chapter 4: Preliminary Results .............................................................................................................21

4.1 Analysis of Data ............................................................................................................................21

4.2 Decision Tree.................................................................................................................................24

4.3 Support Vector Machine ..............................................................................................................25

4.4 Naive Bayes Classifier... ..............................................................................................................25

4.5 Logistic Regression Model ..........................................................................................................26

4.51 Variable Selection ..................................................................................................................26

4.52 Muticollinearity test ...............................................................................................................28

4.53 Receiver Operator Characteristic curve (ROC) and Area Under Cure (AUC) ......................29

Chapter 5: Final Results ........................................................................................................................30

5.1 Analysis of Data and Descriptive Statistics ..................................................................................30

5.2 Decision Tree.................................................................................................................................37

5.21 Decision Tree confusion matrices ..........................................................................................38

5.3 Support Vector Machine ...............................................................................................................39

5.31 Support Vector Machine confusion matrices .........................................................................39

5.4 Naive Bayes Model........................................................................................................................40

5.41 Naive Bayes Model confusion matrices .................................................................................40

5.5 Logistic Regression Model............................................................................................................40

5.51 Multicollinearity test ..............................................................................................................41

5.52 Receiver Operator Characteristic curve (ROC) .....................................................................42

5.53 Logistic Regression Model confusion matrices .....................................................................43

Chapter 6: Conclusions..........................................................................................................................44

6.1 How models could be used in real life scenario.............................................................................45

Bibliography ........................................................................................................................................ 48

Appendices ..........................................................................................................................................50

List of Figures

Figure 3.1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14 Figure 3.2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15 Figure 3.3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16 Figure 3.4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17 Figure 3.5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 Figure 3.6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 Figure 4.11 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19 Figure 4.12 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20 Figure 4.13 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20 Figure 4.14 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21 Figure 4.21 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22 Figure 4.22 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23 Figure 4.31 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23 Figure 4.32 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24 Figure 4.33 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24 Figure 4.41 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25 Figure 4.42 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26 Figure 4.43 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26 Figure 4.44 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27 Figure 5.11 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28 Figure 5.12 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29 Figure 5.13 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30 Figure 5.14 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31 Figure 5.15 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32 Figure 5.16 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33 Figure 5.17 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34 Figure 5.21 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35 Figure 5.22 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36 Figure 5.23 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36 Figure 5.31 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37 Figure 5.32 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37 Figure 5.41 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38 Figure 5.42 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38 Figure 5.51 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40 Figure 5.52 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

Figure 5.53 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41 Figure 5.54 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41 Figure 6.1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44 Figure 6.2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

Chapter 1: Introduction

* 1. Background

The issue of overcrowding with regards to the public health system is a major problem in hospitals across the world and is not unique to any country. The body in charge of healthcare for citizens in Ireland, the Health Service Executive (HSE), is being provided with 26 million euro in order to tackle this problem of overcrowding over the winter months (McMahon, 2019). There are a number of consequences associated with overcrowded emergency departments which have a negative impact on healthcare service users including increased waiting times for patients, debilitated patient evaluation and care, a reduced capability of preserving patient confidentiality and privacy, which can result in sick people leaving emergency departments untreated which can result in spreading of infections/diseases and an increased mortality rate (Moskop et al., 2009). Figures revealed that 2018 was the worst year on record for overcrowding in Ireland as 108,227 patients went without hospital beds, a figure almost double of the number in which this statistic was first recorded in 2006 (Irish Nurses and Midwives Organisation, 2019).

The majority of patients presenting to the emergency department are not admitted, however, this route remains the number one source of all hospital admissions with over half of the total admissions (Hong et al., 2018). The first assessment of patients who present to the emergency department is known as triage. This is a process of evaluating the urgency at which the patient requires medical attention. The patient is then seen by a practitioner who determines the course of action with regards to the caring for the patient, which involves a recommendation of whether or not the patient should be admitted (Cronin, 2003).

Prediction models are widely used in medicine as they can boost the efficiency of a hospital by anticipating outcomes and thereby improving the quality of care provided to patients. The use of machine learning using large data sets enables the prediction of a several clinical questions such as bed demand and disease prediction (Vodenberg, 2009).

The models will be built initially using a synthesised data set provided by my supervisor that contains the variables age, gender, triage category, arrival by ambulance and outcome (admission or non-admission). The models will then be built on a dataset that contains more variables.

* 1. Aims and Objectives

The main aim of this study is to model the factors that predict admissions of individuals presenting to the emergency department at the triage time. Triage is an evaluation technique used in the emergency department where the triage nurse assigns a number from 1 to 5 to each patient. A score of 1 indicates that the patient requires urgent medical attention and is high risk while a score of 5 indicates that the patient is low risk and could be seen elsewhere for example at a general practitioner (GP). Triage is a regularly used evaluation system in overcrowded emergency departments to decide which patients require urgent attention and which ones are at less risk (Hinson et al., 2018).

The data set will be split in two in order to train and test the models. The ratio in which it will be split is 75% of the data to train the models and 25% to test the model’s accuracy. The two things that will be considered when looking at the models are: the accuracy of the outcome variable and the explanatory variables that have the biggest impact on the outcome variables i.e. the explanatory variables that are statistically significant. The ability to predict admissions would enable better decision-making and management and understanding what factors increase likelihood of admission could allow triage nurses to evaluate patients with more precision and allow decision makers to be more efficient in terms of staffing and bed management. . The body in charge of reviewing clinical practices, appraising structures, processes and outcomes to improve the healthcare system is called the National office of clinical Audit (National Office for Clinical Audit, 2020).

The predictive models that will be used in this study are a decision tree, a support vector machine, a Naïve Bayes model and a logistic regression model. The objective of these models will be to arrive at a final prediction of admission or a discharge. All models will be compared using the confusion matrix and the measures associated with it. This will reveal which of the models is the most accurate at predicting admissions. It will also be assessed how these models could be used in real life scenarios in an ED to reduce the number of patients on trolleys and allow for more efficient use of resources.

* 1. Problems obtaining data and ensuring data protection

The data set used to create the models in the preliminary analysis was provided by my supervisor Cathal Walsh and will be used as a preliminary data to build models. An application for a real data set from the HSE was sent which was an application form for ethical review of health-related research studies. This meant to ensure that the HSE GDPR policy for third party research was adhered to which says that the data processor must protect personal data through sufficient technical and organisational security measures. To comply with this part, it was intended to synthesize the data in order to protect the data subject’s rights. The data usage must be in line with local and European data legislation and once the study is finished one must be able to show evidence that the data has been deleted (Health Research Regulations 2018, 2020).

After applying for data and receiving ethical approval, getting external approval was the last step needed. Upon the outbreak of the corona virus, it became clear that it was unlikely the data would be received due to the strains put on the medical industry as a whole. Other options were explored and a suitable data set was found online that contained all the variables required. After exploring this data, the project was back on track and the analysis could commence.

* 1. Bayesian versus Frequentist Approach

There are two main approaches to statistics: frequentist and Bayesian. The frequentist approach holds the assumption that the data is randomly generated from a bigger population. Another assumption is that the measures such as the mean and variances are fixed and unknown values in the bigger population. The frequentist attempts to attain estimates of these measures and then calculates sample statistics. Using the Bayesian approach, a statistician uses past knowledge about the underlying parameters. This knowledge is then combined with the observed data to attain probability distributions (Austin et al., 2001).

Chapter 2: Literature Review

It has been revealed that the number of patients on trolleys in 2019 has reached the 100,000 mark (Irish nurse and Midwives Organisation (INMO), 2020). This is only the second time since they started recording the statistic that it has reached this number, with it coming two weeks before it hit this mark in 2018. This shows that in 2019, the number of patients on trolleys is set to hit an all-time high (Moore, 2019). It is a worldwide problem that is getting worse every year. According to the National patient experience survey 69% of patients who reported their waiting times in the ED had to wait for a period of 6 hours or more and of these people over 4% reported that they had to wait for a period exceeding 48 hours. These increasingly long waiting times can have a negative impact on the patients’ health. In the University Hospital of Limerick 28% of patients were waiting for a period of less than 6 hours, 48% of patients were waiting for a period of between 6 and 24 hours and 24% of patients were waiting for a period of over 24 hours (National Patient Experience Survey - HSE.ie, 2020).

A study carried out in Singapore used routine administrative hospital data to predict hospital admissions at triage. A number of models were used including a stepwise regression analysis and a Bayesian network. Data was collected from patients in the ED and included variables on age, sex, ethnic group, admission in the last 3 months, arrival mode and chronic disease status. Data was collected in 2007 and 2008. Out of the 317,581 patient visits to the ED by 207,069 patients, 30.2% were admitted to the hospital immediately. The area under curve (AUC) for the stepwise regression model was 0.849 which means the model is a good predictor of admissions. The predictive model that was developed, that achieved good prediction of the validation set, allows triage nurses to make an early appraisal of a patient’s likelihood of admission and identify patients in deeper need of care allowing earlier admission which can potentially assist with decreasing ED overcrowding (Sun et al., 2011)

A study called ’A conceptual model of emergency department crowding’ was carried out in order to help researchers and administrators and policy makers understand the factors that influence crowding in EDs to allow them to produce potential solutions to solve this issue. The conceptual model produced splits the ED crowding into three parts. The first is input which includes anything that contributes to the demand of ED services, for example, a person developing a serious illness. The second is throughput which looks at the length of a patient’s stay as an influential factor in driving ED demand. Many things determine the length of a patient’s stay including the ability of nurses and doctors to work together well when looking after the patient. The final component is the output which looks at the admitting and discharging of patients and the factors that affect the ability to do this efficiently, for example, not being able to move a patient from the ED to an inpatient ward due to lack of beds (Asplin et al., 2003).

Another similar study named ’Development and Validation of Models to Predict Hospital Admission for Emergency Department Patients’ developed a model capable of predicting an admission of a patient from the ED at the time of triage. In this study, the data of 14,542 individuals who presented to the ED (which included 2,602 admissions) was collected out of a population of around 500,000 in Southwest Ontario in Canada. This data was used to build two models: a logistic regression model and a Coxian phase type distribution model. The data was split into training and validating sets in order to build and test the models. The models and analysis revealed that the most influential variables in predicting admissions are gender, triage level, number of individuals present in the ED and mode of arrival. The receiver operator characteristic curve (ROC) was used in order to discriminate the models. The logistic regression model had a slightly lower area under curve (AUC) than the Coxian phase type with 0.83 and 0.88 respectively. These scores indicated that both models were good predictors of admission from the ED with the Coxian phase type model slightly more accurate at predicting when compared to the logistic regression model. The Coxian phase type model also had the advantage of including timing of the admission (Xie, 2013).

In November 2000, a study called the PATRIARCH study was carried out which stands for PRISA And Triage In A Regional Children’s Hospital. This study attempted to see if an American prediction model could be used in a children’s ED in the United Kingdom. PRISA was the America prediction model and used logistic regression analysis and bootstrapping validation to generate a model estimating the probability of hospital admission. 701 children were studied for this research. When the PRISA model was applied to these children it produced a risk score. Out of the 701 children, 206 were admitted to hospital. The PRISA model had predicted 206.1 admissions so it appeared extremely accurate. The model was assessed using the ROC curve which gave an AUC of 0.76 which indicates it is a fair predictor of admission. The model was able to predict the outcome with an accuracy of 70%. This was not very high but could be used in predicting admissions in paediatric EDs (Miles, 2002).

A study titled ‘Predicting Emergency Department Inpatient Admissions to improve same-day patient flow’ compared three simple methods of predicting admissions at the triage stage. The three methods were: Naïve Bayes, expert opinion and a logit-linear regression model. Triage nurses estimated the probability that patients in triage would be admitted. The logit-linear regression model and the Naïve Bayes model also estimated probabilities of admission for each patient. The nurses estimated the outcome for 767 patients and the models estimated the outcome for 1160 patients. The predictor variables included were age, provider, chief complaint, mode of arrival, emergency severity index (ESI) and designation. The logit-linear function was the best predictor of total beds needed. The ROC curve had an AUC of 0.887 and an R2 of 0.58 with an error of 0.19 in bed need per day. The Naïve Bayes model had an AUC of 0.841 and an R2 of 0.58 with an error of 2.08 in bed need per day. The nurse’s estimation had an R2 of 0.52 with an error of 1.87 per day in total bed need (Peck et al., 2012).

Chapter 3: Methodology

3.1 Data

The models that will be built in this study will first be attempted using a preliminary data set of synthesised data that was provided by my supervisor. The data set was divided into two subsets: one for training the models and one for validating the models. The training subset contained 75% of the data, while the validating subset contained 25% of the data. The variables included in this data set were as follows:

* Admission – yes/no
* Gender – male/female
* Age – 16 and over and broken down to 5 categories: 16-19/20-39/40-59/60-79/over 80
* Triage score – 1/2/3/4/5
* Arrival by ambulance – yes/no

Over the course of this project, an application was sent to receive real hospital data from the HSE and was in the final stages of completion after receiving ethical approval and just awaiting external approval, but after the outbreak of the Covid-19 pandemic it became clear that it was unlikely this data would become available. After some research, a suitable data set was found that contained all the necessary variables required to build the models. This data contained de-indentified patient records and included all Emergency Department visits between March 2014 and July 2017 to the New Haven Health System. The variables included in this data set were as follows:

* Disposition – admission/discharge
* Age- ranging from 18 to 107
* Ethnicity
* Race
* Language – patient’s primary language: English/other
* Religion
* Marital status
* Employment status – full time/part time/student/unemployed/retired
* Insurance status – commercial/Medicaid/Medicare/self-pay
* Department name – presenting hospital: A/B/C
* Arrival Mode – ambulance/walk-in/car/etc
* Arrival Month – January to December
* Arrival day – Monday to Sunday
* Arrival hour bin – 4 hour timeframes: 23-02/03-06/7-10/11-14/15-18/19-22
* ESI – emergency severity index determined by triage nurse
* Previous disposition – disposition of patient’s last visit to the emergency department
* CC – chief complaint: 200 binary variables with different chief complaints

In order to carry out analysis on the data set a number of observations were removed due to having blank values.

3.12 Analysis of data

Histograms of the predictor variables can be used to visualise the effect they have on the outcome variable. Descriptive statistics can also be used to get an insight into the data being used. A number of measures such as median, mode, mean and inter-quartile range can be used to determine the distribution of the data.

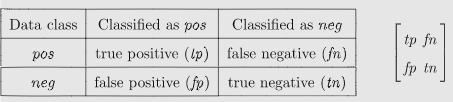
3.2 Binary Classification and Confusion Matrix

A classification problem is when measurements or observations are made on a case or object. These measurements are then used to determine to which of a set of categories it belongs to (Clancey, 1984). The problem we are encountering is a binary classification problem which means that the observations belong to one of two categories: whether the ED visitor will be admitted to hospital or not. Classification problems are widely used in medicine for example predicting whether a patient has an illness or not.

Binary classification results in four possible outcomes.

* True Positives (TP): This is the number of admissions that were correctly predicted as admissions.
* True Negatives (TN): This is the number if discharges that were correctly predicted as discharges.
* False Positives (FP): This is the number of discharges that were incorrectly predicted as admissions. This is known as a type II error.
* False Negatives (FN): This is the number of admissions that were incorrectly predicted as admissions. This is known as a type I error.

These four statistics are used together as metrics to assess the performance of the binary classifier. These metrics include accuracy, misclassification, specificity, sensitivity, and precision (Fawcett, 2006).



*Figure 3.1* (Sokolova and Lapalme, 2009).

* Accuracy: This is the proportion of correctly predicted cases.

***= (TP+TN) / (TP+TN +FP+FN)***

* Misclassification: This is the proportion of incorrectly predicted cases.

***= (FP+FN) / (TP+TN +FP+FN)***

* Specificity: This is the proportion of actual discharges that were correctly predicted.

***= (TN) / (TN +FP)***

* Sensitivity: This is the proportion of actual admissions that were correctly predicted.

***= (TP) / (TP +FN)***

* Precision: This is the proportion of predicted admissions that were actual admissions.

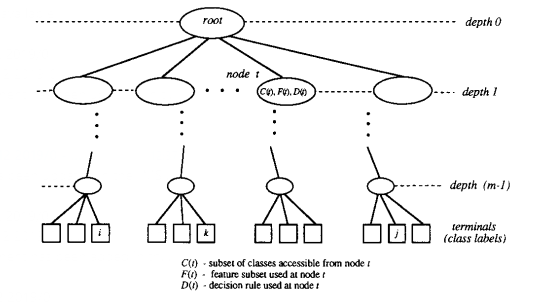
***= (TP) / (TP +FN)***

The above statistics sensitivity and precision are in turn used to calculate the F-score which is an important measure for the predictive power of the model. It is the harmonic mean between of sensitivity and precision of a classifier (Cuadros-Rodríguez et al., 2016). This gives us a score of between 0 and 1.

* F-score ***=[(2)( Precision)(Sensitivity)] / (Precision + Sensitivity)***

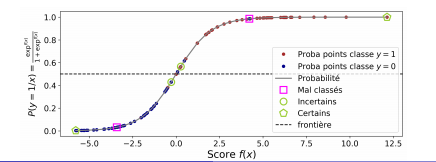
3.3 Decision Tree

A decision tree is a popular method in statistics for tackling problems such as classification using previously recorded data. It is in an inverted tree shape and separates the data into branches. A decision tree is used to predict outcomes by applying a number of decision rules at each node in the tree (Quinlan, 1986). The tree that will be produced is a classification tree due to the outcome variable being binary. The decision tree algorithm that will be used is the CART algorithm which stands for Classification and Regression Trees algorithm. With this algorithm every root node is a dependent variable x. The data is separated so as to fit variables to differentiate new data observations. The layout of a decision tree can be seen in the figure below



*Figure 3.2* (Safavian and Landgrebe, 1991)

3.4 Logistic Regression

Logistic regression is a mathematical model sometimes referred to as the logistic model or the logit model. Its goal is to analyse the relationship between a categorical response variable and predictor variables by predicting probabilities of the data belonging to the categories in question (Park, 2013). In this case the logistic model will try to predict whether an individual will be admitted to hospital or not. When looking at logistic regression, one must first understand simple linear regression: Y = a + BX. This model assumes that the dependent variable Y has a linear relationship with X. But, when modelling admissions the dependent variable Y is a binary variable i.e. it can have only one of two values: 1/0 for admitted/not admitted. Therefore, the simple linear regression will not work as it does not hold the assumption that Y can take any value between −∞ and +∞. The way to model this binary problem is by using probabilities as Y can only be either 1 or 0, p the probability, can be any value between 0 and 1. The probability then gives us an odds ratio = p/1 − p. The odds ratio can be any positive number, we then take the log of the odds ratio = ln(p/(1 − p)) which can be ≥ −∞ and ≤ +∞. A positive value would indicate the probability of admission is greater than 50% and therefore would be classified as an admission. In the figure below the X-axis are the explanatory variables e.g. Age and the Y-axis represents the probability of the outcome admission which in our case will be admission. The cut-off point is 0.5 meaning that if the probability is greater than 0.5 it is rounded off to one indicating an admission and if it less than 0.5 it is rounded down to 0 indicating a non-admission (Gasso, 2019).

*Figure 3.3*

3.41 Testing for multicollinearity

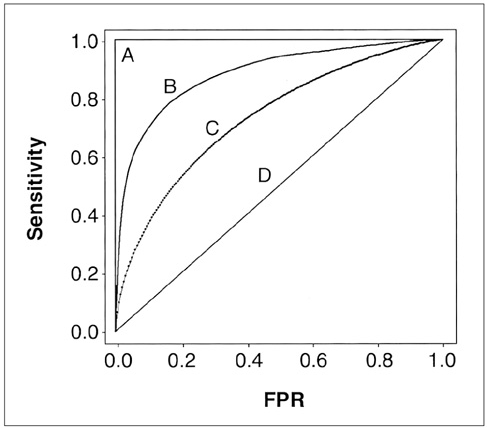
This is a test used to determine if there is a linear relationship between independent variables. The most serious effect of multicollinearity is that it can cause regression coefficients to have large standard errors. This can make coefficients unreliable and decreases the coefficient’s precision (Alin, 2010).

The Variation Inflation Factor is a popular is a popular diagnostic used to test for multicollinearity. If the VIF score exceeds 5 or 10, it indicates regression coefficients are estimated badly due to the presence of multicollinearity (Paul, 2006).

3.42 ROC curve and AUC

The Receiver Operator Characteristic curve (ROC) is a graph used to analyse the performance of classifiers. The ROC curve is a plot of 1-specificity versus sensitivity, also known as 1-false positive rate (FPR) versus the true positive rate (TPR).

The Area Under Curve (AUC) determines the ability of the classifier to discriminate between and admission and a discharge. It is the combined measure of sensitivity and. This statistic can be used to compare classifiers. The maximum score for AUC is 1 which means that the classifier discriminates between admissions and discharges perfectly, while an AUC of 0.5 means that the classifier discriminates between admissions and discharges on complete chance and is considered the minimum level (Tilaki, 2013).

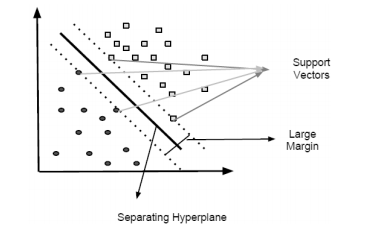


*Figure 3.4*

The figure above shows four different ROC curves with four associated AUCs. D has an AUC of 0.5 which means D distinguishes between outcomes on chance. B and C have an ability to distinguish between outcomes with B being a more powerful classifier. A is a perfect classifier with an AUC of 1 (Tilaki, 2013).

3.5 Support Vector Machine

A support vector machine is a supervised learning algorithm which means that it learns from a set of data. The support vector machine studies the data and its algorithm allow it to classify new data based on the previous data it has learned from. The basic idea of a support vector machine is to separate the data into the two classes by drawing a hyperplane which in simpler terms is a boundary that splits the data. The support vectors are the two data observations that are nearest to the hyperplane. The best hyperplane is the one that has the greatest distance from each of the support vectors. This distance is called the margin (Gandhi, 2018). The figure below shows an example of what a support vector machine does.



*Figure 3.5* (Kuzey, 2012)

3.6 Naïve Bayes Classifier

A screenshot of a cell phone

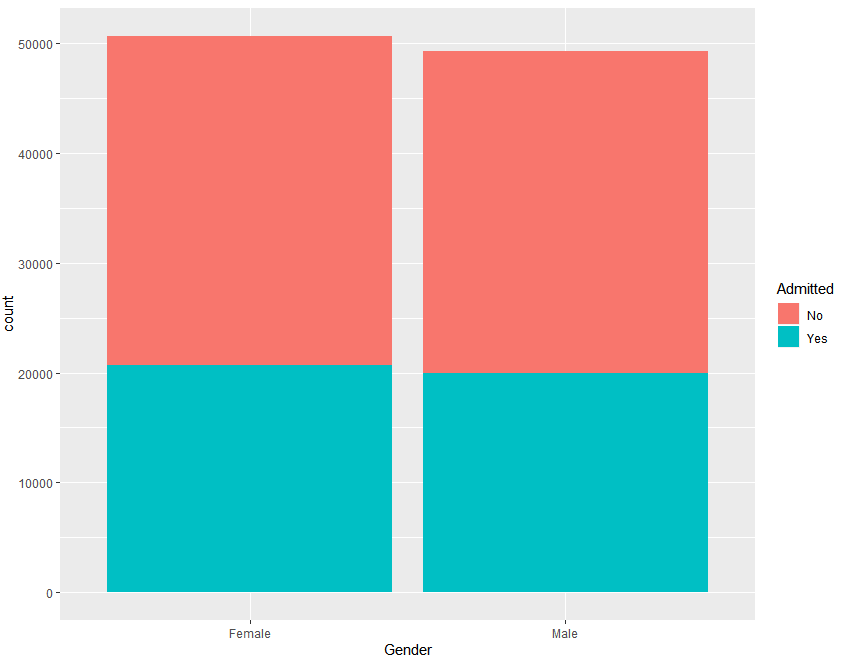
Description automatically generatedThe Naïve Bayes classifier is a classification technique based on Bayes’ theorem concerned with conditional probability. Bayes’ theorem gives a way to update existing probabilities when new information or evidence is made available and it is the foundation of Bayesian statistics. Bayes’ theorem uses prior probabilities which are probabilities before more, or new data is collected to develop posterior probabilities which are probabilities that take into account the new data. The formula for Bayes’ theorem is given in the figure below. Naïve Bayes assumes all explanatory variables are independent of one another; this is where the Naïve part of the name comes from as these explanatory variables are not always independent of one another. It allows classification via probabilities associated with explanatory variables (Medium 2017).

*Figure 3.6* (ThatWare, 2019)

Chapter 4: Preliminary Results

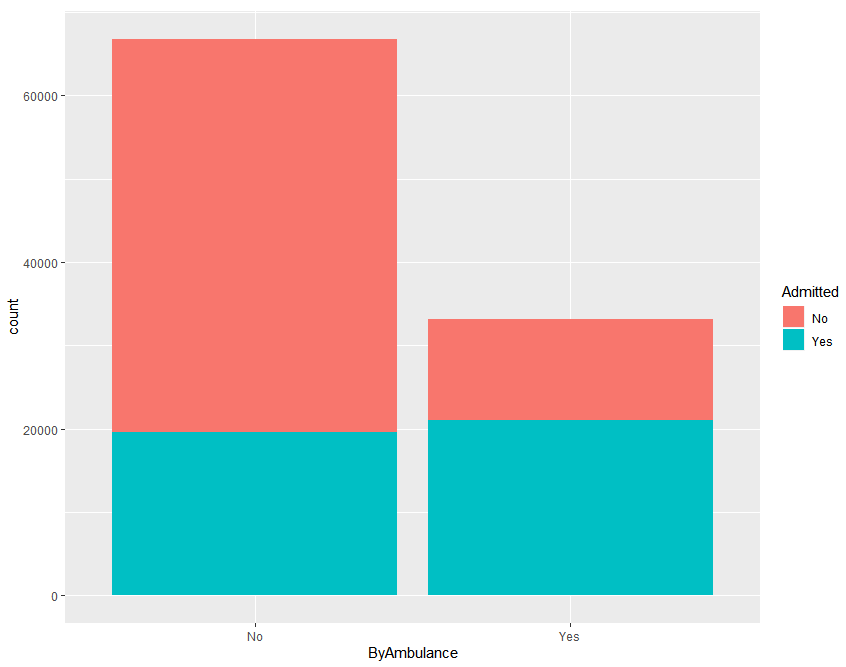
4.1 Analysis of Data

The data was analysed, and histograms were used to show the characteristics of the data. The ratio of admissions was 40.67% of individuals admitted to hospital, leaving 59.33% of individuals not admitted to hospital.



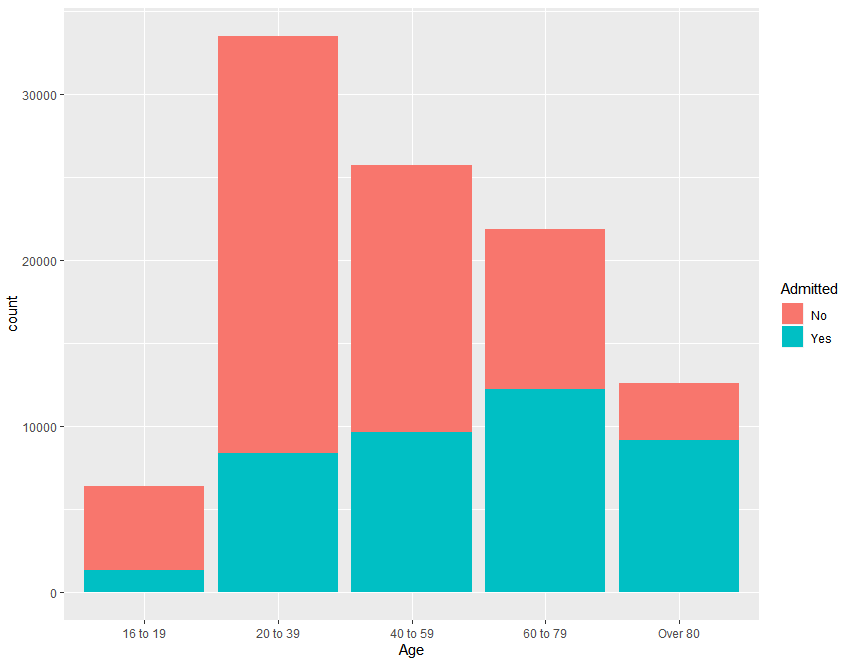
*Figure 4.11*

Figure 4.11 shows the proportion of males and females that were admitted and discharged. Females accounted for 50.68% of the sample while males accounted for 49.32%. The proportion of females admitted was 40.78% and males was 40.56%.



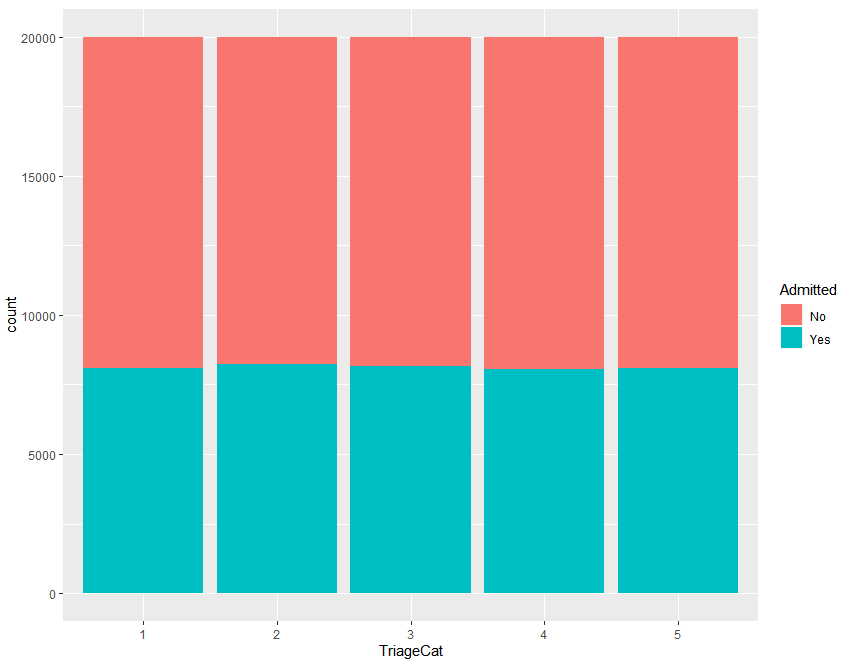
*Figure 4.12*

People who travelled by ambulance to the emergency department accounted for 33.2% while other routes accounted for 55.8% of the total. Arrival by ambulance was associated with a higher rate of admission at 63.36% while other routes had an admission rate of 23.39%.



*Figure 4.13*

The 20 to 39 age group had the highest proportion of total cases with a 33.49% share of total cases, while the 16 to 19 age group had the smallest proportion of total cases with just a 6.37% share. The over 80 age group had the highest admission rate with 72.48% of individuals in that age group being admitted, while the 16 to 19 age group had the lowest admission rate at 20.65%. This could suggest a relationship between age and likelihood of admission.

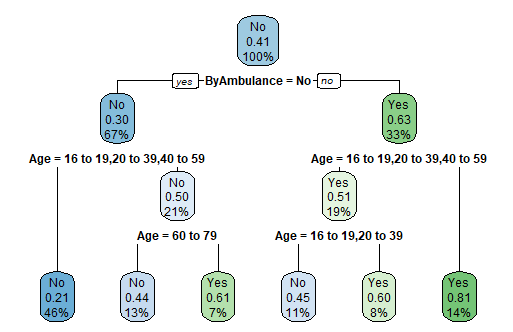


*Figure 4.14*

All triage categories had an equal share of the total (20%) and equal rates of admission due to the synthesised nature of the data set.

4.2 Decision Tree

The functions rpart and rpart.plot were used to produce the decision tree variable selection and plot. The decision tree can be viewed below.



*Figure 4.21*

Interpreting the top row of the decision tree it shows that before any predictor variables are considered, there is a 41% chance of being admitted to hospital from the emergency department (ED). Looking at the next row, if the individual arrived to the ED by Ambulance, they now have a 63% chance of admission from the ED, 33% of individuals took this route. If the individual did not arrive by ambulance, they now have a 30% chance of admission with 67% of individuals taking this route. Looking at the final row, the bottom left node indicates a non-admission with the probability of admission of 0.21. This path was taken by individuals who did not take an ambulance and are a part of the 16 to 19, 20 to 39 and 40 to 59 age groups which accounts for 46% of the individuals recorded. The far-right node indicates an admission to hospital having travelled by ambulance for the over 59 age group. The probability of this was 81% chance of admission and accounts for 14% of all individuals recorded. Similarly, to the log model, the predictors chosen by the decision tree model are ’Age’ and ’ByAmbulance’. The decision tree built was evaluated using a confusion matrix.

A screenshot of a cell phone

Description automatically generated

*Figure 4.22*

The accuracy of this classifier was 71.06%.

4.3 Support Vector Machine

The support vector machine was generated in R using the svm function. The figure below shows the confusion matrix for the support vector machine model.

A screenshot of a cell phone

Description automatically generated

*Figure 4.31*

The accuracy of this classifier was 71.64%

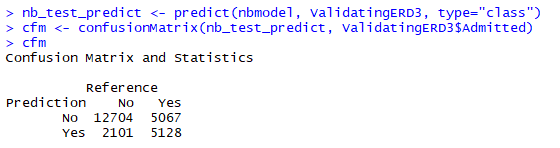
4.4 NaïveBayes Classifier

The Naïve Bayes classifier was generated in using the NaïveBayes function. It produced the following output as can be seen below in the figure. The output produced is a table of conditional probabilities based on Bayes’ theorem. The table shows the percentage of admissions and non-admissions that belong to the subcategories of the explanatory variables.

A screenshot of a cell phone

Description automatically generated

*Figure 4.32*



*Figure 4.33*

4.5 Logistic Regression Model

4.51 Variable Selection

To create the logistic regression model, the statistical software Rstudio was used. The glm function was used to create the model with all independent variables included in the model. The summary function is used to display a summary of the model. The output of the first model can be seen in the figure 4.41.

Interpreting the output from R in the figure, the variables ’ByAmbulance’ and ’Age’ are the strongest predictors of the the outcome variable as they have a p-value of less than 0.05 which means they are statistically significant, whereas the variables ’TriageCat’ and ’Gender’ have pvalues greater than 0.05 which means they have little predictive power in terms of the outcome variable admission. The stepAIC function was then used in order to do a stepwise regression which reduces the number of predictor variables in order to build the best performing logistic regression model. The summary function was then again used to view the output from the new regression model. As can be viewed in the figure, the predictor variables ’TriageCat’ and ’Gender’ were removed during the stepwise regression. The remaining predictor variables ’Age’ and ’ByAmbulance’ all have a p-value of less than 0.05 and therefore have a high significance and are good predictors of admission.

A screenshot of a cell phone

Description automatically generated

*Figure 4.41*

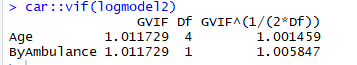
A screenshot of a cell phone

Description automatically generated

*Figure 4.42*

4.52 Multicollinearity Test

In order to check for a linear relationship between independent variables in the final model, a multicollinearity test had to be conducted. The car::vif (Variation inflation factor) function was used to test for this. As can be seen in the figure below there was little to no multicollinearity as the score was very close to one (a score of one indicates no multicollinearity). Since the VIF score was less than 1.1, it can be concluded that there is no issue with multicollinearity between independent variables.



*Figure 4.43*

4.53 Receiver Operator Characteristic Curve (ROC) and Area Under Curve (AUC)

The ROC curve is a graphical representation of two indices used to evaluate the accuracy of the model. The two indices used in this graph are sensitivity and specificity. Sensitivity is also referred to as the true positive rate and is the percentage of admissions that were correctly predicted as such by the log model. Specificity is also referred to as the true negative rate and is the percentage of non-admissions correctly predicted as such by the log model. The ROC curve is then plotted using sensitivity (y-axis) versus 1-specificity (x-axis). The ROC curve was created in R. First, using the predict function to predict admissions and non-admissions using the log model which created a list of admissions outcomes. Then the roc function was used to create the curve with the predicted admissions now attained and the actual admissions from the original data. The R output can be seen in the figure below. AUC stands for area under the curve and gives a measure of performance of the model. The AUC value is between 0 and 1. An AUC of 1 indicates the model predicts with 100% accuracy and an AUC of 0.5 would mean the model would predict on absolute chance. The final log model has an AUC of 0.7549 which means the model predicts a reasonably good accuracy. The ROC curve along with the AUC is shown in the figure below.

A screenshot of a cell phone

Description automatically generated

*Figure 4.44*

Chapter 5: Final Results

The data used to generate the results in this chapter was can be found on kaggle.com and is from a retrospective study that included all Emergency Department visits between March 2014 and July 2017 to the New Haven Health System. The data set is de-identified. The outcomes were either an admission or a discharge. The data had over 500,000 observations and a random sample of 50,000 was chosen to use in the analysis. The data was split into training and validating data at a ratio of 3:1 so the training data set had 37,500 observations and the validating data set had 12,500.

5.1 Analysis of data and Descriptive Statistics

A screenshot of a cell phone

Description automatically generatedA number of the independent variables were analysed in terms of the dependent variable to determine if there was discrepancies between the amounts of patients being admitted among different categories of a variable. This analysis was carried out on the training data set.

*Figure 5.11*

Figure 5.11 shows the proportion of males and females presenting to the Emergency Department with the red showing the proportion that resulted in admissions. The sample contained 20,655 females and 16,845 males. Exactly 5,999 females were admitted, while 5158 males were admitted. Therefore, 29.04% of females were admitted while 30.62% of males were admitted. There are a slightly less proportion of females being admitted.

A close up of a logo

Description automatically generated

*Figure 5.12*

Figure 5.12 above shows the number of cases presenting to the Emergency Department within six 4-hour bins. The most popular time for arrival was from 11AM to 2PM with a 25.69% share of all presentations, while the least popular was from 3AM to 6AM with a 5.77% share of all presentations. The 3AM to 6AM was the period which had the largest proportion of admissions with 34.03% while the period of 11PM to 2AM had the smallest proportion of admissions with just 27.19 of cases resulting in admissions.

A picture containing screenshot

Description automatically generated

*Figure 5.13*

Figure 5.13 above shows the amount of admissions and discharges associated with the mode of arrival to the Emergency Department. Car and ambulance had the highest number of presentations with a 39.74% and 34.23% share. Police had the lowest share.

Ambulance had the highest proportion of admissions with 47.39% of people who used this mode being admitted. Arrival by car (22.05%), walk-in (19.09%) and public transportation (8.95%) had significantly less proportions of admissions.

*A picture containing drawing

Description automatically generatedFigure 5.14*

Figure 5.14 above shows the amount of admissions and discharges associated with the insurance status of the patient.

Medicare insurance is a federal program that provides health coverage if you are 65+ or under 65 and have a disability, no matter your income. Medicaid is a state and federal program that provides health coverage if you have a very low income. Commercial health insurance is defined as a health insurance plan not administered by the government.

People with Medicaid insurance had the highest proportion of presentations with 34.55% share of a total, slightly more than commercial insurance at 32.65%. Self pay had the lowest proportion with just a 0.29share of total presentations. Self pay had the highest proportion of admissions with 96.36% of people self-paying being admitted. People covered by Medicare insurance had the next highest with 52.86% admitted, then commercial with 30.64% and Medicaid with 20.85%. Only 6.36% of people who had ‘other’ as insurance type were admitted.

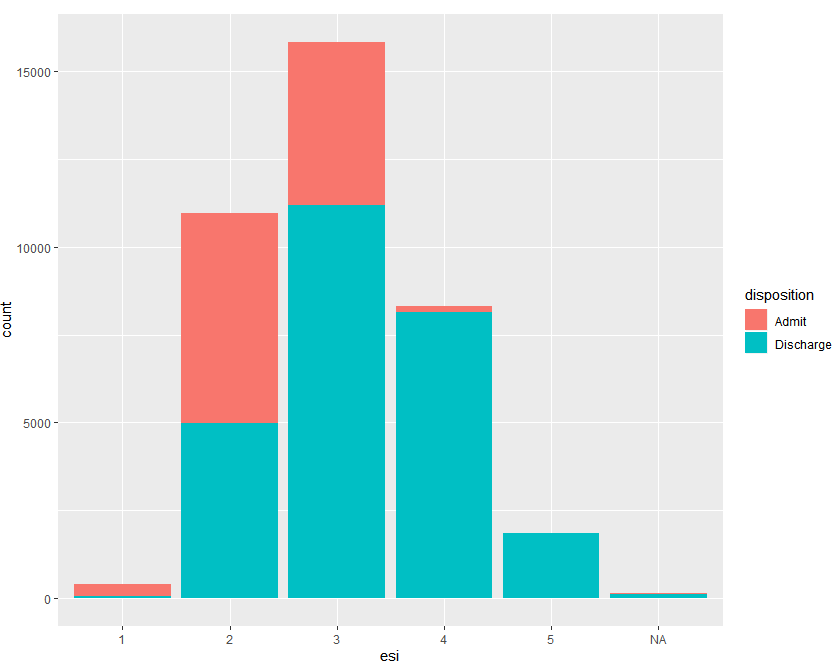
A picture containing pencil

Description automatically generated

*Figure 5.15*

The figure above shows the amount of admissions and discharges associated with the age of the presenting patient to the Emergency Department.

The most popular age groups for presenting to the Emergency department were 24-30 year olds and 50-60 year olds. There appears to be a trend of admission being more likely as age increases as the red part of the bars become more prominent as age increase. This comment is backed by the fact that for the age group of 18-25 year olds, a proportion of 10.11% were admitted, while for those who were 90 years old or older, the proportion of presentations that resulted in admissions was 64.94%.



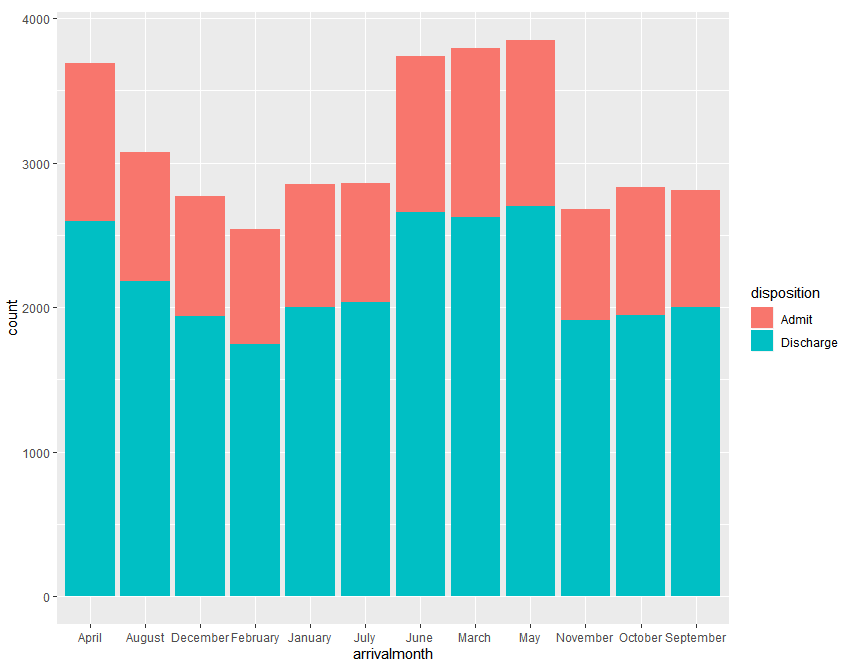
*Figure 5.16*

Figure 5.16 above shows the amount of admissions and discharges associated with the ESI score of the presenting patient to the Emergency Department.

The emergency severity index scale (ESI) ranks patients from 1 to 5, 1 being the category patients who are deemed to be at the highest risk and in greatest need of medical attention, 5 being the lowest risk and in least need of medical attention.

There is a clear relationship between the ESI and the likelihood of admission. The highest risk patients in category 1 had the highest proportion of people admitted with 84.48%, category 2 had 54.47% of patients admitted, category 3 had 29.44% of patients admitted, category 4 had 1.79% of patients admitted and category 5 had just 0.02% of patients admitted.

ESI category 3 had the highest proportion of total patients with a 42.22% share, while category 1 had the smallest share with just 1.08%. There were significant differences in the admission rates between the ESI categories.



*Figure 5.17*

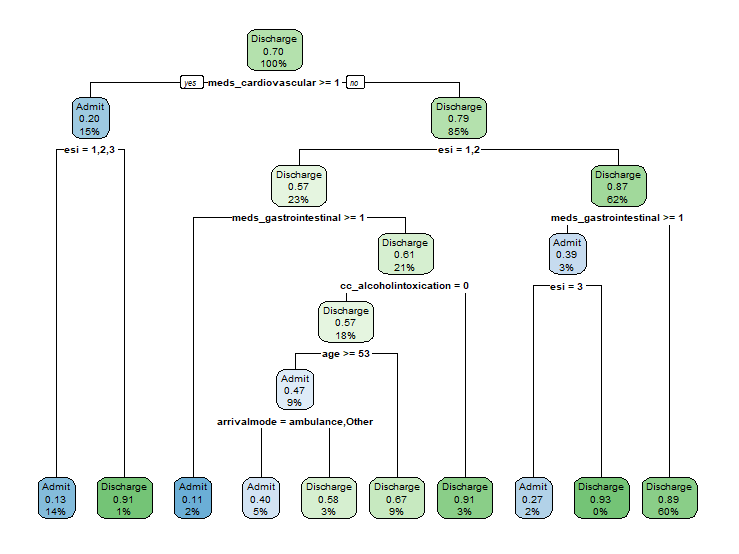
Figure 5.17 above shows the amount of admissions and discharges associated with the month of arrival of the presenting patient to the Emergency Department.

The month of May had the highest proportion of presenting patients with 10.27% of the total, while February had the lowest proportion of the total with 7.38%.

February had the highest proportion of admissions with 31.2% of visitors that month resulting in admissions. July had the lowest proportion of admissions with 28.78% of visitors resulting in admissions. There was very little difference between the admission percentages of the months of the year.

5.2 Decision Tree

The node in question here follows all patients who were not on cardiovascular medication and then gives them a 79% chance of admission with 85% of patients taking this route. Then these patients are split again and where the decision is based on the esi level.



The leaf of the tree indicated by the red arrow denotes all patients who were not on cardiovascular medicine, whose esi was 1 or 2, who were not on gastrointestinal medication, whose chief complaint was alcohol intoxication, who were older than 53 took this route. They have a 47% chance of admission and 9% of patients took this route

*Figure 5.21*

This decision tree model was generated using the rpart function in Rstudio. The plot was generated using the rpart.plot function and figure above shows how at each node a decision is made in order to determine whether the patient will be admitted or discharged.

The model was then run using the both the training and validating data sets. Two confusion matrices of results were generated using the confusionMatrix function in Rstudio.

A screenshot of text

Description automatically generatedThe output below will be used in order to determine the level of performance of the decision tree classifier.

A screenshot of text

Description automatically generated5.21 Decision Tree Confusion Matrices

*Figure 5.22 and figure 5.23*

The confusion matrix output above shows the number of true positives and negatives along with the false positives and negatives that the model predicted on the training data set and validating data set. The accuracy rates of 0.8388 and 0.8318 meant 83.88% and 83.18% of cases were correctly predicted. The sensitivity rates were 0.6175 and 0.6096 which are low. This revealed that 61.75% and 60.96% of actual admissions were predicted correctly. The precision rates were calculated from the matrices and were 0.7948 and 0.7814. This meant that the percentage of cases that were predicted as admissions that were actually admissions were 79.84% and 78.14%. The F-score for the training data was 0.6964 and for the validating data was 0.6472.

5.3 Support Vector Machine

The R package ‘e1071’ was installed in order to use the svm function to fit the support vector machine model to the training data set. The radial kernel was chosen as it was suggested as the best to use on a large data set with a number of independent variables. The support vector machine took a long time to train on the data in Rstudio due the size of the data.

Once built, the support vector machine classifier was assessed using a confusion matrix. The predict function allowed the classifier to make predictions on the outcome of patients in the training and validating data sets. Two confusion matrices of these predictions along with the actual outcomes can be viewed in the two figures below.

The number of predictions is less than the actual amount of outcomes in the data sets due to the classifier being unable to handle data with missing values for the required variables. This meant that the na.omit function had to be used to remove these cases to allow the confusion matrices to be made for evaluation.

A screenshot of a cell phone

Description automatically generatedA screenshot of text

Description automatically generated5.31 Support Vector Machine Confusion Matrices

*Figure 5.31 and figure 5.32*

The accuracy rates of 0.8442 and 0.8357 meant 84.42% and 83.57% of cases were correctly predicted. The sensitivity rates were 0.5803 and 0.5632 which are low. This revealed that 58.03% and 56.32% of actual admissions were predicted correctly. The precision rates were calculated from the matrices and were 0.8715 and 0.8585. This meant that the percentage of cases that were predicted as admissions that were actually admissions were 87.15% and 85.85%. The F-score for the training data was 0.7026 and for the validating data was 0.6802. This classifier performed reasonably well and slightly better than the decision tree.

5.4 Naïve Bayes model

A Naïve Bayes model was fitted to the training data set using the NaïveBayes function which is a part of the e1071 package.

Once built, the Naïve Bayes classifier was assessed using a confusion matrix. The predict function allowed the classifier to make predictions on the outcome of patients in the training and validating data sets. Two confusion matrices of these predictions along with the actual outcomes can be viewed in the two figures below.

A close up of text on a white background

Description automatically generatedA screenshot of text

Description automatically generated5.41 Naïve Bayes Confusion Matrices

*Figure 5.41 and Figure 5.42*

The accuracy rates of 0.6998 and 0.6946 meant 69.98% and 69.46% of cases were correctly predicted. The sensitivity rates were 0.7442 and 0.7502 which are good. This revealed that 74.42% and 75.02% of actual admissions were predicted correctly. The precision rates were calculated from the matrices and were 0.4969 and 0.4939 which are poor. This meant that the percentage of cases that were predicted as admissions that were actually admissions were 49.69% and 49.39% which means predicted admissions had less than complete chance of being correct. The F-score for the training data was 0.5959 and for the validating data was 0.5970. This classifier performed poorly overall.

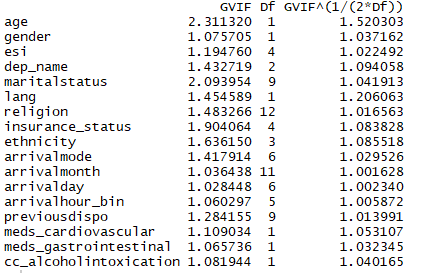
5.5 Logistic Regression Model

As the logistic regression model failed to run a stepwise regression with all the independent variables included, an initial model was ran using all the decision variables from the decision tree along with a number of other variables from the data set that could impact on the likelihood of admission after having discussed this with a member of staff of the UL hospitals group. The extra variables included were gender, department name, language, religion, insurance status, ethnicity, arrival month, arrival hour and previous disposition.

An initial logistic regression model was built using the glm in Rstudio using all the variables mentioned above as well as age, esi, arrival mode, chief complaint being alcohol intoxication and whether the patient is on cardiovascular or gastrointestinal medication. The summary of the model revealed that all independent variables were significant although some were far more significant than others.

5.51 Multicollinearity testing

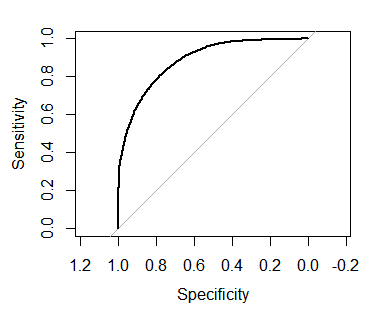
The car::vif function was used to test the logistic regression model for multicollinearity which is a relationship between independent variables. A max value of 5 was determined as the cut off point for the variation inflation factor (VIF) score produced by the above function. A score over 5 would mean that the variable should be dropped from the model and the regression model ran again.

The first multicollinearity test revealed that race had the highest VIF score of 6.155709. This lead to race being removed from the model and the logistic regression model being run again. A new model was produced and a second multicollinearity test was conducted on the model. This test revealed employment status to have a VIF score of 5.3671. This meant that employment status would be removed from the model and the model ran again. The third model was produced and another multicollinearity test was conducted on it. The final test revealed that no variables had a VIF score of 5 or over. This model would be our final logistic regression model. The results of the final multicollinearity test can be viewed in figure below.

*Figure 5.51*

5.52 Receiver Operator Characteristic Curve (ROC)

A plot of the ROC curve allowed the model to be assessed. The ‘pROC’ package was installed in order to use a number of functions to create the plot. The predict function was first used to make predictions of the outcome variable on the validating data set. The roc function was then used to plot the actual outcomes against the predicted outcomes. The ROC plot can be viewed in figure below.



*Figure 5.52*

The auc function was then used to calculate the area under curve. The area under curve was 0.8847 which is a very positive result as it indicates it is a good performance from the classifier.

5.53 Logistic Regression Model Confusion Matrices

Confusion matrices were used in order to further assess the model’s performance. The predict function was used again and probabilities greater than 0.5 were taken as an admission, with probabilities less than or equal to 0.5 taken as a discharge.

A screen shot of a social media post

Description automatically generatedA picture containing knife

Description automatically generated

*Figure 5.53 and figure 5.54*

The accuracy rates of 0.8651 and 0.8507 meant 86.51% and 85.07% of cases were correctly predicted. The sensitivity rates were 0.8407 and 0.8137 which are good. This revealed that 84.07% and 81.37% of actual admissions were predicted correctly. The precision rates were calculated from the matrices and were 0.6939 and 0.6739 which are poor. This meant that the percentage of cases that were predicted as admissions that were actually admissions were 69.39% and 67.39% which means predicted admissions had less than complete chance of being correct. The F-score for the training data was 0.7647 and for the validating data was 0.7377. This classifier performed well overall.

Chapter 6: Conclusions

This project began by identifying the main problem to be a binary classification problem. Another main aim at the beginning of this project was to examine and model the most influential factors associated with individual’s being admitted to hospital from the emergency department. A number of models were created in order to try to predict the outcome of the individual as being an admission or a discharge. The models that were created were a logistic regression model, a support vector machine model, a decision tree model and a Naïve Bayes model. These models were all capable of classifying individual’s outcomes based on their triage data. A literature review of relevant material was gathered, which allowed a decision on the approach of analysis that would be required.

A preliminary analysis was carried out on a synthesised data set was used to determine how well the models would work. Building these models using this data allowed me to improve my coding skills on the software Rstudio. The results obtained from this analysis and a meeting with Fionn McCarthy who works for the planning and performance department of the UL hospitals group lead to the conclusion that there were a number of factors that were not variables in the preliminary data that would have an impact on predicting the outcome of patients presenting to the emergency department. It was clear if more thorough data could be obtained it would improve the models performances.

An initial application to receive real hospital admissions data from the HSE was sent but the process got delayed and due to the corona virus pandemic unravelling it became clear that alternative data would need to be sourced. After research, a suitable data set was found which enabled the planned analysis to be carried out. The first action was to clean the data set, which involved removing any observations that had blank cells. Then an analysis of the distribution of some of the more important factors was carried out. This involved creating histograms if the factors that showed the level of admissions among the different categories of the factors.

A random sample of 50,000 observations was obtained from the data set and this was then split into two parts: a training set and a validating set. The training set was used to build the models and the validating set was used to test the performance of the models on new data. The first model that was trained using the data was the decision tree model. The decision tree model used a number of variables including: ESI, age, arrival mode, whether the patient was on cardiovascular or gastrointestinal medication and whether the chief complaint was alcohol intoxication. To determine the model’s performance it was then ran using the validating data set. This resulted in 83.18% accuracy rate and F-score of 0.6472 which was a good accuracy rate.

A Naïve Bayes model was the next model to be fitted using the training data set. The model built had a much lower accuracy rate (69.46%) and F-score (0.5970) than the decision tree model when tested using the validating data set. This was more than likely due to the size of the data set as Naïve Bayes’ models are more suited to train and predict on smaller data sets.

The next model that was built on the training data set was the support vector machine model. After evaluating using the validating data this model had an accuracy rate of 83.57% and an F-score 0.6802. This model performed well and exceeded the performance of the decision tree model slightly.

The final model that was trained using the training data set was the logistic regression model. I made a number of attempts to train this model using all the independent variables but Rstudio crashed every time. After this, it was decided to use the variables included in the decision tree model along with a number of other variables I that were thought to may have an impact on the likelihood of admission. The variables used were age, gender, ESI, department name, employment status, marital status, language race, religion, insurance status, ethnicity, arrival mode, arrival month, arrival day, arrival hour, previous disposition outcome, medication =cardiovascular, medication= gastrointestinal and complaint=alcohol intoxication. All of these variables were significant so a multicollinearity test was carried out to determine if there were relationships between the independent variables. The first multicollinearity test resulted in the removal of the variable race due to a VIF score exceeding 5. The model was ran again and another multicollinearity test done which resulted in employment status being removed. The model was then trained again and this time no variable had a VIF score higher than 5. This meant this was the final model. The ROC curve was used to assess the model and the AUC of 0.8847 revealed that the model was a very good at predicting the outcome variable. In order to compare the ability of the model it was ran on the validating data set. This model did not have the highest accuracy rate at 81.37% however it had the best F-score with 0 .7377.

In conclusion, the decision tree model, the support vector machine and the logistic regression model all performed well when predicting the patient outcome. They all could be used in a real life Emergency Department scenario to improve bed management and decision making. The easiest model to implement would be the decision tree model as a simple diagram could be followed that would give a prediction at a reasonably high rate of accuracy. The factors that the outcome varied across the most were ESI, mode of arrival, age and insurance status.

This study has shown that modelling the factors associated with hospital admissions could be very beneficial in the planning of hospitals in terms of staffing and bed management – by looking at the times series graphs of arrival day/month/hour, but also in terms of predicting likelihood of admission using the predictive models outlined above. These could be used to make hospitals more efficient in terms of costs and patient care.

6.1 How models could be used in real life scenario

The final part of this project will look at how the models that were built could be used in a real emergency department and how they could improve efficiency. The models could be implemented by inputting the triage data of each visitor and running the models resulting in a risk score. These risk scores could then be used to predict bed demand and could be updated every hour to allow decision makers and bed managers to plan ahead and organise beds equal to the predicted demand. The logistic regression model would work well to do this as it gives probabilities, so for every patient that has a risk score of greater than 0.5 that would equal 1 bed demanded. The decision tree model and support vector machine model would also work.

**Step 1**: Patient data is input into system

Example:

* age
* Gender
* Triage Score
* Arrival mode

**Step 2**: Model produces risk score

Example: 4 patients visited from 1PM to 2PM

* Patient 1: 0.32
* Patient 2: 0.79
* Patient 3: 0.21
* Patient 4: 0.95

**Step 3**: Bed demand calculated

Number of risk scores > 0.5 = bed demand

Therefore bed demand = 2

*Figure 6.1*

Another way in which the models could be used is to use the models in conjunction with a rapid assessment team. The logistic regression model could be used as it gives probabilities. A big problem caused by overcrowding of emergency departments is patient walkouts. These patients are usually low risk patients who are not in need of urgent medical attention. The role of the rapid assessment team of doctors and nurses is to assess patients these patients who are at risk of walking out if not seen to on time. The predicted time of a patient waiting could be predicted using the risk score or triage score (Grant et al.,1999). One method that could be explored is to determine a cut-off point of the risk scores. Patients whose risk score is under this cut-off point would be seen by the rapid assessment team so as to reduce patient walkouts.

**Step 1**: Patient data is input into system

Example:

* age
* Gender
* Triage Score
* Arrival mode

**Step 2**: Model produces risk score and estimated maximum waiting time

Example: 4 patients visited from 1PM to 2PM

* Patient 1: 0.15
* Patient 2: 0.79
* Patient 3: 0.21
* Patient 4: 0.95

**Step 3**: Cut-off point = 0.3

Patients risk score < 0.3 = must be seen by rapid assessment team before max waiting time expires

Therefore rapid assessment team must assess patients 1 and 3 before their max waiting time expires

*Figure 6.2*

Bibliography

Alin, A., 2010. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), pp.370-374.

Asplin, B., Magid, D., Rhodes, K., Solberg, L., Lurie, N. and Camargo, C., 2003. A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42(2), pp.173-180.

Austin, P., Naylor, C. and Tu, J., 2001. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *Journal of Evaluation in Clinical Practice*, 7(1), pp.35-45.

Clancey, W., 1984. *Classification Problem Solving*. Stanford, CA: Dept. of Computer Science, Stanford University.

Cronin, J., 2003. The introduction of the Manchester triage scale to an emergency department in the Republic of Ireland. *Accident and Emergency Nursing*, 11(2), pp.121-122.

Cuadros-Rodríguez, L., Pérez-Castaño, E. and Ruiz-Samblás, C., 2016. Quality performance metrics in multivariate classification methods for qualitative analysis. *TrAC Trends in Analytical Chemistry*, 80, pp.612-624.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861-874.

Gandhi, R., 2018. *Support Vector Machine — Introduction To Machine Learning Algorithms*. [online] Medium. Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

Gasso, G., 2020. *Logistic Regression*. [ebook] Available at: <https://moodle.insa-rouen.fr/pluginfile.php/7984/mod\_resource/content/6/Parties\_1\_et\_3\_DM/RegLog\_Eng.pdf>.

Grant, S., Spain, D. and Green, D., 1999. Rapid assessment team reduces waiting time. *Emergency Medicine*, *11*(2), pp.72-77.

Hinson, J., Martinez, D., Schmitz, P., Toerper, M., Radu, D., Scheulen, J., Stewart de Ramirez, S. and Levin, S., 2018. Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. *International Journal of Emergency Medicine*, 11(1).

Hong, W., Haimovich, A. and Taylor, R., 2018. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, 13(7), p.2.

Hrb.ie. 2020. *Health Research Regulations 2018 FAQ*. [online] Available at: <https://www.hrb.ie/funding/gdpr-guidance-for-researchers/health-research-regulations-2018/health-research-regulations-2018-faq/>.

HSE.ie. 2020. *National Office For Clinical Audit (NOCA) - HSE.Ie*. [online] Available at: <https://www.hse.ie/eng/about/who/qid/measurementquality/nationalofficeclinicalaudit/> [Accessed 2020].

HSE.ie. 2020. *National Patient Experience Survey - HSE.Ie*. [online] Available at: <https://www.hse.ie/eng/about/who/acute-hospitals-division/national-patient-experience-survey/>.

Irish nurse and Midwives Organisation (INMO), 2020. *100,000 Without Hospital Beds So Far This Year*. [online] Available at: <https://inmo.ie/Home/Index/217/13543>.

Irish Nurses and Midwives Organisation, 2019. *Media Press Release 2019 Breaks Record For Most Patients On Trolleys – And It’S Not Even December*. [online] Available at: <https://inmo.ie/Home/Index/217/13549> [Accessed 4 December 2019].

Kuzey, C., 2012. *Illustration Of Support Vector Machine (SVM)*. [image] Available at: <https://www.researchgate.net/profile/Cemil\_Kuzey/publication/260201190\_European\_Journal\_of\_Economic\_and\_Political\_Studies\_Impact\_of\_Health\_Care\_Employees'\_Job\_Satisfaction\_On\_Organizational\_Performance\_Support\_Vector\_Machine\_Approach/links/5c0b3dbd299bf139c746f8e5/European-Journal-of-Economic-and-Political-Studies-Impact-of-Health-Care-Employees-Job-Satisfaction-On-Organizational-Performance-Support-Vector-Machine-Approach.pdf>.

McMahon, A., 2019. HSE’s 26m euro Winter Plan to focus on homecare packages and overcrowding. *The Irish Independent*, [online] Available at: <https://www.independent.ie/breaking-news/irish-news/hses-26m-euro-winter-plan-to-focus-on-homecare-packages-and-overcrowding-38692891.html> [Accessed 22 November 2019].

Medium. 2017. *An Introduction To The Naive Bayes Algorithm (With Codes In Python And R)*. [online] Available at: <https://medium.com/@hackerearth/an-introduction-to-the-naive-bayes-algorithm-with-codes-in-python-and-r-7c85cdb03490>.

Miles, H., 2002. The PATRIARCH Study. Using outcome measures for league tables: Can a North American prediction of admission score be used in a United Kingdom children's emergency department?. *Emergency Medicine Journal*, 19(6), pp.536-538.

Moore, A., 2019. Trolley numbers for 2019 highest since records began, union claims. *The Irish Independent*, [online] Available at: <https://www.independent.ie/breaking-news/irish-news/trolley-numbers-for-2019-highest-since-records-began-union-claims-38738266.html>.

Moskop, J., Sklar, D., Geiderman, J., Schears, R. and Bookman, K., 2009. Emergency Department Crowding, Part 1—Concept, Causes, and Moral Consequences. *Annals of Emergency Medicine*, 53(5), pp.605-611.

Park, H., 2013. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, 43(2), p.154.

Paul, R., 2006. *Multicollinearity: Causes, Effects And Remedies*. [ebook] Available at: <https://www.researchgate.net/profile/Ranjit\_Paul2/publication/255640558\_MULTICOLLINEARITY\_CAUSES\_EFFECTS\_AND\_REMEDIES/links/004635371a7b335d9f000000.pdf>.

Peck, J., Benneyan, J., Nightingale, D. and Gaehde, S., 2012. Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow. *Academic Emergency Medicine*, 19(9), pp.E1045-E1054.

Quinlan, J., 1986. Induction of decision trees. *Machine Learning*, 1(1), pp.81-106.

Safavian, S. and Landgrebe, D., 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), pp.660-674.

Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), pp.427-437.

Sun, Y., Heng, B., Tay, S. and Seow, E., 2011. Predicting Hospital Admissions at Emergency Department Triage Using Routine Administrative Data. *Academic Emergency Medicine*, 18(8), pp.844-850.

ThatWare, 2019. *Naive Bayes : The Definitive Guide*. [image] Available at: <https://thatware.co/naive-bayes/>.

Tilaki, K., 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, [online] 4(2). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/>.

Vodenberg, F., 2009. Predictive and Prognostic Models: Implications for Healthcare Decision-Making in a Modern Recession. *American Health and Drug Benefits*, [online] 2(6), pp.218-222. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4106488/>.

Xie, B., 2013. Development and Validation of Models to Predict Hospital Admission for Emergency Department Patients. *International Journal of Statistics in Medical Research*, 2(1).

Appendices

R Code

#plot and count of gender v admitted

ggplot(ERData) +

geom\_bar(aes(x = Gender, fill = Admitted))

Traincounts1 <- ERData %>%

count(Gender, Admitted)

Traincounts1

#plot and count of age v admitted

ggplot(ERData) +

geom\_bar(aes(x = Age, fill = Admitted))

Traincounts2 <- ERData %>%

count(Age, Admitted)

Traincounts2

#plot and count of byambulance v admitted

ggplot(ERData) +

geom\_bar(aes(x = ByAmbulance, fill = Admitted))

Traincounts3 <- ERData %>%

count(ByAmbulance, Admitted)

Traincounts3

#plot and count of triagecategory v admitted

ggplot(ERData) +

geom\_bar(aes(x = TriageCat, fill = Admitted))

Traincounts4 <- ERData %>%

count(TriageCat, Admitted)

Traincounts4

Traincounts5 <- ERData %>%

count( Admitted)

Traincounts5

# cleaning data and getting random sample

set.seed(16170571)

dfs<-df[sample(nrow(df), 50000), ]

dfs2<-dfs[, -c(301:562)]

dfs3<-dfs2[, -c(328:620]

ERDsplit <- sample.int(n = nrow(dfs3), size = floor(.75\*nrow(dfs3)), replace = F)

TrainingERD <- dfs3[ERDsplit, ]

ValidatingERD <- dfs3[-ERDsplit, ]

#gender admission/discharge count and plot

Train\_counts1<- TrainingERD %>%

count(gender, disposition)

Train\_counts1

g1<-ggplot(TrainingERD) +

geom\_bar(aes(x = gender, fill = disposition))

g1

#arrival hour admission/discharge count and plot

Train\_counts2<- TrainingERD %>%

count(arrivalhour\_bin, disposition)

Train\_counts2

G2<-ggplot(TrainingERD) +

geom\_bar(aes(x = arrivalhour\_bin, fill = disposition))

G2

#arrival mode admission/discharge count and plot

Train\_counts3<- TrainingERD %>%

count(arrivalmode, disposition)

Train\_counts3

G3<-ggplot(TrainingERD) +

geom\_bar(aes(x = arrivalmode, fill = disposition))

G3

#insurance status admission/discharge count and plot

Train\_counts4<- TrainingERD %>%

count(insurance\_status, disposition)

Train\_counts4

G4<-ggplot(TrainingERD) +

geom\_bar(aes(x =insurance\_status, fill = disposition))

G4

#age admission/discharge count and plot

Train\_counts5<- TrainingERD %>%

count(age, disposition)

Train\_counts5

G5<-ggplot(TrainingERD) +

geom\_bar(aes(x =age, fill = disposition))

G5

#esi admission/discharge count and plot

Train\_counts6<- TrainingERD %>%

count(esi, disposition)

Train\_counts6

G6<-ggplot(TrainingERD) +

geom\_bar(aes(x =esi, fill = disposition))

G6

#arrival month admission/discharge count and plot

Train\_counts7<- TrainingERD %>%

count(arrivalmonth, disposition)

Train\_counts7

G7<-ggplot(TrainingERD) +

geom\_bar(aes(x =arrivalmonth, fill = disposition))

G7

# decision tree model

library(rpart)

dtree = rpart(disposition ~ ., data = TrainingERD,method="class")

dtree

library(rpart.plot)

rpart.plot(dtree)

summary(dtree)

predDT<- predict(dtree, TrainingERD, type= "class")

install.packages("caret")

library(caret)

DTCM<- confusionMatrix(predDT, TrainingERD$disposition)

DTCM

#Predict on Validating subset

predDT2<- predict(tree, ValidatingERD, type= "class")

#Confusion Matrix for Evaluation

library(caret)

DT2CM<- confusionMatrix(predDT2, ValidatingERD$disposition)

DT2CM

# support vector machine model

install.packages('e1071')

library(e1071)

classifier1 <- svm(formula = disposition ~ ., data = TrainingERD, scale = FALSE, type = 'C-classification', kernel = 'radial')

library(caret)

# predict on training set

svmpt<-predict(classifier1, na.omit(TrainingERD))

svmpt

ter<-(na.omit(TrainingERD))

#CM train

svmcmt<-confusionMatrix(svmpt, ter$disposition)

svmcmt

# predict on validating set

svmp<-predict(classifier1, na.omit(ValidatingERD))

svmp

qer<-(na.omit(ValidatingERD))

#CM valid

svmcmv<-confusionMatrix(svmp, qer$disposition)

svmcmv

#Naïve Bayes model

library(e1071)

library(caret)

xnb<-TrainingERD[,-12]

ynb<-TrainingERD$disposition

nbmodel = NaïveBayes(ynb~., data=xnb)

summary(nbmodel)

#TRaining data confusion matrix

nbpredictt<-predict(nbmodel, TrainingERD)

nbconfmatt<-confusionMatrix(nbpredictt, TrainingERD$disposition)

nbconfmatt

#Validating data confusion matrix

nbpredictv<-predict(nbmodel, ValidatingERD)

nbconfmatv<-confusionMatrix(nbpredictv, ValidatingERD$disposition)

nbconfmatv

# fitting log model

model <- glm(disposition ~ age + gender + esi + dep\_name + employstatus + maritalstatus + lang + race + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication, family=binomial(link='logit'),data=TrainingERD)

summary(model)

car::vif(model)

# remove race due to high vif

model2 <- glm(disposition ~ age + gender + esi + dep\_name + employstatus + maritalstatus + lang + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication, family=binomial(link='logit'),data=TrainingERD)

car::vif(model2)

#remove employment status due to high vif

TrainingERD$disposition<-as.factor(TrainingERD$disposition)

model3 <- glm(disposition ~ age + gender + esi + dep\_name + maritalstatus + lang + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication+ n\_edvisits +n\_surgeries+n\_admissions, family=binomial(link='logit'),data=TrainingERD)

car::vif(model3)

#plot and count of gender v admitted

ggplot(ERData) +

geom\_bar(aes(x = Gender, fill = Admitted))

Traincounts1 <- ERData %>%

count(Gender, Admitted)

Traincounts1

#plot and count of age v admitted

ggplot(ERData) +

geom\_bar(aes(x = Age, fill = Admitted))

Traincounts2 <- ERData %>%

count(Age, Admitted)

Traincounts2

#plot and count of byambulance v admitted

ggplot(ERData) +

geom\_bar(aes(x = ByAmbulance, fill = Admitted))

Traincounts3 <- ERData %>%

count(ByAmbulance, Admitted)

Traincounts3

#plot and count of triagecategory v admitted

ggplot(ERData) +

geom\_bar(aes(x = TriageCat, fill = Admitted))

Traincounts4 <- ERData %>%

count(TriageCat, Admitted)

Traincounts4

Traincounts5 <- ERData %>%

count( Admitted)

Traincounts5

# cleaning data and getting random sample

set.seed(16170571)

dfs<-df[sample(nrow(df), 50000), ]

dfs2<-dfs[, -c(301:562)]

dfs3<-dfs2[, -c(328:620]

ERDsplit <- sample.int(n = nrow(dfs3), size = floor(.75\*nrow(dfs3)), replace = F)

TrainingERD <- dfs3[ERDsplit, ]

ValidatingERD <- dfs3[-ERDsplit, ]

#gender admission/discharge count and plot

Train\_counts1<- TrainingERD %>%

count(gender, disposition)

Train\_counts1

g1<-ggplot(TrainingERD) +

geom\_bar(aes(x = gender, fill = disposition))

g1

#arrival hour admission/discharge count and plot

Train\_counts2<- TrainingERD %>%

count(arrivalhour\_bin, disposition)

Train\_counts2

G2<-ggplot(TrainingERD) +

geom\_bar(aes(x = arrivalhour\_bin, fill = disposition))

G2

#arrival mode admission/discharge count and plot

Train\_counts3<- TrainingERD %>%

count(arrivalmode, disposition)

Train\_counts3

G3<-ggplot(TrainingERD) +

geom\_bar(aes(x = arrivalmode, fill = disposition))

G3

#insurance status admission/discharge count and plot

Train\_counts4<- TrainingERD %>%

count(insurance\_status, disposition)

Train\_counts4

G4<-ggplot(TrainingERD) +

geom\_bar(aes(x =insurance\_status, fill = disposition))

G4

#age admission/discharge count and plot

Train\_counts5<- TrainingERD %>%

count(age, disposition)

Train\_counts5

G5<-ggplot(TrainingERD) +

geom\_bar(aes(x =age, fill = disposition))

G5

#esi admission/discharge count and plot

Train\_counts6<- TrainingERD %>%

count(esi, disposition)

Train\_counts6

G6<-ggplot(TrainingERD) +

geom\_bar(aes(x =esi, fill = disposition))

G6

#arrival month admission/discharge count and plot

Train\_counts7<- TrainingERD %>%

count(arrivalmonth, disposition)

Train\_counts7

G7<-ggplot(TrainingERD) +

geom\_bar(aes(x =arrivalmonth, fill = disposition))

G7

# decision tree model

library(rpart)

dtree = rpart(disposition ~ ., data = TrainingERD,method="class")

dtree

library(rpart.plot)

rpart.plot(dtree)

summary(dtree)

predDT<- predict(dtree, TrainingERD, type= "class")

install.packages("caret")

library(caret)

DTCM<- confusionMatrix(predDT, TrainingERD$disposition)

DTCM

#Predict on Validating subset

predDT2<- predict(tree, ValidatingERD, type= "class")

#Confusion Matrix for Evaluation

library(caret)

DT2CM<- confusionMatrix(predDT2, ValidatingERD$disposition)

DT2CM

# support vector machine model

install.packages('e1071')

library(e1071)

classifier1 <- svm(formula = disposition ~ ., data = TrainingERD, scale = FALSE, type = 'C-classification', kernel = 'radial')

library(caret)

# predict on training set

svmpt<-predict(classifier1, na.omit(TrainingERD))

svmpt

ter<-(na.omit(TrainingERD))

#CM train

svmcmt<-confusionMatrix(svmpt, ter$disposition)

svmcmt

# predict on validating set

svmp<-predict(classifier1, na.omit(ValidatingERD))

svmp

qer<-(na.omit(ValidatingERD))

#CM valid

svmcmv<-confusionMatrix(svmp, qer$disposition)

svmcmv

#Naïve Bayes model

library(e1071)

library(caret)

xnb<-TrainingERD[,-12]

ynb<-TrainingERD$disposition

nbmodel = NaïveBayes(ynb~., data=xnb)

summary(nbmodel)

#TRaining data confusion matrix

nbpredictt<-predict(nbmodel, TrainingERD)

nbconfmatt<-confusionMatrix(nbpredictt, TrainingERD$disposition)

nbconfmatt

#Validating data confusion matrix

nbpredictv<-predict(nbmodel, ValidatingERD)

nbconfmatv<-confusionMatrix(nbpredictv, ValidatingERD$disposition)

nbconfmatv

# fitting log model

model <- glm(disposition ~ age + gender + esi + dep\_name + employstatus + maritalstatus + lang + race + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication, family=binomial(link='logit'),data=TrainingERD)

summary(model)

car::vif(model)

# remove race due to high vif

model2 <- glm(disposition ~ age + gender + esi + dep\_name + employstatus + maritalstatus + lang + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication, family=binomial(link='logit'),data=TrainingERD)

car::vif(model2)

#remove employment status due to high vif

TrainingERD$disposition<-as.factor(TrainingERD$disposition)

model3 <- glm(disposition ~ age + gender + esi + dep\_name + maritalstatus + lang + religion + insurance\_status + ethnicity + arrivalmode

+ arrivalmonth + arrivalday + arrivalhour\_bin + previousdispo + meds\_cardiovascular + meds\_gastrointestinal

+ cc\_alcoholintoxication+ n\_edvisits +n\_surgeries+n\_admissions, family=binomial(link='logit'),data=TrainingERD)

car::vif(model3)