Predicting Hospital Readmission Using Machine Learning

MSc Research Project
Data Analytics

Zainab Mohamed Ismail Makrani
Student ID: 22190082

School of Computing
National College of Ireland

Supervisor: Jorge Basillio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Zainab Mohamed Ismail Makrani
Student ID:	22190082
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Jorge Basillio
Submission Due Date:	31/01/2024
Project Title:	Title
Word Count:	XXX
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Hospital Readmission Using Machine Learning

Zainab Makrani
22190082

Abstract

Rehospitalization is widespread among various patients and can be exhausting not just financially but also mentally due to the frequent hospital visits. The primary objective of this project is to incroporate machine learning techniques into healthcare, to create an affordable and efficient healthcare environment and structure. Simple machine learning techniques like random forest, xgboost, decision trees, linear regression and ordinary least squares regression are applied as well as neural networks. Random forest is best performing model with an accuracy of 82% and f1 score of 22 and r-squared of 0.04. When compared the dense neural network outperformed the rest with an accuracy of 82.88% and test loss of 0.5 while the rest were mostly overfitting to the model after training.

1 Introduction

Hospital revists refers to instances where patients return to healthcare facilities within a span of 30-90 days for the same or related issue to the one treated before discharge. Baig et al. (2019) noted that at least 1 of 5 patients have been re-hospitalized thereby causing congestion in healthcare facilities and overworking related staff which subsequently leads to an inefficient environment to treat patients. A common measure to determine the quality of care provided is closely monitoring the 30 day recovery period especially among surgical patients requiring critical care as this can directly be linked to post-op and follow up care. Ramírez and Herrera (2019) argue that this period can be indicative of whether or not a patient was fully treated before hospital dismissal. Prevailing cases of readmissions are diabetic and cancer patients where former pose a higher risk of revisits as compared to their non diabetic counterparts while the former tend to revisit due to constant chemotherapy, radiotherapy and surgical removal of tumors as well the occurence of complications after discharge Jung et al. (2023) .

Readmission can impose significant costs and burdens, affecting not just healthcare workers but also impacting patients and their families, who frequently bear the responsibility of care. This emphasizes the critical need for an improved system to reduce readmission rates, particularly within the mentioned patient groups. Furthermore, Jung et al. (2023) sheds light on how to better serve cancer patients who tend to be overlooked in standard medical programs due to the elevated death rate and sophistication of the disease thereby proving the urgency for research in this area and further study the sector and improve healthcare systems. Incorporating machine learning models can facilitate accurate predictions of patients that require vigilant monitoring , thereby making it easier for doctors to timely identify any complications arising and intervene earlier on to prevent escalation of any issue. Moreover, such predictions can allow hospitals to minimise the resources put in the process of readmitting patients and optimising those resources which in turn can help lower the overall costs incurred by patients and by the hospital.

Research in this area is not just advantageous to the healthcare department but also tech as machine learning is constantly advancing. Implementation of machine learning model for prediction can enable accurate problem solving hence making it possible to have a smart, affordable and effective healthcare system.Implementing this research has the potential to empower doctors in more effectively managing chronic illnesses like cancer and diabetes, while also improving the allocation of necessary resources such as chemotherapy equipment, insulin, and other essential supplies.

Studies related to this project offer valuable insights, however they do pose different limitations which may vary depending on the research. Geographical limitations is seen in Jung et al. (2023) where data involved is only specific to Korea as the paper discusses predicitve study of unplanned cancer related readmissions, in this case research can be prejudice in global analysis and research. Additionally Jung et al. (2023) focused on 30 of the main attributes of the top-performing model, XGBoost turned out to be the best model. The models were then assessed using a variety of techniques, including the F1 Score, accuracy, precision, and other techniques. Although the research on cancer-related cases is insightful, its applicability may be restricted to Korean data and methodologies. Further studies Rajput and Alashetty (2022) use a variety of machine learning techniques, such as SVM, KNN, and Gradient Boosting. However, this article simply relied on the predictive accuracy of the models to select the best-performing model, which could be misleading in the event of unbalanced data. Moving on, Teo et al. (2020) blends machine learning and rule-based models, but it doesn't validate the findings in various healthcare contexts, which could limit its application in the real world. It appears that the majority of the time, the restriction stems from the need to validate and assess the models that are used to make predictions; but, in other circumstances, data limitations in a particular region may be the cause. Instead of relying solely on one popular technique, such as accuracy, these circumstances can be managed by simply adapting better assessment and validation procedures; researchers that combine many methods are able to have a better recommendation for a prediction model.

The goal of this research is to provide a straightforward method for performing predictive analysis incorporating several machine learning models such as svm, knn, xgboost, gradient boosting and regression. These models will then be assessed using various techniques. Accuracy, precision, recall, f1 score, and confusion matrix will all be included in the model evaluation models where appropriate.

1.1 Research Question

Understanding the predictive factors influencing readmission is crucial in healthcare management. This study intends to investigate the role of machine learning in predicting readmissions, with the goal of improving healthcare. It specifically targets the factors of readmissions and how they can be used to predict such cases which in return will minimise the healthcare cost thereby making it cost effective.

How can machine learning methodologies be leveraged in healthcare to predict and alleviate hospital readmissions effectively?

2 Related Work

This segment critically asses the current research against similar previously done studies. Papers are categorised into 4 different subsections with similar papers for each section. According to this article [1] , initiating a thorough literature review creates a foundational step in the research process . This comprehensive analysis of existing knowledge serves to solidify one's understanding of the subject and pinpoint gaps in the current literature

2.1 Studies related to general hospital admission

Application of machine learning methods including Gradient Boosting, Decision Tree, Neural Networks and Logistic Regression are used to create a prediction model assessing variables leading to hospital readmission which is an increasingly common practice as suggested by Maddipatla et al. (2015).These predictive models are then evaluated using Area under receiver operating characteristic curve (AUC) and misclassification rate, thereby allowing an insight on how well each model performs. Furthermore, visual analysis has enabled hospitals to compute loss of revenue caused by frequent re-admissions as well as enhancing the quality of care in real time.The research Maddipatla et al. (2015) focused on which socio-economic elements affect re-admission of patients within a span of 30 days. In addition to aforementioned evaluation metrics for this research, other model evaluation methods such as accuracy and precision would be helpful in producing an in-depth evaluationMaddipatla et al. (2015) proposes a cost prediction model, which estimates costs related to readmission and expands on the risk variables, which can be further improvised if the machine learning models are assessed using external validation to make sure that the models predict accurately. Predicting general hospital re-admissions can be approached differently where a study can aim to highlight the main challenges faced when predicting re admissions as highlighted by Wang and Zhu (2022) which suggests that the issue can be further divided into 4 main categories in which, each of the issues is handled accordingly. To do this, the study analysed existing dataset and projects as well as comparing what models were implemented in similar studies thereby allowing an insight into how different variables can affect readmission. While the research is significant in understanding the challenges in prediction of readmission, it can be further improved by comparing the models with external validation rather than just using pre existing models for prediction and evaluation.

Another research Zhu et al. (2015) implemented conditional logistic regression(CLR) and classification models with the purpose to look into implementation of CLR to improve accuracy of prediction and is motivated by not only high expenses incurred in healthcare caused by readmission but also the discomfort and toll taken on patients in the processing of re-hospitalization. It was noticeable that the CLR models outmatched the standard classification techniques including regular Logistic Regression and SVM, by an enhanced accuracy of about 20%. The study Zhu et al. (2015) suggested that proper implementation of CLR models will enhance advancements in prediction models, accurately identifying patients requiring extra care. The purpose of this study Zhu et al. (2015) is to raise awareness about the collection of data and the progress made in studying the prediction of readmission risks.Additionally, Length of Stay, Acuity of Admission, Charlson Comorbidity Index, and Emergency Department Use (LACE) -index and Patient At-Risk of Hospital Readmission (PARR) have been used to authenticate performance of predict-

[1] Writing a literature review : `https://doi.org/10.1177/2051415816650133`

ive models, which the study Zhu et al. (2015) considered to be the most commonly used models for risks of readmission. Machine learning models like XGBoost, Adaboost and Random Forests models are used to predict readmissions within a span of 30 days and are evaluated using F1 score, precision and positive predictive value(PPV) and validated using the PARR and LCE models. The research aimed to further add non clinical data to get a better insight into future works and have better predictions of risk. While the researchZhu et al. (2015) was able to create a machine learning models with a considerably greater accuracy, it also implemented a generic model which does not prove to be as significant where the goal is to implement minimal prediction variables in prediction of hospital readmissions.

All in all, each of these studies have a unique perspective of involving machine learning techniques into prediction of general readmission where some focus analysis of pre existing models and their challenges while others are keen on building and evaluating the models. Most commonly used technique for prediction is XGBoost and logistic regression methods with the most preferred being Conditional Logistic Regression.

2.2 Research on the basis of unplanned readmission

Unplanned readmission are the most common and tend to occur commonly among low risk patients and pose harder to forsee since they do not follow conventional trends as those discussed in general readmissions. More often than not they are used to give an insight into the quality of care and effeciency of certain healthcare facilitiesJung et al. (2023). XGBoost implemented by Jung et al. (2023) was declared to have the best performance, where the model included the top 30 features with the main aim being to greatly reduce number of re-hospitalization incidents. While this study incorporated various model evaluation techniques, applying more focus and care on feature selection would enhance better models as out of 4,437 variables only 30 are chosen and comparison of models with different features besides the top 30 can help bring better insights to the evaluation of prediction. Incorporation of different elements and machine learning techniques can allow researchers to maximise on their goal to reduce cases re-hospitalization with a slightly better accurate result.Further models such as time series would allow the researches to provide vast insights thereby giving room for enhanced model selection.

The identification of high risk patients, which is an important first step in enhancing treatment, encourages the implementation of the proper strategies to reduce readmission. Four machine learning techniques applied to a research Lo et al. (2021) to predict re admissions within 14 days included logistic regression, extreme gradient boosting, random forest and categorical boosting and then evaluated using precision, F1-score, recall, area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC). The study by Lo et al. (2021) is very similar to the approach intended for this thesis on machine learning approcahes and models that can be implemented to minimise re-readmission incidents. The machine learning models from Lo et al. (2021) can be applied to identify people that are at a higher risk of readmission allowing for early intervention by healthcare workers.

In summary, both studies are pertinent to this current thesis where one majorly focuses on XGBoost and the other focuses on multiple machine learning techniques. Each of these researches aforementioned assist in making informed decisions regarding feature engineering from large datasets to effectively using the data in prediction of hospital re-admissions

2.3 Research related to specific diseases

As highlighted earlier diabetic patients are at a higher risk of readmission as compared to their non-diabetic counterparts. Prediction aimed especially towards diabetic patients can aid in improving overall health of patients by detecting risks of re-hospitalization in good time to control the effects of the disease.Simple machine learning techniques like logistic regression, svm, neural networks and random forests, where the last one had a comparatively better performance, are applicable independently for successful prediction without needing deep learning techniques as proven by Ramírez and Herrera (2019). Various evaluation metrics applied are proof to the well performing models. This approach is aspired to be used in the current thesis to ensure better performing and accurate models. Random forest is seen to be the least implemented technique in comparison to other studies discussed earlier. A key improvement that could potentially enhance the study by Ramírez and Herrera (2019) is the process of variable selection, where the machine learning models are not only inclined towards considering the patients' first visit only. High glucose levels can also be taken into keen consideration when implementing diabetic readmission models as it is directly associated to the disease. This can be achieved by application of different machine learning techniques such as knn, logistic regression, gradient boosting (GB), and Gaussian naive bayes (GaussianNB) to establish models' accuracy as done by Rajput and Alashetty (2022) . Form all models applied, Gradient Boosting appeared to be the best, however, this could have also been caused due to improper fitting of data since accuracy was the only model evaluation methodology relied upon. This research Rajput and Alashetty (2022) can be further improved by adding more evaluation methods and taking them into consideration rather than solely relying on accuracy of the models, which may in some cases be biased.

Overall, these studies are significant to the current project as they give an overview on disease based readmission, which can prove useful in enhancing model performance for creation of an inclusive model that takes into consideration important variables, which may otherwise be overseen, such as the link between high glucose levels and number of times the person has been rehospitalized.

2.4 Research Niche

The main focus of this project is to improve the quality of healthcare by applying data analytics with the goal of reducing the need for patients to be readmitted to the hospital. Previous works have utilized machine learning techniques such, as regression, knn, svm, Random Forest and xgboost. To ensure predictions and a deeper understanding, each model should be evaluated using different evaluation methods simultaneously. This initiative is crucial as it will help minimize hospital readmissions leading to cost friendly healthcare systems while relieving the burden on healthcare providers. In the future this study is expected to make contributions, in various fields as well as making healthcare systems more effective by reducing patient re-admissions.

3 Methodology

The research methodology outlines the approach and techniques used to analyze the dataset, on hospital re-admissions and extract insights. By explaining how data was collected, prepared and analyzed this section aims to offer an understanding of the research process.

The measures taken to ensure that the findings are valid and reliable.The sub-sections will outline the elements of the research methodology explaining the approach used to tackle the research goals and contribute to the existing knowledge, in the field of hospital re-admissions.

3.1 Environment and Tools

Python is used as the primary programming language due to its readability and extensive library support. Utilizing key libraries such as NumPy, pandas, and Matplotlib for numerical operations, data manipulation, and visualization, respectively, ensured effeciency. Kaggle's, Integrated Development Environment (IDE), was chosen for its collaborative features and pre-configured environments as the local coding environment on personal laptop did not have suffecient resources. The platform's Jupyter Notebooks allowed an interactive and document-oriented workspace, integrating code, visualizations, and explanatory text. The combination of Python and Kaggle aligns with current industry standards, ensuring transparency, reproducibility, while adhering to best practices in data science.

3.2 Data Collection and Sources

Kaggle [2] is an open source platform where datasets from various fields are uploaded and can also be used by others depending on the license among other uses. Data for this project, is obtained from dataset on kaggle Omnamahshivai (2018) and contains a common creative license ensuring proper licensing to be used for this project. It contains 18 columns including race, gender, age, time in hospital, number of lab procedures, number of procedures, number of medications, number of outpatient occurrences,number of inpatient occurrences, number of diagnosis and history of readmission. It further has unique id columns to ensure patient anonymity and 3 different diagnosis columns.Data is presented in binary form and is used as data perceived from electronic health records of relevant healthcare services.

3.3 Data Pre-processing

Data pre-processing is the most important step when preparing the coding, as it ensures that the data is clean and the format is suitable for training machine learning models as well as analysis. It involves many smaller activities including cleaning of data which is very crucial, which if skipped can lead to inaccuracies in training leading inconsistent analysis of data. While it majorly involves cleaning of data, data visualization and understanding is also critical to build understanding of how the data behaves. This extensive process can be further sub-divided.

3.3.1 Data Cleaning

The main goal in this step is to correct any issues within the data, ensure data integrity by eliminating inaccuracies and inconsistency. Here, missing data is checked, if any, applying methods like imputation or removal of incomplete entries that may not affect overall performance of models. By doing so, it ensures that the analysis is accurate by

[2] Kaggle: https://www.kaggle.com/

providing an inclusive dataset. Another reason this step is important is to eliminate any duplicate entries in the dataset, removal of redundant entries ensures analysis is not misrepresented and creation of machine learning models are trained accurately trained. All in all, the main focus of this step is to handle missing and redundant values to prep data for data exploration.

3.3.2 Data Exploration

In this step, data is scrutinized to understand its patterns and useful insights that can help in determining skewed or un-normalized data. Its main purpose is to understand data characteristics, identify patters and discover potential insights that may aid subsequent analysis. Statistical and visualization techniques such as bar plots and pie charts used to visualize categorical values, histograms used in visualization of numerical elements, line graphs are used to gain a comprehensive understanding of the data structure in the dataset as well as its distribution. Furthermore, it involves calculation of descriptive statistics such as mean, standard deviation and median to summarize numerical features.Correlation analysis allows for an insight into how closely the variables are related to one another thereby providing insights into potential dependencies and correlations within the datasets. Outlier detection done at this stage is useful to identify data points that significantly vary from the overall data pattern. It also involves the evaluation of summary statistics among different categories within the data, highlighting variations and trends within specific subgroups. This allows an understanding of the data quality which can later be useful in informed model selection.

3.3.3 Transformation

Primary objective in this phase is to convert raw dataset into a more suitable format for model creation and analysis which involves normalization and standardisation. These methods ensure that the variables with large difference do not adversely influence or affect the machine learning models thereby allowing accurate assessments of the analysis. Categorical variables such as ”yes”, ”no” can be handled in this stage by encoding them to numerical variables like ”yes” being 0 and ”no” being 1. However, since this dataset, is in binary form, this step may be skipped as it is not necessary since all the data is already in its numerical form but, this technique may be applicable in other datasets with categorical data. In this case, data transformation is mainly focused on normalization and standardization as well as handling any outliers missed in exploratory phase to ensure data is not skewed to allow for accurate model training and analysis in later stages.

3.3.4 Normalization and Standardisation

This step is necessary as it mitigates the impact of outliers by bringing the data to a common scale ensuring extreme variables have less of an impact on overall model creation and analysis. There are different methods of testing normalization including the Shapiro-Wilk test which tests the null hypothesis of a sample obtained from normally distributed population and is especially useful when testing for normality after standardisation. Another technique is using the Anderson-Darling test is similar to Shapiro-Wilk test but emphasizes more on the tails of distribution. Such methods are used to test of normalization of the data, after this, normalization techniques like min-max scaling which transforms numerical variables into a specific range usually between 0 and 1, this is done by subtracting minimum value from each data point and the result divided by the difference between maximum and minimum value. This technique eliminates the effect of the original data scale, thereby making it easy to analyze features with different measurement scale.

Additionally, Z Score normalization is a method used to transform data and involves centering the data around an average of 0 and scaling it to have a deviation of 1. This is achieved by subtracting the average from each data point and then dividing it by the standard deviation. The resulting Z Scores indicate how many standard deviations each data point is away, from the average. It is a commonly employed technique in analyses and machine learning tasks as it helps remove any biases related to scale in datasets. This allows room for comparisons and gaining deeper insights, from the transformed data. Moreover Z Score normalization enhances stability and performance of algorithms when dealing with datasets.

3.3.5 Feature Engineering

This process aims to select features most relevant in the dataset to highlight valuable information and improve the model's prediction by creating new or modifying existing features to enhance model performance. In this project, recursive feature elimination (RFE) is implemented to select the 10 most important features for training machine learning models. This is done by recursively removing the least important features based on the contribution of these values to the performance of the model,where, the model is trained while ranking features by their importance and eliminating the least important features. By doing so, it ensure that models are not overfitted thereby aiding in a better analysis of model performance and understanding.

3.4 Statistical Analysis

Descriptive statistics, including measures like mean and standard deviation, offer insights into the central tendency and variability of the data. T-tests and chi-squared tests are implemented for statistical analysis where the former is mainly applicable with numerical variables and latter is applied to categorical data.T-tests are especially beneficial in this project as they can help determine whether readmission was caused by diagnosis 1, 2 or 3 presented in the dataset. It helps in understanding where there is a real effect and a meaningful observation rather than just mere coincidence in the behaviour of data patterns.While both tests are meaningful, t-test are more preferred in this project as data is presented in numerical form.

3.5 Creating Machine Learning Models

In this project, a variety of six different machine learning techniques were chosen to predict readmission based on previous research mentioned in related work. Logistic regression, chosen for its simplicity and interpretability, is employed for binary classification. XGBoost, an ensemble learning algorithm, sequentially builds decision trees to enhance predictive accuracy. Naive Bayes, adept at handling textual data, is applied to predict readmission probabilities based on medical diagnosis. Linear regression establishes linear relationships for predicting numerical outcomes related to readmission. Random Forest, an ensemble method, handles the complex dataset by constructing multiple decision trees which highlights factors influencing readmission.

3.6 Model Evaluation

Model evaluation is important in assessing model performance, and key metrics like F1-score, accuracy, precision, and area under the curve (AUC) play integral roles. The F1-score offers a balanced measure, considering both precision and recall, making it particularly valuable for imbalanced datasets. Accuracy, a fundamental metric, calculates the ratio of correct predictions to total instances but may not be suitable for imbalanced datasets. Precision, focusing on true positives in positive predictions, is crucial when minimizing false positives is a priority. AUC, determined by the area under the ROC curve, evaluates a model's ability to distinguish between classes, with higher values indicating superior discrimination.These evaluation metrics allow for informed analysis of the models before choosing most applicable models for prediction of readmission.

3.7 Data Analytics Method Implemented

Knowledge Discovery in Databases is applied as the data analytics methodology for prediction of hospital readmission which involved systematically selecting and preprocessing comprehensive patient records. Exploration through statistical analyses and visualizations uncovered data patterns, while feature engineering refined variables. Machine learning algorithms, including logistic regression, were employed in the modeling phase, assessed using metrics like F1-score. The final model, integrated into the healthcare system, enables real-time predictions, offering valuable decision support to mitigate hospital readmissions. This concise KDD process facilitated a comprehensive understanding of healthcare data and contributed to the development of a practical predictive model.

4 Design Specification

This section discusses key elements in implementation of the project, focusing particularly on the selected programming language and key libraries used in the prediction of hospital readmission. This acts as a foundation to understanding the implementation of different machine learning models discussed at a later stage in the report.

4.1 Machine learning methods implemented

In this research, after looking into previous studies, the most common methods applied include random forest, decision trees and gradient boosting, however, it is very few studies like Zebin and Chaussalet (2019) that applied CNN to create readmission risk models which were later evaluated using different measures such as AUC, accuracy and precision. Based on the study, it is noticed that neural networks like CNN can be applied to creating predictive readmission models. Simpler machine learning models like Jiang et al. (2018) implemented the use of random forest and is similar to the application in this project as the primary focus is on prediction of re-admissions. This shows that random forest has been researched and makes a good choice for prediction in this project. Applied models can be divided into two categories where one is simple machine learning methods and the other is neural networks. The former includes statsmodel, random forests, xgboost, linear regression and decision trees. The latter includes Multi-Layer Perceptron Neural Network,

4.2 Programming language

Python is a versatile, programming language known for its simplicity and powerful libraries that can be used for data analysis and web development among other uses. The Python version used to develop predictive machine learning models is 3.10.12. It was packaged by conda-forge, a community-driven collection of conda packages, on June 23, 2023. The compiler used for building is GCC 12.3.0 in Kaggle Integrated Development Environment(IDE) . For this project, python is selected over R mainly due to its extensive libraries such as Keras and Sklearn among others. This provides extensive application of statistical application and data analysis thereby helping to create accurate machine learning models for predictions and other applications. Furthermore, these libraries offer user friendly interfaces and algorithms that are efficient.

4.3 Libraries

The two main python libraries implemented in this projected include Keras and scikit-learn(sklearn) where the former is applicable for ANN and the latter is used for simple machine learning models for data analysis. These libraries can be further divided into three categories;Keras Libraries, Sklearn libraries and General python libraries.

4.3.1 Keras libray

This library is implemented to create neural networks like simple neural network. The "Sequential" and "Dense" modules from Keras act like building blocks for creating neural networks in Python where the former creates sequential layering of neurons thereby creating a structured and systematic approach to the models design while the later plays a foundational role of defining fully connected layers in the network where each of the neurons is linked on different levels. These imports "from keras.models import Sequential ; from keras.layers import Dense" show display the versatility of Keras as a high level neural network API for the training and designing of the Artificial Neural Networks such as SNN implemented in this project. All in all this library and its imports make implementation of neural networks for predictive analysis easy to integrate into the project as it requires minimal coding where it is easy to add multiple layers in top of the models to ensure a well fitting data and avoid over fitting of the data.

4.3.2 Tensorflow

This library is specifically implemented to create a Convolutional Neural Network for prediction of readmissions as a comparison to the other neural networks created using different python libraries namely:SNN created by Keras library and Multi-Layer Perceptron Neural Network created using Sklearn library. Each of these models have a different performance which is further discussed later in the report.

4.3.3 Sklearn Library

This library imports included the features selection , ensemble, linear model, t-test, chi test, neural network ,regressor and pre-processing libraries. The neural network library from sklearn is used to create Multi-layer Perceptron neural network which is then compared to the neural network models created from the keras library, thereby allowing identification of complex patterns in the data. Different regrossor imports specifying the machine learning models are then implemented from the library to create the predictive models where in this case it includes Random Forest, Gradient Boosting and Decision Tree, however for xgboost it is imported independently under the *import xgboost as xgb*. The pre-processing library is used for standardisation of features by removing the mean and imputing it to the variance thereby ensuring a consistency in the size of the data. Furthermore, model selection is imported from this library for selection of model through splitting of training and testing data as well as cross validation. The t-test and chi2 tests are used to statistically analyse the behavior of the data where the former is used for numerical data and the latter for categorical data such as gender. Another import is yellowbrick which is used for visualization of residuals in the regression models, this helps analyse how well a certain regression model has fit the model. Finally, the feature selection library is also imported when selecting the top 10 most applicable features for model creation through recursive feature elimination that highlights the best performing variables and removes the least important depending on the performance of the model. Besides the MLP model, Dense Neural Network is also created using this library as it is more suitable with binary data as compared to CNN, which is better suited to images with data. However since CNN is applied in previous studies, it was implemented in this project to see how it will behave with the healthcare data. It is seen that there are issues when reshaping the data which hinders its evaluation thereby making it unsuitable for this project.

5 Implementation

This section discuss the steps involved in creating the predictive models for hospital re-admissions. The methodology section discussed earlier is used as a guideline from preparing the data all the way to evaluating models. This gives a better understanding of the procedure involved in creating predictive models using different libraries and machine learning algorithms. Python is used as the programming language and kaggle as the IDE.Results and performance of the models are further discussed in the evaluation method while in implementation, just a brief overview of the models s presented to highlight what ML techniques are utilised in the research.

5.1 Data Pre-processing

The project is developed in Kaggle IDE, and data is loaded directly from the dataset available on kaggle. In this stage visual representation of data allows for better understanding of data patterns and relations among the variables. To ensure data is successfully loaded, it is assigned to a dataframe and the first 6 rows of the dataset are displayed. Thereafter, the data dimension is checked which reveals that there are 59557 rows and 18 columns where none of them have missing values. To further understand the data, a countplot of gender and readmitted cases is created which shows while gender does not play a crucial role, men tend to have a slightly higher rate of readmission as compared to women1

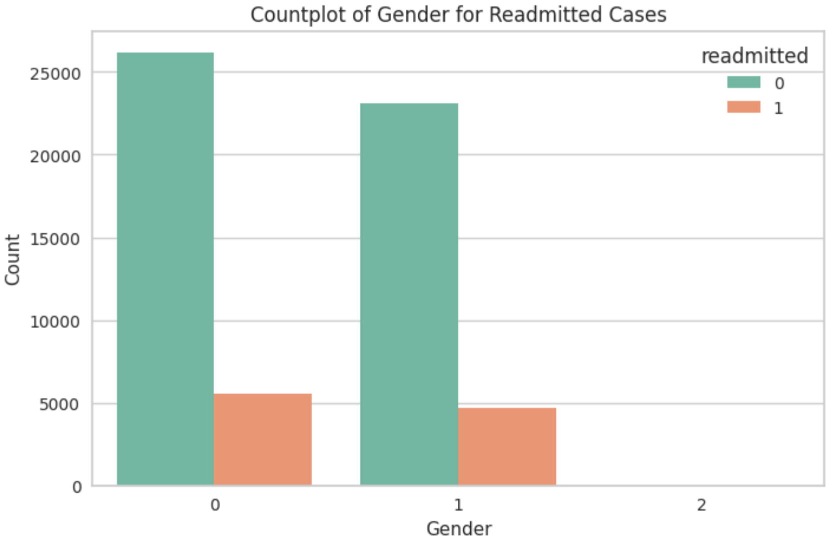

Figure 1: Counterplot of Gender and Readmitted cases

A barplot 2of time spent in hospital against readmission instances depicts that patients that spent longer in the hospitals tend to have been readmitted alot more than the ones that have spent less. This leads to assumptions that some of the patients spending alot of time in the hospital may be due to recurring or chronic illnesses such as diabetes and heart diseases among others.

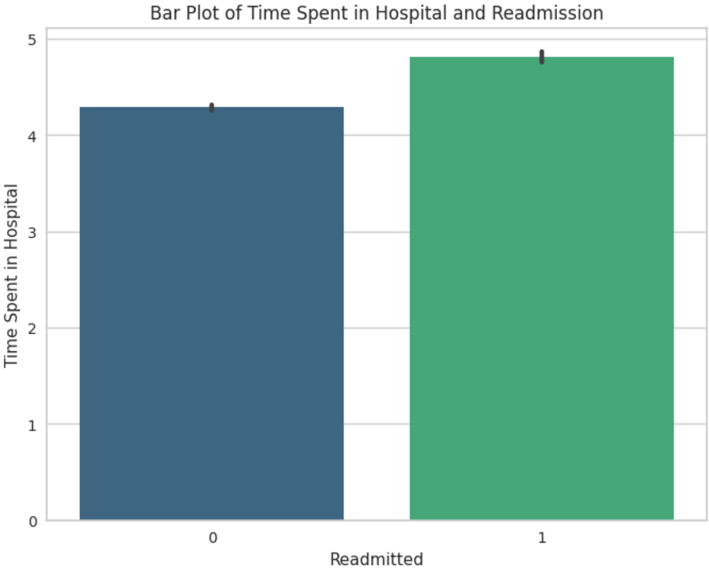

Figure 2: Barplot of Time Spent in Hospital and readmission

Additionally, checking the number of medications taken by patients and relating it to the number of re-admissions seems like a natural next step, as from the previous graph 2 it is assumed that patients that spent longer in hospitals may have chronic illnesses, so exploring the relationn between the medication and readmission rate is inghtful. From 3 it is noticeable that, patients that have shown a pattern of readmission are also taking in a fairly higher number of medication as compared to thei counterparts with less frequent visits.

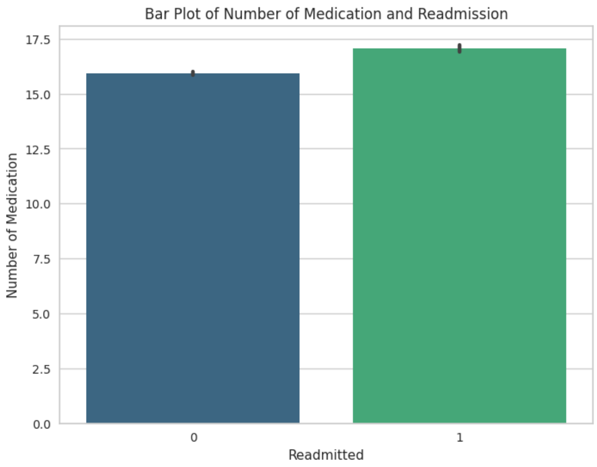

Figure 3: Barplot of number of medications and readmission

Moving on, distribution of the numerical values is checked firstly through histograms which reveal that data is not normally distributed. However, in the case of this dataset, this happens because different patients have different needs based on their disease and age. Whilst it is usual for the data to not be normally distributed due to nature of difference in healthcare data in this project data is not normalised or standerdised to ensure that all the different data points are captured accurately. Normalization test including shapiro wilk and anderson darling test are still conducted nonetheless to confirm this decision. Furthermore due to the choices in implementing machine learning models such as random forest and decision trees, model performance is not affected as these methods are not sensitive to the different data scale. The histograms below 4 show that data from the dataset is not normally distributed

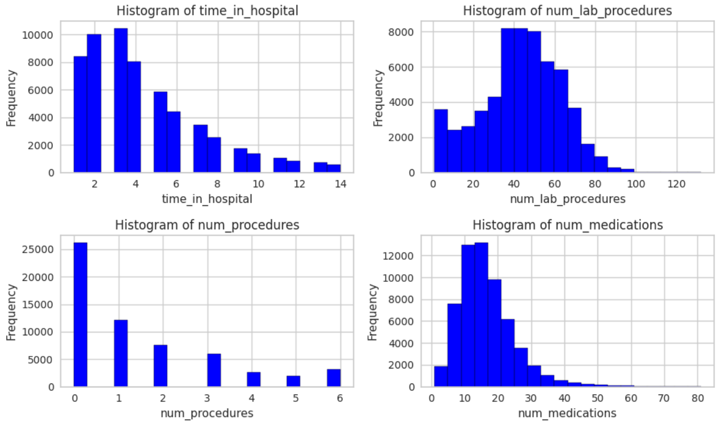

Figure 4: Histograms of numerical variables

Moving on, a heatmap 5 is created to check how closely the variables are related to each other where the values closer to 1 indicate a strong positive correlation while values near 0 suggest a weak or no correlation. The heatmap is used to identify patterns in the data and visualize the correlation matrix of variables in the dataset.It is notable that the number of medications has a high correlation with the amount of time spent in hospital.

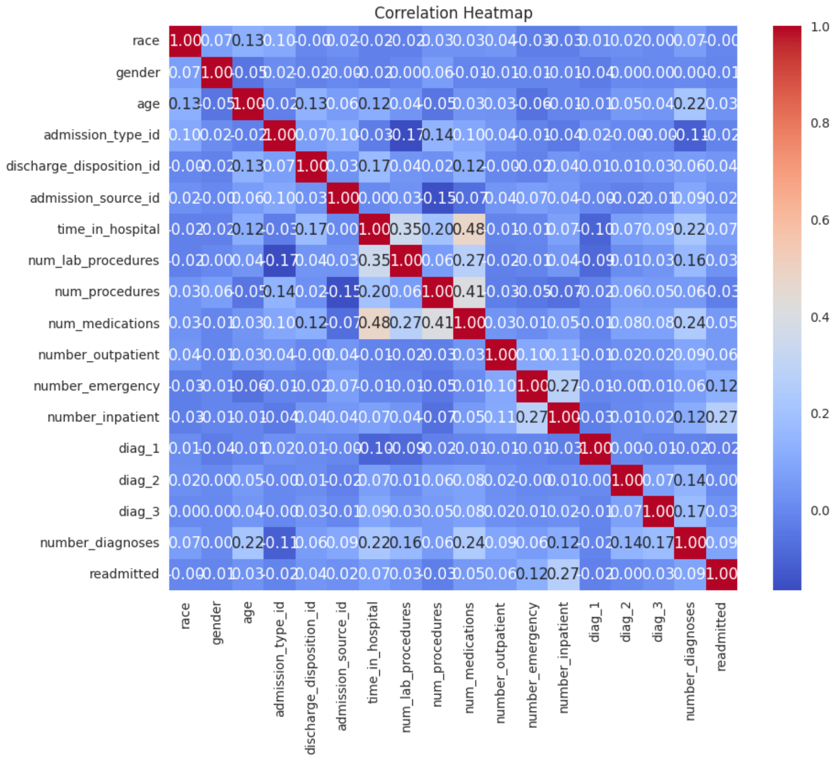

Figure 5: Heatmap

For optimisation of model performance feature engineering is conducted by the recursive feature elimination which is used to identify the most important features and removing the least important. Top 10 features are selected and different models are experimented with including linear regression, lasso regression and gradient boosting. The last model appeared to be the best for RFE as it had the lowest mean squared error of 0.128 . Gradient boosting is chosen for RFE and the 10 best variables are selected and then split into training ad testing data for model creation. 80% of the data is used for training models and 20% is used for testing.

5.2 Model Creation

Model creation for this project as mentioned before can be categorised into simple machine learning methods and neural networks. As for the former; Random Forest, Decision Trees, Linear Regression, XG Boost, Ordinary Least Squares regression are implemented while the latter, MPNN, SNN, CNN and DNN are implemented. The dense layer model performed best with an accuracy of 82.88% from the neural networks while CNN is the least applicable as the library applied for it is mainly used for image datas in data analytics. Each model is created, trained and then used for prediction. Once prediction is completed, model is evaluated using different methods including mean squared error, R-squared, F1 Score, Accuracy, AUC, Precision , Confusion matrix and class report. Different evaluation methods are applied to different models. The application and performance of the models are discussed in details in the evaluation section 6 of the project. All in all a total of 8 machine learning models are implemented, where CNN is least applicable to this project but was explored as it had been suggested in previous studies such as Zebin and Chaussalet (2019) which explored the implementation of deep recurrent machine learning models for prediction.

5.3 Statistical Analysis

Chi squared test is done to check the association between whether a person has been readmitted and age. Based on the test, there is a significant association unlike that between gender of the person and readmission. T test is also conducted which indicated that the expected mean was not so different from actual mean suggesting that there is no solid reason to reject the null hypothesis. The target variable used in this research is ”readmitted” from the dataset.

6 Evaluation

This section discusses the output of the model performance implemented to predict hospital re-admissions as evaluation metrics are applied to understand performance of the model. 8 different models are evaluated and the best performing is discussed to better highlight the implementation of machine learning in healthcare.

6.1 Ordinary Least Squares

This model suggests that age, number of emergency admissions and number of times a person has been inpatient is positively in association of readmission. The R-squared is 0.0080 indicating that roughly 8% of the variability in readmitted is explained by the independent variables available in the model thus suggesting a poor fitting of the model. However after analysing the prediction, the model has a high accuracy of 0.83 which may be due to imbalance in the classes since it ability to predict the positive classes is limited as suggested by the low recall of 0.99 and f1-score of 0.91. It further has a false positive of 68 and only true positive of 145 concluding that this model is not a good fitting model for predictions despite its high accuracy. In contrast, the precision is high for class1 meaning that when the model does predict, it is correct in most cases while the low recall points towards an imbalances in the classes.

6.2 Linear regression model

This model has an accuracy of 83.10% and precision of 68.08% which if used the only evaluation metrics make it seem like a good predictive model. Upon further investigation, it is noted that the f1 score is 12.6 which is low suggesting that the models ability to balance recall and precision is low. The AUC of 0.53 points to a limited ability to distinguish between classes and false positive of 68 from the confusion matrix also does not indicate to a well performing model. The low recall for class 1 based on the class report indicated that the model might not be capturing a significant portion of the positive instances. To better understand the models performance, a residual plot is used6 .While the plot doesnot suggest overfitting, it does indicate the presence of outliers which affect the models performance in prediction thereby explaining the low f1 score and high accuracy caused by class imbalances. The spread of residuals systematically increases suggesting there isnt constant variance of residuals highlighting heteroscedacity.Based on this information, linear regression doesnt not seem to be the best model for this dataset for predictions.

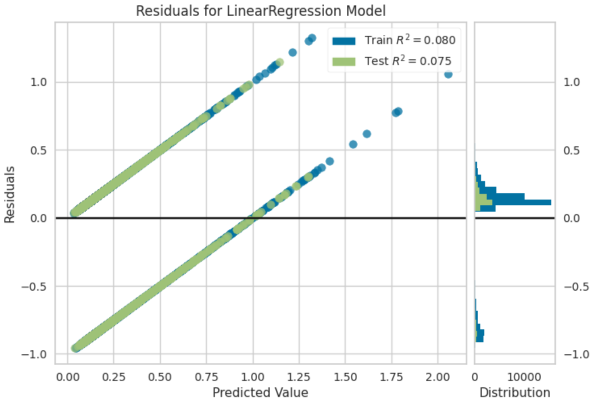

Figure 6: Linear Regression Residual Plot

6.3 Random Forest

This model has an accuracy of 82.68% and precision of 52.83 suggesting that the model was able to predict the values without being affected by any class imbalances. The AUC of 0.549 indicates a better discrimination among classes suggesting that the model performance is moderately good with no outliers as shown in the figure below 7. It had an MSE of 0.1319 and MAE of 0.2554 which are both considerd low once again indicating the model performed well. The f1 score is 19.64 which s higher compared to the first two model with true negatives of 9598 and true positives of 252 once again the false negative and positives are lower as compared to earlier models. This so far has been an improvement and better performing model.

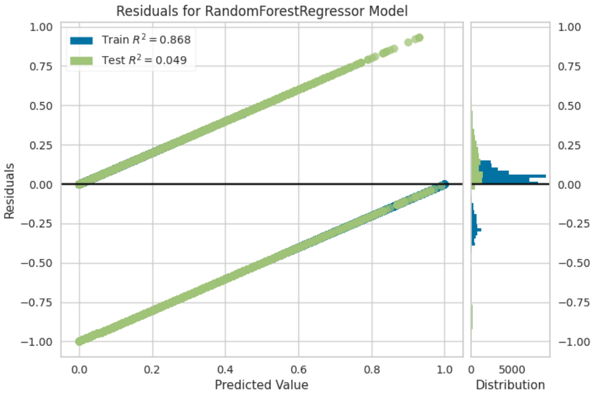

Figure 7: Random Forest Residual Plot

6.4 XG Boost

This model indicates a good performance of predicting readmissions, with an accuracy of 82.69%. The Mean Squared Error of 0.1319 and Mean Absolute Error of 0.2554 suggest a relatively low level of error in the predictions. The R-squared value of 0.0882 highlights that the model explains a small portion of the variance in the dependent variable. The F1 Score, is at 19.64, displaying moderate effectiveness in correctly identifying positive instances whilst minimizing false positives. The Precision for class 1 is 52.83%, indicating the proportion of correctly identified positive instances among all predicted positives. The

AUC value of 0.5489 suggests a moderate ability of the model to distinguish between positive and negative cases. The confusion matrix suggests that the model correctly predicts a 9598 negative cases, but does not do so well with false negatives of 1837 and has a limited ability to identify positive cases 252. While the model seems to be performing fairly based on the evaluation metrics, the r-squared for the test data is higher than training r-squared with the former being 88% and latter only 20.5% which indicates overfitting of the model.

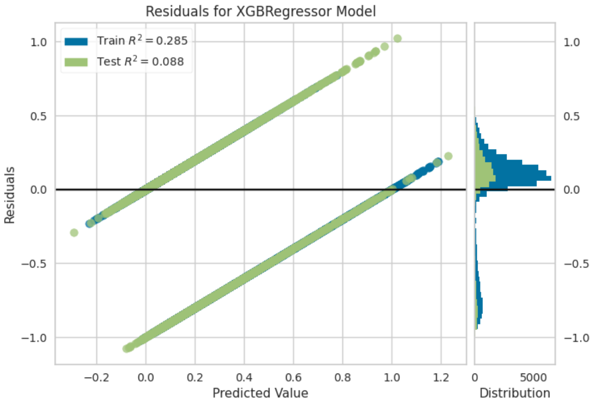

Figure 8: Residual plot for XG Boost model

6.5 Decision Trees

This model displayed a relatively higher f1 score of 27.22 , lower accuracy of 73.48 and r-squared of -0.8335 indicating that model does not adequately capture the variance in the dependent variable. The precision is 26.24% which all together indicates that this model is not a good fit for the data as it also produced a false positives of 1661 and false negatives of 1498 in the confusion matrix,, contributing to the model's classification challenges. The accuracy is the lowest in comparison to all the earlier discussed models and presents not to be suitable model for the data.

6.6 Multi-Layer Perceptron Neural Network

This model had a high MSE of 0.3058 and very low r-squared of -1.1149, where the negative r-squared suggests that the model did not fit data and had potentially worse performance than a naive model that predicts the mean of the target variable. This also means that the neural network is not effectively identifying the patterns in data but rather training on the noise in the data.

6.7 Simple Neural Network

This model performed with an MSE of 0.1368 the low R-squared value of 0.0539 suggests that the model shows only a small portion of the variance in the target variable thus suggest moderate levels of error in predictions. The F1 Score is 18.67, indicating moderate effectiveness in identifying positive instances while minimizing false positives. The accuracy of 82.59% implying that the model correctly classifies the majority of instances. The AUC is 0.5456, implying an ability of the model to distinguish between positive and negative cases. Precision at 51.63% while the confusion matrix reveals 9600 correctly predicted negative cases but highlights challenges with false negatives of 1851.To better understand the performance, a graph is implemented which shows that the training loss which shows that the MSE of training loss decreases while validation loss mse increases thus indicating overfitting

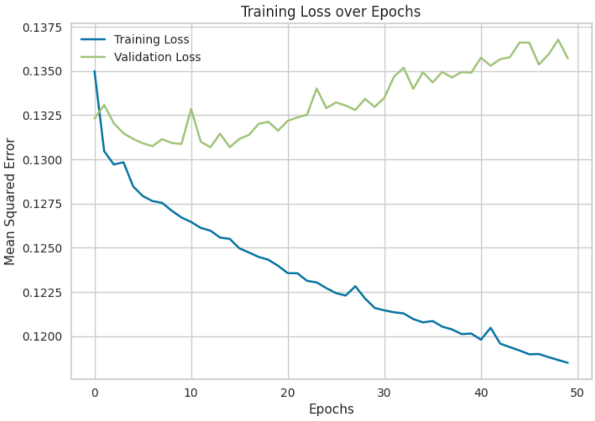

Figure 9: Simple Neural Network performance

6.8 Dense Neural Network

This model is trained over 10 epochs and is evaluated differently from all the other models. It is noted that the accuracy improves from 77.95% to 80.79% over the epochs. The training loss significantly reduces from 0.7471 to 0.5386. It also achieved the highest accuracy of 82.99% with a minimum loss of 0.4665. After this the model is then evaluated once again to check its performance on a different test set which yield an accuracy of 82.88% with a test loss of 0.5230. This model appears to have the best performance overall in comparison to the other neural networks whether the data is not overfitting or underfitting as displayed from the graphs below10

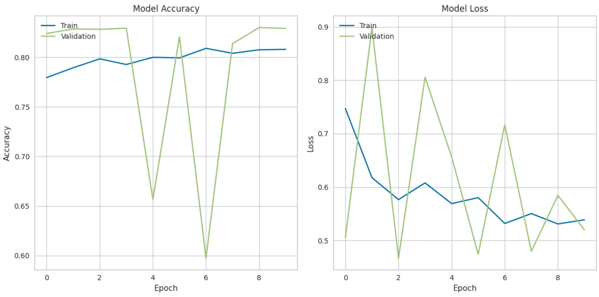

Figure 10: Dense Neural Network performance

6.9 Discussion

The linear regression model showed ability in explaining the differences, in the target variable with an R value of 0.0755. Precision of 68.08% is demonstrated by the model to correctly identifying patients. The decision tree model performed better with an accuracy of 73.48%, however the linear regression model was sensitive to the scale of data and was not the best fitting model for this project. If relied only on accuracy it would make a compelling case of 83.1%. Random forest model achieved an accuracy of 82.69%

outperforming other models as it was able to not only have a high accuracy but also and maintained a good balance between precision and recall. It particularly excelled in predicting cases as shown by the confusion matrix. XGBoost model had the same accuracy rate of 82.69%, however, its precision, for positive instances was slightly lower suggesting a potentially better performance. The neural network models, displayed varying degrees of performance. The simple neural network achieved an accuracy of 82.59% but displayed limitations in terms of precision and recall for positive instances. The Multi-Layer Perceptron Neural Network demonstrated overfitting concerns, as evidenced by a relatively low R-squared value of -1.11 and a high test loss of 0.3058. Of all the neural network models, the dense network model outperformed the rest with an accuracy of 82.88% and minimal test loss of 0.5. All in all from the exploratory data, it was noticed that 2 time spent in hospitals can affect wether a person will be readmitted. To predict cases of readmission the dense neural network and the xgboost displayed to be the best performing and if implemented will lead to higher accuracy where the former has an edge over the latter.

The dense neural network can be used to accurately predict the instances of readmission among patients as it had a high accuracy and low test loss meaning that it can be used in the industry. The main objective of this project is to aid healthcare facilitators in recognizing patients at risk before its too late and readmission is one of the main occurences among various patients. By applying DNN for prediction, it can help minimise the costs of healthcare, timely identify at-risk patients, minimse costs and DNN gives an insight into what features are most relevant to prediction of readmissions.

7 Conclusion and Future Work

In conclusion, incorporating machine learning in healthcare to predict hospital readmission can significantly lower the financial impact it has not only on healthcare facilities but also patients whilst reducing the burden on healthcare workers. This research was effective in identifying what machine learning models work best when predicting the cases of readmission where the dense neural network is best applicable and the CNN being least effective due to binary nature of the dataset.In the future applying a base model to the models can be beneficial for cross validation and further enhancement of the performances.Additionally, analysts can remove outliers after model creation and retrain the model to compare for differences between both as well use standardized data. A scope for improvement in this case would be cross validation of models to pre existing models.

References

Baig, M. M., Hua, N., Zhang, E., Robinson, R., Armstrong, D., Whittaker, R., Robinson, T., Mirza, F. and Ullah, E. (2019). Machine learning-based risk of hospital readmissions: Predicting acute readmissions within 30 days of discharge, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2178–2181.

Jiang, S., Chin, K.-S., Qu, G. and Tsui, K. L. (2018). An integrated machine learning framework for hospital readmission prediction, *Knowledge-Based Systems* **146**: 73–90. **URL:** *https://www.sciencedirect.com/science/article/pii/S0950705118300443*

Jung, H., Park, H. W., Kim, Y. and Hwangbo, Y. (2023). Machine learning-based prediction for 30-day unplanned readmission in all-types cancer patients, *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 132–135.

Lo, Y.-T., Liao, J. C., Chen, M.-H., Chang, C.-M. and Li, C.-T. (2021). Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms, *BMC Medical Informatics and Decision Making* **21**(1): 288.
URL: *https://doi.org/10.1186/s12911-021-01639-y*

Maddipatla, R. M., Hadzikadic, M., Misra, D. P. and Yao, L. (2015). 30 day hospital readmission analysis, *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2922–2924.

Omnamahshivai (2018). Hospital Readmissions Binary, `https://www.kaggle.com/datasets/omnamahshivai/dataset-hospital-readmissions-binary`.

Rajput, G. G. and Alashetty, A. (2022). A machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique, *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pp. 173–177.

Ramírez, J. C. and Herrera, D. (2019). Prediction of diabetic patient readmission using machine learning, *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pp. 1–4.

Teo, K., Wee Lai, K., Wai Yong, C., Pingguan-Murphy, B., Huang Chuah, J. and Tee, C. A. T. (2020). Prediction of hospital readmission combining rule-based and machine learning model, *2020 International Computer Symposium (ICS)*, pp. 352–355.

Wang, S. and Zhu, X. (2022). Predictive modeling of hospital readmission: Challenges and solutions, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19**(5): 2975–2995.

Zebin, T. and Chaussalet, T. J. (2019). Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records, *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–5.

Zhu, Lou, Z., Zhou, J., Ballester, N., Kong, N. and Parikh, P., K. (2015). Predicting 30-day hospital readmission with publicly available administrative database. a conditional logistic regression modeling approach, *Methods Inf Med* **54**(6): 560–7. Epub 2015 Nov 9. PMID: 26548400.