# 

FEATURE ACCOUNTABILITY FOR 30-DAY HOSPITAL READMISSION PREDICTION FOR DIABETES PATIENTS

SHRIDEVI BELLATTI

Mid Thesis Report

MARCH 2021

**TABLE OF CONTENTS**

[ABSTRACT i](#_Toc67140614)

[LIST OF TABLES v](#_Toc67140615)

[LIST OF FIGURES v](#_Toc67140616)

[LIST OF ABBREVATIONS vi](#_Toc67140617)

[Chapter 1 INTRODUCTION 1](#_Toc67140618)

[1.1 Background of the Study 1](#_Toc67140619)

[1.2 Problem Statement 4](#_Toc67140620)

[1.3 Aim and Objectives 6](#_Toc67140621)

[1.4 Research questions 6](#_Toc67140622)

[1.5 Scope of the Study 6](#_Toc67140623)

[1.6 Significance of the Study 7](#_Toc67140624)

[1.7 Structure of the Study 7](#_Toc67140625)

[Chapter 2 LITERATURE REVIEW 8](#_Toc67140626)

[2.1 Introduction 8](#_Toc67140627)

[2.2 Risk factors associated with the readmission 8](#_Toc67140628)

[2.2.1 Demographic characteristics 8](#_Toc67140629)

[2.2.2 Comorbidities 9](#_Toc67140630)

[2.2.3 Length of Stay 10](#_Toc67140631)

[2.2.4 Insulin Therapy 10](#_Toc67140632)

[2.2.5 Ways of Discharge 10](#_Toc67140633)

[2.3 Applications of Interpretability in healthcare 10](#_Toc67140634)

[2.4 Interpretable Machine learning Models 11](#_Toc67140635)

[2.5 Deep learning Models for Prediction 16](#_Toc67140636)

[2.6 Black Box Model Explanation Problems 20](#_Toc67140637)

[2.6.1 Global Interpretation 20](#_Toc67140638)

[2.6.2 Local Interpretation 21](#_Toc67140639)

[2.6.3 Transparent Box Design 24](#_Toc67140640)

[2.7 Class Imbalance Techniques 25](#_Toc67140641)

[2.8 Summary 26](#_Toc67140642)

[2.9 Discussion 27](#_Toc67140643)

[Chapter 3 RESEARCH METHODOLOGY 28](#_Toc67140644)

[3.1 Introduction 28](#_Toc67140645)

[3.2 Dataset Description 28](#_Toc67140646)

[3.3 Data Pre-processing 28](#_Toc67140647)

[3.3.1 Missing Value Imputation 29](#_Toc67140648)

[3.3.2 Data Cleaning 29](#_Toc67140649)

[3.3.3 Data reduction 30](#_Toc67140650)

[3.3.4 Class Imbalance 31](#_Toc67140651)

[3.3.5 Feature Engineering 31](#_Toc67140652)

[3.4 Modelling 34](#_Toc67140653)

[3.5 Model Agnostic Interpretable Methods 37](#_Toc67140654)

[3.6 Proposed Method 41](#_Toc67140655)

[3.7 Evaluation 42](#_Toc67140656)

[3.8 Summary 44](#_Toc67140657)

[REFERENCES 44](#_Toc67140658)

[APPENDIX A: RESEARCH PLAN 51](#_Toc67140659)

[APPENDIX B: RESEARCH PROPOSAL 51](#_Toc67140660)

# ABSTRACT

A Hospital readmission is where a patient gets readmitted to the hospital post-discharge within a specified interval of time. Readmission is the key measure of the quality of treatment provided by the health centres to the patients during hospitalization and has proven to be immensely expensive. According to the Affordable care Act (ACT) 2010, hospitals are penalised with low reimbursement for the patients admitted within 30-days of discharge and the cost of readmission is estimated to be around $25 billion yearly in U.S hospitals. Studies have shown that patients with diabetes are more susceptible to readmission within 30-days of discharge. So, predicting the readmission is beneficial for patients and healthcare centres. Current practice of identifying diabetes patients for readmission is subjective, where the clinician will assess what should be the appropriate care given to the patients. But, research have shown that with the help of predictive modelling it is slightly better which learns the records of various patients. Considering this, the main goal of this study is to build a predictive classification model and explain the predictions by identifying feature importance. This study benchmarks the existing machine learning models and proposes MultiLayer perceptron (MLP) with extensive pre-processing and transformation which includes feature selection using Random Forest feature selection, feature engineering, balancing the data using SMOTE to handle the noisy, inconsistent and imbalanced dataset of diabetes patients from 130 US Hospitals for the year 1999-2008. The proposed model will predict the readmission of diabetes patients within 30-days and derive meaningful insights of the attributes contributing towards the readmission by applying LIME, SHAP and DeepLIFT algorithms. Identifying the readmission of hospitalised diabetes patients, uncovering the features, helps healthcare centres to improvise the treatment and reduce the hospital expenditure. Model performance is evaluated using Accuracy, precision, recall, Area Under Receiver Operating characteristics (AUROC) and the interpretability of the model is evaluated, comparing the important features with the existing machine learning suggested features.

# LIST OF TABLES

[Table 2.1 Summary of machine learning models 15](#_Toc67125373)

[Table 2.2 Summary of deep learning models 20](#_Toc67125374)

[Table 3.1 Newly added features 32](#_Toc67125375)

# LIST OF FIGURES

[Figure 2.1 (a): De (Zhou and Feng, 2017) Figure 2.1 (b) Single forest view De (Zhou and Feng, 2017) 15](#_Toc67132515)

[Figure 2.2 Feed-forward architecture 17](#_Toc67132516)

[Figure 2.3 (a) Explanation of SHAP (b) Explanation of LIME (Alvarez-Melis and Jaakkola, n.d.) 23](#_Toc67132517)

[Figure 2.4 MNIST digits with maximum perturbations and their explanations (Alvarez-Melis and Jaakkola, n.d.) 23](#_Toc67132518)

[Figure 2.5 ﻿An interpretable approach, decision sets (left), and a decision list (right) learned from the same medical diagnosis dataset. Decision sets (left) are set of if-then rules and are independent of each other. While Decision lists (right) are if-then-else rules, where the rule depend on the decision above it. For, humans readability, decision sets are more comprehensible. 25](#_Toc67132519)

[Figure 2.6 Different methods of addressing class imbalance 27](#_Toc67132520)

[Figure 3.1Simple Random Forest Classifier 35](#_Toc67132521)

[Figure 3.2 Interaction of features with the outcome 35](#_Toc67132522)

[Figure 3.3 Single layer perceptron, multilayer perceptron 37](#_Toc67132523)

[Figure 3.4 Explanation pipeline 38](#_Toc67132524)

[Figure 3.5 Steps of LIME [https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe] 40](#_Toc67132525)

[Figure 3.6 Backpropagation approach, yellow=input, green=output 41](#_Toc67132526)

[Figure 3.7 saturation problems faced by perturbations and gradient based methods 41](#_Toc67132527)

[Figure 3.8 Architecture of Explainable Model 42](#_Toc67132528)

# 

# LIST OF ABBREVATIONS

ACT Affordable Care Act

AI Artificial Intelligence

ANN Artificial Neural Network

AUC Area Under Curve

AUROC Area Under Receiver Operating Characteristics

AXAI Adversarial Explainable Artificial Intelligence

BDL Bayesian Decision List

CBA Classification Based on Association

CI Confidence Interval

CHF Congestive Heart Failure

CMS Centres of Medicare and Medicaid Services

CNN Convolution Neural Network

DBFN Deep Belief Network

DeepLIFT Deep Learning Important FeaTures

DEERT Diabetes Early Readmission Risk Indicator

DT Decision Trees

FN False Negative

FP False Positive

GAMs Generalized Additive Model

GB Gradient Boosting

GBM Gradient Boosting Machine

GLM General Linear Model regression

HRRP Hospital Readmission Risk Program

IDS Interpretable Decision sets

IR Imbalance Ratio

KNN K-Nearest Neighbor

LACE Length of stay, acuity of admission, Charlson comorbidity index and Emergency visit

LASSO Least Absolute Shrinkage and Selection Operator

LIME Locally Interpretable Model Explanation

LORE Local Rule-based Explanations

LR Logistic Regression

LSTM Long and Short Term Memory

MLP Multi-Layer Perceptron

MWBS Modified Weight Boosting-Stacking

NB Naïve Bayes

OR Odds Ratio

PIDD Pima India Diabetes Dataset

PReLU Parametric ReLU

ReLU Rectified Linear Unit

RBF Radial Base Function

RF Random Forest

SB Sparse Bayesian

SHAP Shapley Adaptive explanation

SGD Stochastic Gradient Descent

SMOTE Synthetic Minority Oversampling technique

SVM Support Vector machine

TN True Negative

TP True Positive

UCI University of California Irvine

VFDT Very Fast Decision Tree

XGBoost eXtreme Gradient Boosting

# 

# INTRODUCTION

## **Background of the Study**

Diabetes has become an Epidemic in today’s world with moving towards urbanization and changing lifestyle of people around the world. The global diabetes prevalence was increased to 9.3% in 2019 and estimated to increase to 11% by 2045. Diabetes can be of different types, type 1, type 2 and gestational diabetes. Type-1 diabetes is due to the body not producing enough insulin and is seen mostly with children’s. Type-2 diabetes is when the body becomes resistant to insulin produced in the body, this is seen in population above 45 years. Gestational diabetes occurs during the pregnancy. Type-2 diabetes is most chronic among all these types. Diabetes can lead to various health issues like, kidney failure, cardiovascular diseases, hearing loss, brain damages, skin infections, sexual disabilities and hence treating diabetes is most important. At the same time, with poor quality of treatment there are high chances of patients getting readmitted after the discharge in < 30 days, >30 days, >60 days or 1 year due to same cause or different causes or in some cases leads to mortality. This calls for a better hospital operational efficiency and better care of Diabetes patients. Readmission is one of measure of hospitals operational efficiency. According to Affordable Care Act (ACT), Centres of Medicare and Medicaid Services (CMS) has enforced policy on the hospitals to measure the readmissions, wherein hospitals enrolled under this policy will be penalized by reducing the reimbursement payments to hospitals with high 30-day readmission for targeted conditions. Although Diabetes is not among the six targeted conditions identified by CMS, diabetes readmission is high compared to any other diseases.

Various measures have been implemented to reduce the unplanned readmission of inpatients, like identifying the patients at high risk before discharge, improving the Transition care model, proper handoffs between caregivers. From 2010 to 2015, through various patient intervention efforts, hospitals have reduced readmissions by an estimated total of 565,000 patients. Even with this good work, there’s still much room for improvement: the federal government estimates that readmissions cost $26 billion per year, and that 65% of those readmissions are avoidable (Graham et al., 2019). So, predicting the unplanned 30-days readmission in Diabetes patients is of great importance.

Identifying the patients with high risk of readmission by health practitioners, looking at history of patients is time consuming and requires a domain expertise. Now, with the advancement in technology of, “Big Data Era”, hospitals are able to produce huge amount of data. Predictors include those quantifying clinical treatment, demographic and maternal factors. With this, medical data is heterogeneous, sparse, inconsistent, noisy and imbalanced. Gaining insights into all this available information and building a highly predictive model is quite challenging (Kansagara et al., 2011; Artetxe et al., 2018). Deep Learning helps to learn these patterns consistently from this large data by developing a good decision support system which helps the practitioners to make decision in faster way.

Many machine learning models have been developed to predict the readmission of diabetes patients within 30-days. Models vary from simple Logistic Regression, Decision Trees, Random forest which are popular classification algorithm to more complex models like Deepforest, ensemble models and Deep learning. Compared to machine learning models, deep learning looks promising with accuracy in predicting the readmission. But, making predictions alone is not sufficient, it is also important to provide evidence to the decision made. It is important for the patients and the decision makers to understand “why” the model made the decision, what are the factors influencing the decision, how can you prevent readmission. Deep learning model with its high ability to work with large data and learn the features iteratively through its multiple layer architecture can increase the predictive accuracy of the model, but does not provide interpretable ability. There is an inherent tension between machine learning performance (predictive accuracy) and explainability. Often the best-performing methods (e.g., deep learning) are the least transparent, and the ones providing a clear explanation (e.g., decision trees) are less accurate (Bologna and Hayashi, 2017).

Model fairness and interpretability is required for data-scientists, researchers, developers to bring in value for their findings and support the accuracy. Interpretability is also important to debug machine learning models, understand the model bias and make informed decisions on how to improve them. Model Interpretability is important in critical fields like medical, financial domains where not just the model prediction becomes important but also the need to understand why the model made this prediction is important. For example, by looking at the size of the cancer cell, if the models predicts that the patient is likely to have cancer, it’s a human tendency to understand what factors are contributing to the decision. Models like Logistic Regression, Linear Regression, decision trees exhibit intrinsic property of interpretability while other complex models like deep neural networks don’t, and are called black box models. These requires explicit interpretability techniques to be used which are known as post-hoc interpretations. However, machine learning models are still of important because of its superior popularity of model interpretability and its ease of use, and the black box nature of neural network takes away the model interpretability as explained by the authors in (Dreiseitl and Ohno-Machado, 2002). Due to some of the commonly seen challenges with deep learning models like intense-computation, increase in parameters, interpretability and incorporate domain knowledge, clinical applications of deep learning models is still not being considered (Miotto et al., 2017; Rudin, 2019). According to European Union’s General Data Protection and Regulation law and “right to explanation”, a user can ask the reasons for the decision made against them and hence models without interpretability are not considered. There is lot of research conducted in these areas and open the doors to use deep-learning in healthcare, especially with the interpretability models being developed (Ribeiro et al., 2016). (Swartout, 1983) emphasized: Trust in a system is developed not only by the quality of the results but also by clear description of how they were derived, a user, should be able to ask for a description of the methods employed and the reasons for employing them.

Class Imbalance is another most commonly seen challenge in medical data where large number of instances belong to one class called majority class and the other is the minority class with fewer number of instances which are the targets of the model. This always lead to biased decision towards majority class. Balancing the data is an important factor to consider while model building. SMOTE is the common and most suggested methods by various researchers.

To build an explainable AI model in medical domain, knowledge base and neural concepts should work together to get the interpretability and high-precision model to support the medical experts (Holzinger et al., n.d.). So, in this study we use a large medical dataset, and use medical evidence in feature engineering, balance the class, discuss the model functionality and evaluate the performance with accuracy, precision and recall obtained on different models. This study suggests developing holistic and meaningful interpretable architectures to bridge deep learning models and human interpretability. Models like Random Forest, Multi-layer perceptron along with post-hoc interpretability methods, SHAP, DeepLIFT, LIME are used on 130-US hospitals for years 1999-2008 Data Set, (UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set, 2021) to predict the readmission of diabetes patients within 30-days of discharge and identify the risk factors explaining the prediction. Effective risk prediction models and better understanding the risk factors will enable the development of discharge planning which is more effective, targeted treatment, better patient education interventions and improved risk prediction tools for high-risk patients. Some of the risk factors identified after a comprehensive review of literature are “gender”, “race”, “age”, “length of stay”, “discharge disposition”, “comorbidity”, “insurance type” and “insulin therapy”.

## **Problem Statement**

Diabetes is a growing burden in US population. According to the National Diabetes statistics report 2020, 10.5% (34.2 million people) of US population has Diabetes (National Diabetes Statistics Report, 2020 | CDC, 2021). Prolonged diabetes damages most of the body tissues leading to several complications like cardiovascular diseases, kidney failure, poor eyesight, hearing loss. Due to high prevalence of diabetes and its complications, it has been a common comorbid condition among the hospitalized patients. This leads to frequent procedures and patients with diabetes, experience increased the length of stay (Collins et al., 2017), leading to increased cost of treatment to patients, health care providers and policy makers. Compared to patients without diabetes, patients with diabetes were more likely to get readmitted with some of the comorbidities like Heart failure, Depression, Respiratory Diseases, Renal Diseases, Hypertension. (Eby et al., 2015; Rubin et al., 2016, 2018a; Caughey et al., 2017; Enomoto et al., 2017). In recent studies, 30-day readmission rate for hospitalized patients with diabetes was estimated to be (14.4% - 22.7%)(Robbins and Webb, 2006), much higher than total hospitalized patients (8.5% - 13.5%) (Friedman et al., 2008). In 2012, CMS has launched a program, HRRP, to reduce rehospitalization by penalizing hospitals with high readmission rates for specific diseases. In addition, HRRP has fueled in intense interest in identifying the predictors and identifying the causes of readmission, however due to poor performance of models, lack of domain knowledge and wide range of predictors, hospitals lack the ability to identify high risk patients. With a wide range of predictors available it becomes difficult for the health practitioners to diagnose and analyze the disease quickly. No doubt, data collected from patient and expert’s opinion are important factors to build an intelligent model in healthcare, it’s not easy and trustable due to possible errors that can be induced by inexperience experts. To increase accuracy and avoid the misclassification cost, researchers have developed cost-sensitive models (Yu and Xie, 2020). To help experts make decisions in short time, classification models with high performance and good understanding can be used to develop a decision-support-system.

Several models have been developed around UCI-130 hospital Diabetes dataset using traditional machine learning models (Tamin and Iswari, 2017; Reid, 2019) and deep learning models. Deep learning models are found to provide promising results, with the increase in accuracy but compromising on interpretability (Hammoudeh et al., 2018; Goudjerkan and Jayabalan, 2019; Hu et al., 2019; Reddy et al., 2020). To build a model in healthcare, interpretability of the model is of great importance, (Golmohammadi and Radnia, 2016; Yang et al., 2016) rejected to use neural networks only because it was lacking to explain the decisions of the model. Providing explanations by identifying the risk factors causing the readmission helps to achieve targeted treatment, better discharge plans and hence reduce the readmission.

With the advancement in data gathering with the available technology, “Big Data”, hospitals are able to provide large amount of data. However, medical data is noisy, heterogeneous, sparse and imbalanced. Class imbalance is one of the major problems in medical data, due to which decision of the model is biased towards the majority class (Artetxe et al., 2018). This misclassification can be a serious issue in high-stake decisions like healthcare, crime, finance where the life and death of a person is involved. Understanding the characteristics of the data and using the correct balancing technique is very important. Balancing not always improves the performance of the model , but can also reduce drastically (Lu et al., 2019).

Traditional machine learning algorithm with intrinsic ability to explain the decision and provide risk factors, logistic regression, decision trees are most popular classification algorithm used across any of the classification task. Deep learning is also used in the several classification task, but with the scarcity of interpretability. Several researchers tried to explain the black-box model, (Bhuvan et al., 2016) using rule based methods, ExplainD (Poulin et al., 2006) , LIME (Ribeiro et al., 2016). But these are not applied on the UCI-130 hospitals diabetes dataset. As this study focuses in explaining the high performance models using this dataset, the factors contributing to readmission suggested by others authors are compared with the factors predicted by any of these post-hoc methods.

## **Aim and Objectives**

The aim of this study is to propose an Interpretable model to predict the readmission of Diabetes patients in less than 30-days of discharge. Interpretability achieved by identifying the risk factors helps health practitioners to mitigate the risks quickly, providing targeted treatment and reduce the readmission cost to both patients and hospitals.

Following are the research objectives planned towards achieving the aim of the study:

* To suggest a suitable balancing technique to be applied on a class imbalance dataset.
* To compare between the decisions of predictive models with high interpretability to identify the influencing factors of readmission in diabetes patient.
* To evaluate the performance of predictive models based on the model interpretability.

## **Research questions**

In this research project, there are two questions that are focused based on the literature survey conducted. Many researchers proved neural networks provide better performance with the exception of model interpretability.

* Can the machine learning model, or a deep learning model be used by healthcare providers to predict the readmission of diabetes patient within 30-days of the discharge?
* Can the model performance and interpretability (black box) be achieved using model agnostic interpretation methods (SHAP, LIME, DeepLIFT) to make the healthcare professionals trust the prediction of the model?

## **Scope of the Study**

The dataset used in this study is of US population, and the prediction of readmission of diabetes patients depends on certain demographic information of the patients, so the scope of this study is limited to diabetes patients within US. And the dataset is highly class imbalanced, and not too large in size to reduce the size of the data with under-sampling, in order to leverage the information and retain data acceptable for deep learning models, SMOTE balancing technique is used.

With the recent growth in Artificial Intelligence, and understanding the body of literature, complex models are developed to gain high accuracy, compromising interpretability of the model. Along with it, research has grown towards developing new methods to interpret these complex models. DeepLIFT, SHAP has shown good explain-ability compared to LIME, Permutation feature importance and other interpretable methods. So, this study will perform analysis on how to leverage these methods with black-box models and make the models more acceptable and reliable.

## **Significance of the Study**

With the massive growth in digitization, large amount of clinical data is collected for the researchers to analyse and predict the readmission of hospitalized diabetes patients within a period of 30-days. Machine learning models and deep learning models are developed to make this prediction and achieve high accuracy models. But as we move towards high accuracy using the complex models, the model interpretation is lost. When there is a need to deploy the model, understanding why the model makes certain prediction becomes crucial in the field of medicine. The main purpose of this research is to develop a Neural network model with high accuracy and use model agnostic interpretable methods, LIME, SHAP and DeepLIFT to interpret the model. This study facilitates healthcare professionals to predict the early readmission of diabetes patients and improve the treatment of the patients by understanding the risk factors contributing to the predictions, establishing trust and confidence on the predictions of the model.

## **Structure of the Study**

This research is explained well by formulating the study in different sections. Chapter 1, focuses on the “Introduction of study”, explaining the domain, background of the study, problems seen with other researchers, aim and objectives, scope and significance of the study. Chapter 2, is the “Literature Review” which explains the review on the readmission prediction, showing the various models, pre-processing techniques, class imbalance techniques, evaluation methods studied by other researchers on the same dataset or an different dataset. Chapter 3, “Research Methodology”, explains the models, the work flow, pre-processing applied, evaluation metrics followed in this study.

# 

# LITERATURE REVIEW

## **Introduction**

With high Risks involved with readmission of Diabetes patient, identifying the patients at high risk of readmission has become a prime area of research. At the same time, with lot of data available from Hospitals, many researchers have tried various machine learning models with low-medium accuracy, with good understanding of the model decision and also several deep learning models with high accuracy without providing insight of the model. In the recent years, with evolving deep learning methods, models with high accuracy and interpretability are studied specially in the high-stake decision systems, like healthcare, finance, crime. In this section, we review the classification models used for readmission predictions on the UCI diabetes dataset and other datasets, class imbalance techniques, feature extraction methods, interpretable methods adapted and risk factors identified by various other studies.

## **Risk factors associated with the readmission**

Readmission of Diabetes patients is relatively high compared to any other diseases. Predicting the readmission and risk factors associated with unplanned readmission helps the patients to receive better inpatient treatment, helps health professional to guide the patients with proper discharge ways. Risk factors brings in evidence with the prediction and builds trust in the decision. Lot of research is conducted on widespread population of diabetes around the world. The common predictors cited across research are ‘comorbidities’, ‘length of stay’ and ‘previous admissions. ‘laboratory tests’ and ‘medication’ variables had more weight in the models for cardiovascular disease and medical conditions in relation to readmissions.

### **Demographic characteristics**

Gender:

Effect of gender on the readmission risk prediction among diabetes was studied various researchers using multivariate logistic regression. (Rubin et al., 2016) developed a tool, “Diabetes Early Readmission Risk Indicator (DERRT)” to identify high risk readmission diabetes patient using univariate logistic regression model and handling data cleaning with missing value imputation. The study was based on only the internal samples, later the same author validated DERRT model on external instances (Rubin et al., 2018a) and showed gender has no significant influence on readmission. This was also supported by other studies, (Chen et al., 2012; Emons et al., 2016; Caughey et al., 2017), while other studies found gender is an influencing factor with male gender being at higher risk of readmission, supported with Odds ratio (OR) and Confidence interval (CI) (OR=0.94, 95% CI 0.91 to 0.97) (Collins et al., 2017; Sonmez et al., 2017).

Age

(Chen et al., 2012; Eby et al., 2015; Caughey et al., 2017; Collins et al., 2017; Sonmez et al., 2017) showed older age group population >65 years are highly associated with 30-day hospital readmission. (OR=1.23, 95% CI 1.08 to 1.40)

Race

Authors of (Rubin et al., 2016, 2018b; Enomoto et al., 2017; McCoy et al., 2017) found ‘Race’ to be highly correlated to 30-day readmission in diabetes patients. A consolidated survey (Gek Sang Soh et al., 2020) shows the predictors of unplanned readmission of diabetes patients in which White population is considered to be at low-risk compared to non-white population. Also, the quality of care, the number of emergency visits was considered high in non-white population (Chin et al., 1998)

Health insurance type

(Collins et al., 2017; Enomoto et al., 2017) studied the diabetes population with Insurance plan including, Medicare, Medicaid and other private insurance plans and showed the patients with Medicare or Medicaid plans had high chances of getting readmitted with 30-days of discharge compared to any other privately insured patients.

Number of outpatient visits

(Chen et al., 2012) shows number of outpatient visits did not have significant impact.

### **Comorbidities**

Heart failure, Depression, Respiratory Diseases, Renal Diseases, Hypertension are some of the comorbidities associated with diabetes patients likely to get readmitted according to previous studies. (Eby et al., 2015; Rubin et al., 2016, 2018a; Caughey et al., 2017; Enomoto et al., 2017).

### **Length of Stay**

(Collins et al., 2017) supported the number of days for inpatients stay is associated with higher likelihood of readmission. (Enomoto et al., 2017) discovered that the diabetes patients are more likely to get readmitted and have longer length of stay.

### **Insulin Therapy**

Compared to non-insulin treated patients, patients using insulin were more likely to be readmitted specially for sever dysglycemia (McCoy et al., 2017). A survey of the risk factors predicting 30-day readmission in general diabetes population and subset of diabetes population also supported insulin therapy to increase the risk of readmission.(Robbins et al., 2019)

### **Ways of Discharge**

(Graham et al., 2019; Reid, 2019) performs feature selection to reduce the feature space and show discharge method impacts the readmission heavily. (Transitional Care Interventions Prevent Hospital Readmissions For Adults With Chronic Illnesses | Enhanced Reader, 2021) studied specially the impact of post-discharge on readmission and proved that with better discharge ways readmission can be prevented. (Hospital Readmissions Reduction Program (HRRP) | CMS, 2021) started programs to better post-discharge and showed a relatively good results with reduction in readmission during 2012-13.

## **Applications of Interpretability in healthcare**

Model-specific approaches of ML based models are used in healthcare for more than a two decades due to its ease of use and high level of interpretability Some of the real-world applications are discussed (Stiglic et al., 2020), like urology, toxicology, endocrinology, neurology, cardiology, or psychiatry. However, due to increase in non-homogeneous, non-linear data, these model specific approaches like linear regression or naive Bayes models fails to predict the outcome.

Complex models like deep learning are proven to provide better accuracy, but with low interpretability. (Vellido, 2020) states that improving model interpretability increases the adaption of ML models in healthcare. Local and model agnostic interpretability methods helps achieving this. For example, SHAP was used in interpretation of predictions for the prevention of hypoxaemia during surgery (Lundberg et al., 2018) which increased the anaesthesiologists anticipation of hypoxemia events by 15%. ﻿Model Understanding through Subspace Explanation (MUSE) was used to help in explaining decisions of the neural network model built on depression diagnosis dataset by generating sets of if-then rules that describe the model decisions on a global level.

﻿Machine learning models are effective in identifying risk factors and early diagnosis of various diseases. Predictive models allow healthcare professionals in healthcare resources allocation, better care, and patient outcomes.

## **Interpretable Machine learning Models**

With the Initial study it was found the LACE index (Length of stay, acuity of admission, Charlson comorbidity index and Emergency visit) was considered to be the most preferrable metrics to predict the readmission in Congestive heart failure (CHF), Diabetic patients (Wang et al., 2014; Mingle, 2017) because of its ease of use by health practitioners. But it was implemented using small dataset. So, there was a need for further research.

Recent study (Work, 2017) has used Diabetics Dataset from 130-US Hospital from 1999-2008 dataset which is 1,01,766 records with 50 attributes, which focuses on predicting Readmission of Diabetes patients and a model providing more information to doctors on the factors affecting the readmission. Random Forest provided an accuracy of 89% with promising feature weights that explained the model prediction. But, due to class Imbalance the model accuracy is lost to some extent. The same Random Forest model was tried by (Zhu et al., 2017) with small modification to Random Forest by adding randomness to each feature and rank the features based on meanDecreaseAccuracy. The randomness gives high confidence by eliminating the bias in variables and selecting the high-risk factors influencing readmission. Discharge disposition, number of inpatient visits, emergency visits, number of lab procedures, primary diagnosis were considered as most significant variables for unplanned 30-day readmission. While the initial study (Strack et al., 2014) on the same dataset tried to fit the relationship of HbA1c measurement on readmission of diabetes patients using multivariate logistic regression, controlling the covariates such as demographics, severity and type of the disease and type of admission. The statistical results suggested that the relationship between probability of readmission and HbA1c measurement depends on primary diagnosis.

According to the study (Yang et al., 2016), High predictive power, interpretable results and prediction confidence constitutes a comprehensive framework to predict and understand hospital readmissions. Three models, LASSO, GBM and MLP was used to predict the readmission of Asthma patients wherein all three showed no difference in the prediction accuracy, but MLP was dropped in the final analysis because of the non-interpretability and its high computation cost. LASSO and GBM were finally selected to predict and interpret the model to derive the risk factors. The model’s probabilities are used to construct 95% confidence interval and LASSO achieved a confidence interval of [0.77-0.95] with average readmission probability of 0.875 and GBM with [0.696-0.845] confidence interval and average readmission probability of 0.781. This confidence interval adds Trust to the model prediction results. (Golmohammadi and Radnia, 2016) predicted the patients at high risk of readmission using neural network, Classification and Regression model and Chi-squared Automatic Interaction Detection models with all the models performing at an accuracy above 80%. However, the latter two models had an advantage of providing users with the explanations by deriving the important features contributing to readmission, namely, age, sex, number of previous medications, length of stay, place of service and number of previous claims. The authors used C4.5 algorithm to lean the patterns in history and demographic records of the readmission patients.

CMS (Center of Medicare & Medicaid) started a program HRRP (Hospital Readmissions Reduction Program (HRRP) | CMS, 2021) to avoid excess readmission and improving the patient care with better discharge plans. According to section 3025 of Patients prediction and Affordable Act, Inpatient Perspective payment system hospitals will be reimbursed at lower rate for excessive readmissions. The article (Hospitals are avoiding admitting Medicare patients to dodge financial penalties, study suggests, 2021) stats that in FY-2019, 82% of the hospitals engaged in the program were penalized for readmission. (Maddipatla et al., 2015) aims to predict the 30-day readmission along the cost prediction model and the contributing risk factors associated with it. The study compared the performance of models Decision Tress, Gradient Boosting, Neural Network and Logistic Regression. Based on the AUC results, Decision Tree with 0.95 AUC was chosen to be the better model. And later, with the significant variables obtained, linear model was built to predict the revenue loss to the hospitals which shows financial impact on the hospitals.

A case study on predicting hospital readmission showing the application of Explainable Models in healthcare domain (Vucenovic et al., 2020) using Lasso model, which inherently selects the important variables, and achieved an AUROC of 0.795, when compared to logistic regression model used by Canadian Institute for Health Information, Hospital Admission Risk Prediction (HARP) with AUROC of 0.678 (Canadian Institute for Health Information, 2013).

(Reid, 2019) created a web services to predict the 30-day readmission of diabetes patients using Boosted Decision Tree algorithm, treating the imbalance dataset using SMOTE, selecting the features using Permutation feature importance method, which assigns scores to each of the column and selects the predictor variables with the highest scores. Boosted Decision Tree model showed better performance with binary classification of readmission with 86% accuracy compared to Logistic regression, Neural networks, Decision Tree and Support vector machine. (Graham et al., 2019) worked on the same dataset with a different approach in feature selection, recursive feature elimination, to reduce the feature space and used logistic regression, support vector machine to develop a model and identify 11 important features contributing to readmission. The variable are identified as “age”, “discharge id”, “insulin”, “admission type id”, “admission source id”, “time in hospital”, “number of diagnosis”, “race” and “being on diabetes medicine or not”. (Hu and Sokolova, n.d.) studied the same diabetes dataset considering three class outcomes for readmission prediction using Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting, Decision Trees, Random Forest and Logistic Regression. The author performed pre-processing using the information provided by medical experts and thus optimizing the features and their values which is one of the challenges faced by machine learning models in medical domain (Vellido, 2020) and handled class imbalance using SMOTE, discussed the model functionality and procedures for hyperparameter tuning and showed the 23 medication features highly impact the Recall of five out of six model studied.

One other study on the same UCI diabetes dataset was conducted using C4.5 algorithm, pre-processing the data by replacing the missing values, removing the fields like ‘patient\_id’, ‘encounter\_id’, ‘payer code’ and removing data with Readmitted=NO helped to attain highest accuracy of 74.5%, compared to other pre-processing techniques applied on the same dataset. (Tamin and Iswari, 2017). All these studies comprehend knowledge-base and machine learning concepts to build a better performing model as discussed in the study (Holzinger et al., n.d.).

A comparative study was conducted to diagnose diabetes using Pima India Diabetes Dataset (PIDD) and 130-US Hospital diabetes dataset. The study used Random forest, Naïve bayes, J48, KNN and compared the behaviour of individual models with ensemble of these models on small (PIDD) and large dataset. 10-fold cross validation was conducted and Ensemble of models with stacking outperformed with an accuracy of 88.56% with proper feature selection applied (Alehegn et al., 2019).

(Yu and Xie, 2020) built a cost-sensitive model using Modified Weight Boosting-Stacking (MWBS), improved version of Ada-boosting algorithm with stacking. There are two main contributions of the study. First, to build a high performance predictive model, including feature engineering, data sampling and ensemble learning to alleviate the class imbalance and the high dimensionality and event sparsity of medical codes to overcome the limitations of traditional readmission predictive modelling approaches and improve predictability. Secondly, apart from improving the model predictivity, misclassification cost on the models are analysed to improve the model recall by setting different weight parameter in the model. To reduce the feature space, non-code variable selection was done using Pearsons Chi-square test independence and medical-code variables was handled using six different dimensionality reduction methods, namely, LASSO, ICD-9 code clustering, Relief algorithm, Sparse Bayesian method, Number of codes. After comparing the performance of each of the sampling technique and feature selection method on six different machine learning models, (LR, NB, DT, SVM, ANN, RF) Random forest with LASSO and SMOTE was used as final prediction model. To get the better performance, the model is Boosted and Stacked with Modified weight model. Yet, another approach of cost-sensitive deep-learning was applied to predict the readmission of various diseases in a study (Wang et al., 2018). The study used different data, clinical, demographic, lab tests which is in the form of time-series and built the model using CNN to automatically derive the features of lab-tests holding the data as time-series and converting categorical variables to numerical data using look-up table, then providing these variables as input to MLP (with multiple hidden layers) and output layer formulating a cost-function (Bayesian Optimal Decision system) to minimise the cost of wrong predictions. The model outperformed other baseline models on the same dataset with AUC 0.70 and accuracy of 89%. The performance is boosted because of its extensive pre-processing and the feature engineering applied.

In the recent years, a new method, Deep Forest, was used as an alternate to deep learning model because of the complexity, high computation, hyperparameter tuning and requirement of large training data with deep learning models. Figure 2.1(a) shows the model of Deep forest and Figure 2.1(b) shows the how each of the forest contributions are averaged. This new method was tested on various domains like Image classification, face recognition, music classification, sentiment classification, and showed good results when compared to deep neural networks (Zhou and Feng, 2017). Deep forest was also used in the prediction of readmission on the same dataset as this study (Hu et al., 2019), in which the authors tried to use wavelet-transform for pre-processing and attained an AUROC of 0.726, when compared to other baseline models, Random forest, SVM, Stochastic Gradient Descent (SGD) and Logistic regression.

Diagram

Description automatically generated

A picture containing diagram

Description automatically generated

Figure 0.1 (a): De (Zhou and Feng, 2017) Figure 2.1 (b) Single forest view De (Zhou and Feng, 2017)

Table 0.1 Summary of machine learning models

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Dataset | Models | Results |
| Wang et al., 2014; Mingle, 2017 | Diabetes,CHF |  |  |
| Zhu et al., 2017 | UCI-130 diabetes dataset | Random Forest + randomness | Risk Features |
| Work, 2017 | UCI-130 diabetes dataset | Random Forest | 89% accuracy |
| Strack et al., 2014 | UCI-130 diabetes dataset | Multivariate Logistic regression | Risk Features |
| Yang et al., 2016 | UCI-130 diabetes dataset | LASSO, GBM, MLP | LASSO = 0.87  GBM 0.781 |
| Maddipatla et al., 2015 | UCI-130 diabetes dataset | DT, Gradient Boosting, NN, LR | DT=0.95 AUC |
| Vucenovic et al., 2020 |  | LASSO | 0.795 AUC |
| Canadian Institute for Health Information, 2013 |  | LR | 0.678 AUC |
| Reid, 2019 | UCI-130 diabetes dataset | Boosted Decision Tree algorithm | 86% accuracy |
| Tamin and Iswari, 2017 | UCI-130 diabetes dataset | C4.5 algorithm, | 74.5% accuracy |
| Golmohammadi and Radnia, 2016 |  | neural network, Classification and Regression model and Chi-squared Automatic Interaction Detection models | 80% accuracy |
| Alehegn et al., 2019 | Pima India Diabetes Dataset (PIDD) and UCI-130 diabetes dataset | Random forest, Naïve bayes, J48, KNN and compared the behaviour of individual models with ensemble of these models on small (PIDD) and large dataset | 88.56% accuracy |
| Hu et al., 2019 | UCI-130 diabetes dataset | Deep Forest | 0.726 AUC |
| Yu and Xie, 2020 |  | Ensemble Model (Random forest using Modified Weight Boosting-Stacking (MWBS)) |  |
| Wang et al., 2018 |  | CNN+MLP | 0.7 AUC, 89% accuracy |

## **Deep learning Models for Prediction**

Deep Learning models are receiving more demand compared to other Machine Learning classifiers in computer vision, image processing and medical classification. The main characteristics of deep learning models to attain this high demand is its power of processing massive amount of data generated by “Big Data Era” technology. Gaining insights into this complex, heterogenous and high-dimensional data remains a challenge in transforming healthcare. Specially, using the traditional data mining and statistical approaches, where explicit features engineering is done and then build the model. The recent advances in deep-learning technologies overcomes these challenges providing end-end learning (Miotto et al., 2017). So, the other most advantage of deep learning model is its ability to iteratively learn the features without the need for domain knowledge and core feature extraction which is an important building block of explainable AI in medical domain (Holzinger et al., n.d.).

Deep learning models provides extremely robust approaches for learning complicated, non-linear data, trying to establish relationship between new patterns to improve the performance of the model. The basic building blocks of deep-learning models is the feed-forward neural network, as shown in the Figure 2.2, constructed using three layers, an “input layer”, reading the predictors, a “hidden layer”, processing inputs and a “output layer”, predicting the decision.

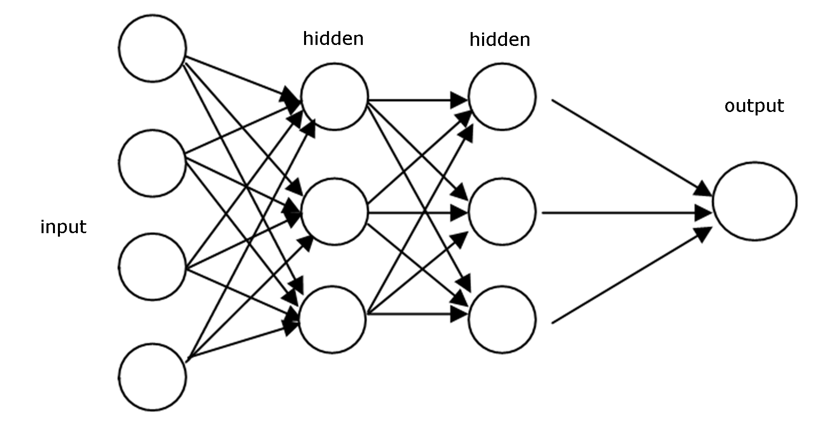


Figure 0.2 Feed-forward architecture

Neural networks works on the model of the brain, where neurons are connected to each other transforming the information from one neuron to the other. Neural networks which is built of several layers, consisting of multiple neurons at each layer. A fully connected networks connects each of the neuron in one layer to every neuron in its next layer. Each of these neuron learns the features, weights, and transfers the output to the next layer to strengthen the relationship between the connections of neurons. For example, if you need to identify a cat in image, the first layer learns the edges, corners and next layer looks at ears, nose, hence increasing the layers to increase the prediction accuracy. This way neural networks increases the number of layers and the number of neurons in the model, leading to increase in number of parameters which are cumbersome and difficult to understand. This limitation stops the application of ML models in healthcare, but researchers have used various other methods to interpret the model to produce meaningful insights.

(Bhuvan et al., 2016) conducted research on US 130 Hospital Dataset to predict 30-day readmission and identified the risk factors using black-box models, Naive Bayes, Bayesian Networks, Random Forest, Adaboost and MultiLayer Perceptron. MultiLayer Perceptron and Random Forest performs better when compared to other classifiers, with slight prevalence in Random Forest favour with AUC of 0.650. And feature analysis was performed to derive the risk factors influencing the readmission of diabetes patients and help medical practitioners gain more insights to understand why patients got readmitted within 30 days of the discharge. Ablation study of risk factors and Association Rule Mining was used for feature analysis and concluded with “inpatient incidents”, “discharge disposition” and “admission type” as most important features.

A study (Futoma et al., 2015) was conducted to predict the readmission of five different diseases using deep-learning algorithm and the other traditional models like logistic regression, logistic regression with multi-step variable selection, logistic regression with penalizing large coefficients using LASSO and Ridge, SVM, Random forest. In the initial study, penalized method showed better outcome compared to other models, but deep neural networks outperformed in 3/5 populations with high AUROC compared to penalized models.

A study to predict the rehospitalization of infants compared the predictive ability of models constructed using LASSO and Random forest with expert-driven logistic regression model (Reed et al., 2021). Logistic regression was built with 10 predictors derived from expert opinion and LASSO and random forest was built using all the 75 predictors. The goal of the study was to compare the impact of considering all the 75 predictors Vs the expert driven ten predictors and showed that Random forest model demonstrated well with AUROC of 0.65 compared to LASSO and traditional logistic regression. Random forest aggregates the predictions of all the decision-trees, retaining large number of predictors without over-fitting the model, when compared to experts driven selection of predictors where some of the information is lost. And also, since the data was not linearly separable, Random forest captured the non-linear relationship between the predictors and the targeted outcome variable, which improved the performance of the model. Apart from retaining the additional features, researchers tried to tune the traditional models, (Sidey-Gibbons and Sidey-Gibbons, 2019) conducted a study to evaluate impact of performance of machine learning algorithms by tunning the algorithms to predict breast cancer. The algorithms include regularized General Linear Model regression (GLMs) with L1 regularisation LASSO, Support Vector Machines (SVMs) along with reducing the high-dimensional feature space using a well-known kernel transformation method, (Radial basis function) RBF kernel, and single-layer Artificial Neural Networks (ANN). SVM attained the maximum accuracy of 0.96 and AUROC=0.97, while LASSO also performed well preserving the ability to understand the features impacting the decision and ANN did not perform well generating high-parameters. However, SVM, which is a ‘black-box’ is still considered due to its high predictive performance in the applications where the reason for the decision is not important like image processing, speech recognition with large features. This looks for research in the area of improving the predictive performance using neural networks and even deep learning.

A research has been published (Hammoudeh et al., 2018; Goudjerkan and Jayabalan, 2019; Hu et al., 2019; Reddy et al., 2020) in predicting readmission of diabetes patients using neural networks with the main aim to improve the performance of classifier at the expense of interpretability of the model. The researchers of (Goudjerkan and Jayabalan, 2019) also used Diabetes Dataset from 130-US Hospital from 1999-2008 dataset considered in-depth data pre-processing, performing, Missing value imputation using Hot-Deck imputation, Data reduction using clustering, handling data inconsistency, Outlier treatment, Class Imbalance using SMOTE along with Feature Engineering and achieved the model accuracy of 95% using Multilayer Perceptron to predict the 30-days Readmission for Diabetes patients. The study focuses more on accuracy and not interpretability of the model which is essential in health care as shown in (Yang et al., 2016). (Hammoudeh et al., 2018) compared traditional artificial neural networks (ANN), convolution neural networks (CNN) and other machine learning models and resulted in CNN hitting ~92% accuracy, ~95% AUC. Neural networks performance depends on size of the data, so data engineering was handled appropriately, applying SMOTE to balance the data, feature creation by adding two new features from drug attribute, using one-hot encoding for feature transformation and removing the duplicate records. While other study (Sarthak et al., 2021) on the same dataset used deep neural network achieving an accuracy of 95.2%. Authors of (Reddy et al., 2020) compared the results of proposed algorithm DBFN (Deep Belief Network) with other machine learning models, Logistic regression, Ada-boost, Gradient Boosting, Random Forest, Decision Tree. DBFN resulted in 69% accuracy compared to other low accuracy machine learning models. Data Imbalance and reduction in data size might be the reason for low accuracy with models. One other study used under-sampling for balancing the dataset and used three incremental learning methods like Naive Bayes, Hoeffding Tree (Very Fast Decision Tree, VFDT) and Hoeffding Adaptive Tree to build a predictive classification model and showed that Naïve Bayes performed better with an AUC of 0.655.

A comparative study (Futoma et al., 2015) between machine learning models and deep neural networks to predict early readmission of patients on five different diseases was conducted. The study compared Logistic regression, Support vector machine, Random forest with deep neural network model and concluded with deep neural network performing better with all five different population.

LSTM was used to predict the rehospitalization in lupus patients, which outperformed compared to traditional classic ANN and penalized logistic regression model. The ability of LSTM to study the complex relationship and capture the patient’s prior visits clinical information in the memory may have enabled the higher performance with LSTM.

Table 0.2 Summary of deep learning models

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Dataset** | **Models** | **Results** |
| (Bhuvan et al., 2016) | UCI-130 Diabetes | Random Forest, MLP | Random Forest, AUC=0.65,  Risk Features |
| (Futoma et al., 2015) | Five different diseases | Deep neural networks |  |
| (Reed et al., 2021) | Infants | Random Forest | ACU = 0.65 |
| (Sidey-Gibbons and Sidey-Gibbons, 2019) | Brest cancer | SVM | AUC = 0.97 |
| (Goudjerkan and Jayabalan, 2019) | UCI -130 diabetes | MLP | 95% accuracy |
| (Hammoudeh et al., 2018) | UCI – 130 diabetes | CNN | 92% accuracy |
| (Sarthak et al., 2021) | UCI – 130 diabetes | Neural network | 95.2% accuracy |
| (Reddy et al., 2020) |  | Deep Belief network | 69% accuracy |
| (Futoma et al., 2015) | Five diseases | Deep neural network |  |
|  |  | Naïve bayes | AUC = 0.655 |
|  | Lupus | LSTM |  |

## **Black Box Model Explanation Problems**

Methods applied to explain a model depends on the problem at hand. The problems can be of different types, (1) Model Explanation or the Global Interpretation (2) Model Outcome Explanation (Local Interpretation) (3) Transparent Design (4) Model Inspection. Details on each of the problems and the respective methods applied is shown in the survey (Guidotti et al., 2018b). In this study we focus the Model Outcome explanation problem, where it is required to provide evidence for a patient’s prediction to high risk of readmission.

### **Global Interpretation**

The first move towards getting better accuracy and achieving model interpretability was presented by (Lou et al., 2012) using GAMs (Generalized Additive Model). GAMs hold good with only Univariate terms and was able interpret regression splines of both linear and logistic regression along with single trees, ensembles methods like bagging, boosting, random forest. The explanation was presented as the importance of contribution of each individual feature towards the decision together with Shape functions. Same authors refined the study by including the pair-wise interactions and univariate terms, resulting in a model Generalized Additive Model plus Interactions (GA2 M) (Lou et al., 2013) gaining better accuracy as the complex models along with interpretability. This study was evaluated by predicting the pneumonia risk and hospital readmission with 30-days of discharge (Caruana et al., 2015) where it has uncovered few patterns that the complex neural nets failed to do even with highest accuracy. For example, the rule-based methods learned that patients with history of asthma are at low-risk, because asthmatic patients were directly admitted to ICU and with better treatment, the survival of these patients was mush higher when compared to general population. But the way data is presented, models tend to learn the patten in a negative way, that means asthmatic patients with pneumonia are at low risk. So, with methods providing reasoning to decision, like the rule-based methods, it was easier to detect the rule leading to wrong decisions and it can it mitigated by removing the rule or removing the feature, but complex models without interpretability it is difficult to trust the model. (Rudin, 2019) also tries to bring in the importance of understanding the model in healthcare.

### **Local Interpretation**

These are explanation methods that are applied to any black box models and are generalized in nature. The first agnostic method applied to a black box model is referenced in (Poulin et al., 2006) ExplainD (Explain Decision). This study brings out the graphical representation of the decision process, visualizing the feature importance of the decisions, capable to speculate the changes in the data and whenever possible dive deep to get to the source of evidence. ExplainD is specially designed for few classification methods like, Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and can be generalized for any black box models. ExplainD uses the concept of additive models to weight the importance of the features of the input dataset.

Then, (Štrumbelj and Kononenko, 2010) presented a generalized explanation method, explaining individual predictions of any classification model using the fundamental concepts of coalitional game theory and the explanation is depicted by the contribution of each feature value towards the prediction.

To add interpretability to neural networks, the book (Interpretable Machine Learning, 2021) shows the model interpretation methods and its importance. Various research has been done on the adapting of these methods. (Shrikumar et al., n.d.) used model agnostic method of interpretation, DeepLIFT (Deep Learning Important FeaTures) on MNIST dataset and simulated genomic data. DeepLIFT learns the important features by decomposing the output prediction of neural network for any given input by backpropagating the contributions of all neuron to each of the input feature and assigning scores to each feature comparing the difference of the activation with its ‘reference activation’. (Ribeiro et al., 2016) presented Local Interpretable Model-Agnostic Explanations(LIME) to explain the predictions using local models. Author (Lundberg et al., n.d.) compared LIME, SHAP, DeepLIFT to explain the model and presented SHAP to perform better in consideration with human explanations, while the recent study (Fidel et al., 2019) introduced a variation of SHAP, SHAP DeepExplainer to explain the DNN models. SHAP was also used to explain the predictions from Long Short-Term Memory (LSTM) algorithm. For 30-day readmission prediction, “length of stay” was largest impact along with “medications” and “total cost” (Chi et al., 2020).

An explainable method should possess three important properties, (1) Easily understandable by humans (2) Easy to apply and provide quick explanations (3) It should be efficient and consistent. Although LIME is considered as one of popular explainable methods for DNN, it experiences high computation power while training linear models for the perturbed samples around the target variable to produce the explanations and is considered to be very sensitive to changes in the input. Even a small change in the input which the model prediction does not capture, LIME tends to show significant change in its explanation. The robustness of the these interpretable method is argued by authors of (Alvarez-Melis and Jaakkola, n.d.). As show in Figure 2.3 (a), the study shows the performance of LIME and SHAP using various UCI classification dataset. Figure 2.4 shows the explanations of MNIST digit with maximum perturbation. To overcome the robustness, a recent study (Modzy and Modzy, n.d.) introduced a novel algorithm to explain the predictions of black-box models by using the behaviour of Adversarial attack on the models and called it as AXAI (Adversarial Explainable Artificial Intelligence). The author compared the explanations on ImageNet using LIME and SHAP and showed LIME performs slower compared to other two methods. This was because LIME iteratively forward propagates the perturbated inputs through DNN. Whereas, LIME was proven to perform better in explaining a clinical data used in the study (Barr Kumarakulasinghe et al., 2020). The author evaluated the method at “Application level” by comparing the explanations provided by physicians. However, there are few limitations associated it, exhibiting robustness, unstable results with high-dimensional dataset as seen in (Visani et al., 2020), the author used LIME to explain the predictions of the credit payers.

Table

Description automatically generated with medium confidence

Figure 0.3 (a) Explanation of SHAP (b) Explanation of LIME (Alvarez-Melis and Jaakkola, n.d.)

A picture containing text

Description automatically generated

Figure 0.4 MNIST digits with maximum perturbations and their explanations (Alvarez-Melis and Jaakkola, n.d.)

Rule-based methods are considered to be most prominent among the interpretable models. A model agnostic rule-based method capable of providing faithful and interpretable explanation was presented in the study Local Rule-based Explanations (LORE) (Guidotti et al., 2018a), overcoming LIME and Anchor in terms of prediction accuracy and quality of explanations. LORE first gets trained using the local interpretable predictor, decision tree, on synthetic neighbourhood generated through genetic algorithm approach. Then, from the logic of local interpretable predictor it derives the explanations consisting of decision rules, explaining the reason of the decision, and a set of counterfactual rules, suggesting the changes in the instance’s feature leading to different outcome. The study was experimented on three datasets, (1) adult dataset, predicting the income of the person (2) compass dataset, predicting the defendants and their risk (3) German dataset, classifying a person as ‘good’ or ‘bad’, over black box models, Random Forest, Support vector machine and multilayer neural networks.

Then compared the explanations using hit scores derived by applying LIME and LORE over the combination of dataset and models. Results showed that LORE outperforms the explanations of LIME, where LORE automatically figures out the important features participating in the decision, but LIME requires the user to be provide the number of features composing the explanation. And also compared the results with Anchor, an extension of LIME which provides explanation in the form of decision rules.

### **Transparent Box Design**

This section reviews the approaches used to solve the classification problem using transparent methods that are locally or globally interpretable on its own. Linear regression, logistic regression, decision trees are considered as transparent models but with the compromise on the accuracy compared to complex models. But still given the importance of interpretability specially in healthcare domain, or critical decision support system, Rule-based models are popular among all these. There is lot of research done in improving these rule-based methods.

A study (Lakkaraju et al., 2016) proposed a variant of rule-based method, Interpretable Decision sets (IDS) for generating predictive models with high accuracy and high interpretability. Decision sets are the set of if-then rules in Figure 2.5 (left), that are independent of each other, which makes it easier to understand, when compared to other rule-based models with if-else rules, in Figure 2.5 (right). In general, decision sets are short, non-overlapping, accurate rules covering the entire feature space. As part of evaluation, the model was used in the prediction of three real world problems (1) diagnosis of patients with six different diseases (2) identifying students at risk of dropping out from attaining their diploma degree on time (3) identifying the defendants released on bail are suspected to commit violent or non-violent crime. The model outperformed in all these scenarios with high AUROC, Disease diagnosis (61.19), Bail (69.78), Student performance (75.12) when compared to other rule-based classification models: Bayesian Decision List (BDL), Decision Trees (DT), Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Classification Based on Association (CBA).

﻿If Respiratory-Illness=Yes and Smoker=Yes and Age≥ 50 then Lung Cancer If Risk-LungCancer=Yes and Blood-Pressure≥ 0.3 then Lung Cancer If Risk-Depression=Yes and Past-Depression=Yes then Depression If BMI≥ 0.3 and Insurance=None and Blood-Pressure≥ 0.2 then Depression If Smoker=Yes and BMI≥ 0.2 and Age≥ 60 then Diabetes If Risk-Diabetes=Yes and BMI≥ 0.4 and Prob-Infections≥ 0.2 then Diabetes If Doctor-Visits ≥ 0.4 and Childhood-Obesity=Yes then Diabetes

﻿If Respiratory-Illness=Yes and Smoker=Yes and Age≥ 50 then Lung Cancer Else if Risk-Depression=Yes then Depression Else if BMI ≥ 0.2 and Age≥ 60 then Diabetes Else if Headaches=Yes and Dizziness=Yes, then Depression Else if Doctor-Visits≥ 0.3 then Diabetes Else if Disposition-Tiredness=Yes then Depression Else Diabetes

Figure 0.5 ﻿An interpretable approach, decision sets (left), and a decision list (right) learned from the same medical diagnosis dataset. Decision sets (left) are set of if-then rules and are independent of each other. While Decision lists (right) are if-then-else rules, where the rule depend on the decision above it. For, humans readability, decision sets are more comprehensible. (Lakkaraju et al., 2016)

## **Class Imbalance Techniques**

Medical data is the real time data collected from various hospitals which is always accompanied by high-dimensionality, noisy, missing and imbalanced data. Class Imbalance is where one class of data is more than the other. For example, in the detection of loan defaulters, credit card fraud detection, telecom churn, the number of legitimate transactions is always higher than fraudulent transactions. The class with more data points is the majority class and other is the minority class. The level of class imbalance is measured by the Imbalance Ratio (IR), that is a ratio of 1:10 says for everyone sample of positive class there are 10 samples of negative class. Unlike any other machine learning model’s linear regression model is not affected by class imbalance due to its linear separability feature, but most of the classification model favour the majority class when the dataset is imbalanced, with high accuracy for majority class and low accuracy for minority class (Lin and Chen, 2013). This can be treated in three different ways, re-sampling, hybrid sampling, cost-sensitive learning, as shown in Figure 2.6 (Lu et al., 2019) shows the impact of Imbalance on the model performance and treatment of imbalanced data not always improves the performance but can reduce as well. So, deciding to balance the data is important, and this depends on data characteristics, like overlapping of minority and majority class data (Napierala and Stefanowski, 2016). (Zhou, 2013) recommended different sampling ratio for different data size, while (Loyola-González et al., 2016) used the automatic selection of sampling rate with respect to the imbalanced ratio. After detailed study of sampling methods used by various authors presented in (Artetxe et al., 2018), resampling was mostly used to overcome class imbalance. The performance of models using re-sampling was studied to predict student’s performance (Ghorbani and Ghousi, 2020) and concluded Random Forest classifier performed well using SVM-SMOTE (76.83% Accuracy), compared to others, Decision Tree, Artificial Neural Network, XG\_boost, Logistic Regression and Naïve Bayes. Some insights from these papers: 1). When there are hundreds of minority observations in the dataset, an under-sampling method was superior to an over-sampling method in terms of computational time. 2). When there are only a few dozen minority instances, the over-sampling method SMOTE was found to be a better choice. 3). If the training sample size is too large, a combination of SMOTE and under-sampling is suggested as an alternative. 4). SMOTE is slightly more effective in recognizing outlier.

## **Summary**

Random forest classification algorithm uses multiple decision trees and bagging to merge predictions across the multiple trees. The advantages of random forest include the efficient consideration of larger predictor sets, a reduced risk of overfitting by considering the collective decision and an ability to manage non-linear relationships between predictors and predicted probabilities more effectively (Reed et al., 2021). Neural networks is another model with high performance which learns the non-linearity by iteratively passing through the layers of the network. However, more complex models like CNN, LSTM did not show better performance, so a single hidden layer MLP is considered and gave a high performance compared to any other models.

Quality of the data highly impacts the model performance, so pre-processing the dataset and making it suitable for model building is very important. To trust, maintain fairness and transparency of a specific model and its predictions, it is important that we understand different approaches to model interpretability. Interpretability can be model specific or model agnostic, locally or globally interpretable. Model interpretability is as important as the accuracy of the model. With the recent developments to build explainable AI, model agnostic interpretable methods are more popular. This bridges the gap between the prediction of the model and the end user.

Models for class Imbalance

Resampling

Ensemble

Cost-sensitive

Undersampling

Oversampling

Hybrid

Hybrid

Cost-sensitive

Boosting

Bagging

Figure 0.6 Different methods of addressing class imbalance

## **Discussion**

Readmission prediction in diabetes patients is one of the measure of hospital efficiency and identifying the patients at high risk is important for patients, policy makers and care givers.

Several machine learning models are used, Logistic regression, Decision tree, SVM, Gradient boosting, Random forest along with deep learning model, multi-layer perceptron, CNN, LSTM. On some datasets, machine learning algorithms shows good performance while most of the time Deep learning models performed better in predicting readmission in diabetes patients or Congestive heart failure, pneumonia patients. But still, some researchers still go with the machine learning models due to its simplicity in understanding predictions. Models like Logistic regression, decision trees, random forest has inbuilt methods to provide feature importance, while deep learning models need external methods to explain the model. In few cases like medical, finance domain interpretability is as important as accuracy, so a model with high accuracy and high interpretability is required to build trust and confidence in the model. Some of the features like Discharge disposition, number of inpatient visits, emergency visits, number of lab procedures, primary diagnosis were considered as most significant variables for unplanned 30-day readmission in UCI dataset, whereas race, gender, insulin, insurance type are generic factors identified by various authors using logistic regression.

# 

# RESEARCH METHODOLOGY

## **Introduction**

This section will highlight the methods and algorithms used in this research to accomplish the objectives. Each of the following subsections explains Dataset used, data cleaning, data engineering techniques, SMOTE as a technique for balancing the data, feature selection using Random Forest, and the model classifier Random Forest, Multilayer Perceptron are trained and tested. Finally, the model interpretation methods LIME, SHAP, DeepLIFT are explained and used with models. The model performance is evaluated using accuracy, precision, recall, AUC and the decisions of the models are explained with the supporting features of prediction.

## **Dataset Description**

The Dataset used in this study is a de-identified abstract of the Health Facts database (Center Corporation, Kansas City, MO) (UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set, 2021). The dataset is collection of diabetes patients from 130 US Hospitals over a span of 10 years (1999-2008) which includes 1000000 records with 55 attributes. The attributes hold the following information,

* It is a hospital admission.
* It is diabetic encounter, that is one of 3 diagnosis is identified as diabetes.
* The length of the stay between 1-14 days.
* Medications offered at the time admission.
* Laboratory tests conducted during the encounter.
* Discharge deposition
* Demographic information such as race, gender.
* Target variable ‘readmitted’. (> 30 days, < 30 days, No)

## **Data Pre-processing**

The real world data is not gathered with respect to any specific purpose in mind. The data is noisy, inconsistent, heterogeneous, with large amount of missing values. Specially health care data includes lot of attributes, patient demographic information, medications, ways of discharge, laboratory tests, inpatient/outpatient details, large number of ICD codes and so on. Pre-processing is a process of converting the raw data suitable for the model. So, before starting to build the model, it is required to pre-process the data appropriately and have the relevant info to achieve better performance of the model (Khurana and Kumar, 2019). As seen in the literature survey (Mingle, 2017; Hammoudeh et al., 2018; Goudjerkan and Jayabalan, 2019; Hu et al., 2019) model performed better with data pre-processing applied, data transformation done, sampling techniques adapted, reducing dataset and optimizing it (Duggal et al., n.d.). In case of large feature space, model tends to overfit, and increases the computation cost, in few cases the data has to normalized to avoid any irregularities in the model. In this study we perform feature selection, missing values imputation, removing outliers, feature engineering, balancing class, data reduction, data cleaning and data transformation. As such, Random Forest feature selection, ICD-9 code clustering, Hotdeck imputation are the key pre-processing tasks.

### **Missing Value Imputation**

With this real world dataset of diabetes patients, there are a lot of missing values, ‘weight’ (97%), ‘payer code’ (40%) and ‘medical speciality’ (47%). As a general rule of thumb, a feature with more than 50% missing values should be removed. As such, ‘weight’ is dropped with highest percentage of missing value and does not influence this study. However, based on the previous research and understanding the background ‘payer code’ that shows the socio- economic condition of the patient is considered as important factor to predict readmission and ‘medical speciality’ is also highlighted to be the great influencer and its missing values are encoded as ‘missing’. In addition, few variables with low missing values and irrelevant to the study are removed from the dataset. Others with average missing value, imputation was performed in order to retain enough data to test and train the model. Imputation is a method to replace the missing value with the existing data in the dataset. There are multiple ways of imputation, Mean Imputation, Logistic Regression Imputation, HotDeck Imputation. Imputation of each variable depends on its significance in the study, from the previous studies and recommendations, HotDeck imputation was considered to one of the reliable and commonly used imputation technique (Rithy, 2016). HotDeck imputation is a procedure of replacing missing value using values from one or more similar instances.

### **Data Cleaning**

This second step of pre-processing refers to cleaning of duplicate or inconsistent data that alter the performance of the model. From the body of literature, this dataset has some unstable features that need to be taken care of. For instance, the dataset has multiple encounter of the same patient. So, only one record is retained while the others are deleted to avoid the bias in the model, the first encounter is selected (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021), while the last encounter would have lot of imbalanced data.

Discharge disposition refers to location or status, where the patient is moved after the admission. There are patients who died during their hospitalization and there is no probability of these patients to be readmitted, so exclude the records of the patients with ‘expired’ discharge disposition, 11, 19, 20, 21. Few others like ‘encounter\_id’, ‘patient\_nbr’ is removed. Features like ‘examide’ and ‘citoglipton’ are listed as ‘No’ for all records, and will be deleted from further analysis.

### **Data reduction**

After the data is cleaned replacing the missing values and excluding the records leading to bias, improve the data by reducing the number of unique values in categorical variables. For instance, ICD-9 codes are used for ‘diag\_1’, ‘diag\_2’, ‘diag\_3’. The code ranges from 001-1000. These numerical values are difficult to interpret and make it unusable in the model. (Yu and Xie, 2020) studied various data reduction methods on medical code, like LASSO, ICD-9 code clustering, Relief algorithm, Sparse Bayesian (SB) method, Number of codes. LASSO and SB performed better, but on different dataset. ICD-9 Clustering technique used in (Mingle, 2017; Goudjerkan and Jayabalan, 2019) on the same UCI-130 diabetes dataset, is applied to cluster the diagnosis code into nine categories, “Circulatory, Digestive, Diabetes, Injury, Genitourinary, Respiratory, Musculoskeletal, Neoplasms and Others. The clusters(group) are formed based on the ICD-9 code ranges to specific diseases and if there were less than 3.5% of the encounter, those were categorised as “Others”. Other features with high unique values, “medical\_speciality” , “admission\_type\_id”, “admission\_source\_id”, “discharge\_disposition\_id”, are converted to categorical values and will be clustered to enrich the dataset and interpret the model (Ahmed and Kabir, 2019; Goudjerkan and Jayabalan, 2019). “admission\_type\_id” has 8 unique values, is converted to 4 categorical variables, by considering the description mentioned. “Emergency”, “Urgent”, “Trauma Center” were marked as “Emergency”, similarly 5,6,8 are grouped as “Not available”, and remaining 3,4 are considered the same. “medical\_speciality” has 73 unique values, converge them based on the semantic and general understanding. “admission\_source\_id”, “discharge\_disposition\_id”, has 26, 29 unique values respectively, will be clustered based on percentage of encounters and its semantics.

### **Class Imbalance**

The target variable considered for this study, “readmitted”, is found to be imbalanced with higher rate of “No” (~53%). In case of predictive modelling, Imbalanced classification results in poor performance, especially with respect to minority class. As the study focuses on predicting readmission within 30 days, which is a minority class, only holds ~11% of the data. Classification of minority class (“interested”) is more important than majority ones. In this case, the cost of misclassification of minority class is more compared to majority class. For example, if a patient who has to be readmitted is classified as “not readmitted”, it increases the readmission rate. Hence, treating class Imbalance is essential for the study. There ae different sampling techniques discussed in the literature review, resampling, cost-sensitive-learning and ensemble method. Oversampling method, Systematic Minority Over-sampling technique (SMOTE) (Chawla et al., 2002) is most preferred technique, (Goudjerkan and Jayabalan, 2019; Ghorbani and Ghousi, 2020; Sarthak et al., 2021) and is used to balance the data, generating synthetic records of minority class in the vicinity of its nearest neighbours, to balance the output classes. This makes the decision region of the minority class to be more generalized. While to use under-sampling, the data is not large enough to train and maintain the model performance.

SMOTE process:

1. Identify the feature vector (sample) under consideration and its nearest neighbours.
2. Take the difference between the sample and its nearest neighbour.
3. Multiply the difference with a random number between 0 and 1.
4. Add the result to the sample under consideration.
5. A random point is selected on the line segment joining the two specific features.
6. Repeat this for all the selected feature vector.

### **Feature Engineering**

Feature Engineering is the most important step for successful data-analysis. Specially with the medical data which contains lot of features with varied ranges. The study will perform feature creation, feature encoding, feature selection to gain significant insight of features and Log transformation to standardise the data.

Feature Creation:

The services utilized by the patients over the last year are measured using, “number\_inpatient”, “number\_outpatient” and “number\_emergency”. Adding all these three variables, a new feature “service\_utilization” will be engineered. Medication change was another feature of importance as predicted by (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021; Sarthak et al., 2021). There are 24 features of medication in the dataset, the feature indicates whether the drug was prescribed or was there a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no’ if the drug was not prescribed. A new feature, “medication\_change” is engineered by counting the change in medication during the patient encounter. And, to account for total number of medications used, “count\_of\_medication” will be added as new variable, which will be the sum of all medications used during the stay in hospital. Higher the number, higher will be the severity of the patient. Table 3.1 shows the added features with the computation. Most of the researchers tend to remove the original features and replacing it with new features, but this study considers all the features and statistically eliminate the features.

Table 0.1 Newly added features

|  |  |
| --- | --- |
| Feature | computation |
| service\_utilization | number\_inpatient + number\_outpatient + number\_emergency |
| medication\_change | Up+down (change in medication) |
| count\_of\_medication | Total number of medications |

Feature Encoding:

The target variable “readmitted” is of three category, < 30-days, >30-days, “NO”. The study is focussed on readmission within 30-days, so above 30-days are considered as no readmission and it accounts for (~34%) and actual no-readmission is (~54%). Ideally the observations above 30-days is discarded, but it leads to reduction in the size of data and affects the model accuracy. Thus the output class is encoded as binary ‘1’ for a readmission and ‘0’ for “non-readmission”. A change in the medication represented by column, “change” will be encoded as ‘0’ for no change in medication and ‘1’ denoting a change in medication for values “No” (no change) and “Ch” (change) respectively. A column describing if the diabetes medicine was prescribed or not, “diabetesMed” denoted with values ‘yes’ or ‘no’ are encoding to ‘1’ and ‘0’ respectively.

Log transformation:

According to the data exploration conducted by previous researchers, numerical variables such as “number\_inpatient”, “number\_outpatient”, “number\_emergency”, “service\_utilization” are highly skewed and log transformed to ensure normal distribution (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021). In order to retain the zero values in the columns, log(x+1) is used, log(x) converts the zeros to -inf. After achieving the normal distribution of numerical variables from log transformation, standardise the data using z-score scaling, where the difference in the mean of the variable and the variable itself is divided by standard deviation. This results in the dataset have equal mean(0) and standard deviation (1).

Standardization (X) = , where μ is the mean of the features and σ is the standard deviation of the features.

Feature selection:

Feature selection is essential pre-requisite of data modelling (Ogedengbe and Egbunu, 2020) and has a great impact on model performance. Not all the features available are relevant to the study(classification), and their relevance is not unknown at the beginning, may be the domain knowledge can help in understanding few features, but to a non-expert is it a difficult. With large features, the computation time is high and model performance decreases. So, the contribution of each feature towards the decision is assessed. Some of the researchers use p-value, Chi-square independence test or use domain experts decision to select the relevant features. This study uses the Random forest feature selection method, like other authors (Kursa and Rudnicki, 2010; Ahmed and Kabir, 2019). Previous studies have used Random forest for identifying the risk factors of readmission. It is an ensemble of various week classifiers – Decision tree classifier. The factors identified are based on the voting of decisions of these classifiers and hence removing the bias of individual prediction. Random Forest is selected for its accuracy, speed of execution, efficient on high-dimensional huge data without overfitting and not requiring data pruning.

Scikit-learn provides the inbuilt method with the model, “feature\_importances\_”, to easily get the importance of the features in the form of scores.

Outlier Removal:

Outliers are the observations or a set of observations which stands inconsistent from the rest of the observations of the same group. Outlier leads to miss classifications problems, if not treated. Once the log transformation and standardisation is done, the outliers can be treated by deleting the data outside 3 standard deviation range (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021).

## **Modelling**

The main objective of this study is to build the model with high predictive performance along with explaining the features influencing the readmission. Logistic regression or linear regression are good classification models with high ability to explain the model prediction but the model fails when the data is non-linear. Models like Random forest, decision trees, neural networks can handle the non-linearity. Based on the performance studied by other researchers, and the purpose of this study to build an interpretable model, will include two black box models, Random Forest with Gini and Multilayer perceptron (MLP). Then, use the model agnostic interpretation methods SHAP, DeepLIFT, LIME to explain the model via the contribution of features.

1. Random Forest Classifier:

Random Forest classifier as it indicates is an ensemble of large number of individual trees- Decision trees to form a forest and the term “Random” is used due to the randomness in the: (1) subset of data selected with replacement (2) random features selected, to train each decision tree. This technique of training each of decision tree in the forest with the random subset of data is known as Bagging (Bootstrap Aggregation). Each Decision tree is set of internal nodes and leaves. Each of the internal node is a feature, based on which the dataset is split. And, the feature at each node is selected based on certain criterion, known as Gini index. Gini index is the measure of impurity at each node. Features that lead to highest decrease in the impurity are selected at each node. Each of these individual trees makes a prediction of the target class and the class with the maximum votes is elected for the final prediction of the model as shown in Figure 3.1. Random selection of the dataset at each independent decision trees, makes the model more generalised and ensures that the collective prediction attains the highest accuracy compared to single models prediction. The other most advantage is the speed of computation, where all the required values are computed at the training.

Diagram

Description automatically generated

Figure 0.1Simple Random Forest Classifier

Interpretability:

Model interpretability can be explained easily by traversing from the root node through all its edges till the leaf node. The leaf node tells the predicted outcome and all the edges are connected through “AND”. For example: In the Figure 3.2, with two features X1=”Does it Swim” , X2=”Have four legs”, to predict the outcome that ‘x’ is a ‘bird’, the decision is explained by its edges connecting the root node till the leaf node, if (‘Swim’==’No’ AND ‘have four legs’== ‘No’).

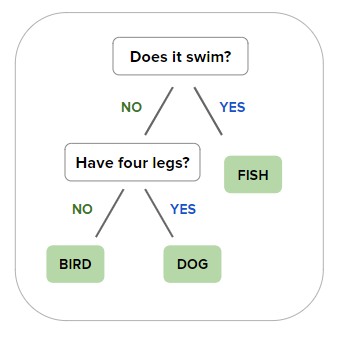


Figure 0.2 Interaction of features with the outcome

Feature importance:

Another important advantage of random forest is its ability to measure the contribution of features towards the prediction. The overall contribution of the feature in the decision tree can be computed by going through all the splits where the feature was used and measure by how much it has reduced the impurity (Gini index) (1) compared to the parent node. The sum of these importance is scaled at 100 and the average on all the decision trees in the forest computes the overall importance of feature in the model. Scikit-learn provides the method, “feature\_importances\_”, to easily get the importance of the features in the form of scores. The ones with the highest score are considered as important features.

Gini = 1 - ∑i=1𝑐(pi)2 ------------ (1)

Where, c is the number of class variables, pi is the probability of each class.

In this study of binary classification, c=2, and the probabilities will be, p(readmitted), p(not readmitted)

There are other ways of computing the feature importance using model agnostic methods, Permutation Feature Importance and SHAP as explained below.

* Permutation Feature Importance

Permutation based feature importance is used to overcome the drawback of default feature selection with mean decrease in Gini index. The concept here is to find the importance based on the increase in prediction’s error caused by permuting the features. A feature is considered to be “important” if the model’s prediction error increases with shuffling the values of features, because model relies on this feature. An “unimportant” feature is one where the model prediction does not change with the shuffling of values. It is implemented in scikit-learn as a method, “permutation\_importance”. The drawback with this is high cost of computation and highly co-related features get dropped as unimportant.

perm\_importance = permutation\_importance(rf, X\_test, y\_test), where ‘rf’ is a classifier, ‘X\_test’ and ‘y\_test’ is the test data.

* SHAP (SHapley Adaptive exPlanations)

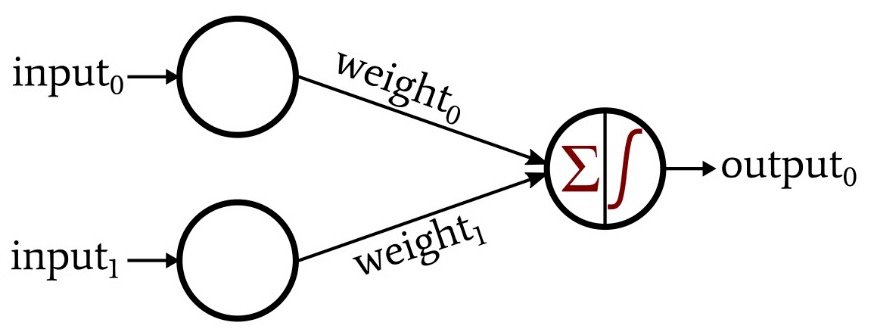
SHAP method works by computing the shapley values from game theory to estimate the contribution of each feature towards the prediction. This can be easily installed using ‘pip install shap’ and used with Scikit-learn Random Forest classifier. More details in the next section.

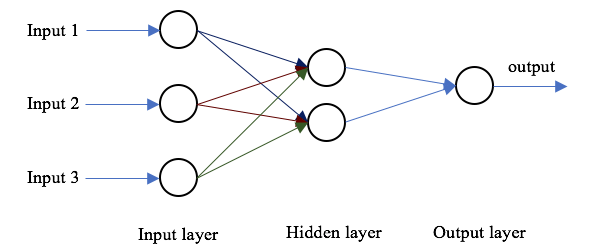
explainer = shap.TreeExplainer(rf)

shap\_values = Explainer.shap\_values(X\_test)

1. Multi-layer Perceptron:

Artificial Neural Networks (ANN) are one of the most powerful classifiers because of its ability to process large amount of training data and reduce the diagnosis time, resulting in high performance. The basic form of ANNs with one input layer and output layer is the single-layer perceptron (SLP), as shown in Figure 3.3 (a), but it is proved that SLP cannot handle non-linearly separable parameters. To handle the non-linearity, hidden layers are added between input and output layer, with each layer comprising of one or more neurons and each layer fully connected to next layer as in Figure 3.3 (b). This multilayer network is the Multilayer perceptron.

(a)Single layer perceptron



(b) Multilayer perceptron

Figure 0.3 Single layer perceptron, multilayer perceptron

MLP is a class of feedforward artificial neural networks, which accepts multidimensional input data, processes it through a weighted linear sum, followed by non-linear activation function and transforms the values to output layer to classify the output values (2). Multilayer perceptrons are often used in supervised learning problems, where it trains the model by building the relationship between input-output pairs. Training involves adjusting the parameters, weights and biases to minimise the error. Backpropagation is used to adjust the weights and bias relative to the error, calculated by Root Mean Squared Error. This results in back and forth propagation and continues util the error, no longer goes down. The performance of the model can be varied with number of layers added, adding more neurons and selecting the proper activation function.

## **Model Agnostic Interpretable Methods**

Interpretability of a model is finding the “contributions” or “relevance” or “attributions” of each input feature towards the prediction outcome. Attribution methods can be classified as gradient based attribution or perturbation based attribution. Gradient-based methods compute the attributions by some forward and backward propagations while perturbation based methods perturb the input and compute the changes in the output with respect to original input. Consider the network trained with several input features, the goal of attribution method is to find the real value of input feature corresponding to the target neuron of interest. A target neuron is identified by the tensors that gets activated at hidden or output layer. For a binary classification task, the target neuron might be the neuron corresponding to the correct class or the neuron of misclassified class. When the attributions of these target neurons are arranged together to have same shape as the input it becomes a attribution map. The attribution map helps to visualize the features contributing towards the prediction of the model. Figure 3.4 shows the explanation pipelines of interpretable model.

Black-box Model

user

Explainable Method

Figure 0.4 Explanation pipeline

Model Agnostic interpretability Methods proposed for this study:

1. Shapley values :

Shapley values -- a method from coalitional game theory -- tells us how to fairly distribute the prediction among the features. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. As the number of features increases, number of possible coalitions increases leading to expensive computation. (Interpretable Machine Learning, 2021)

1. SHAP (Shapley Additive explanations):

SHAP is a solid theoretical foundation in game theory. The prediction is fairly distributed among the feature values. We get contrastive explanations that compare the prediction with the average prediction. SHAP is good for tree based models unlike Shapley values with high computational cost. With the fast computation of Shapley values, Global interpretation can be easily achieved, holding the three most important properties of Additive feature Attribution 1. Local accuracy 2. Missingness 3. Consistency (Interpretable Machine Learning, 2021).

1. LIME :

Local Interpretable Model Agnostic Explanation model (LIME) are used explain the individual predictions of the black-box models. LIME as such was presented in the paper (Ribeiro et al., 2016) where the authors proposed a concrete implementation of local surrogate model. LIME tests the difference in the predictions with varied data set. So, it generates new dataset with perturbed samples in the neighbourhood of the sample of interest and the corresponding predictions from black-box. LIME trains interpretable models on this new dataset. Interpretable models can either be a Linear regression, Lasso or decision tree as shown in figure 3.4. The learned model should provide good approximation of machine learning predictions locally, even if not globally. This accuracy is called the local fidelity and one of property of interpretable models. The main advantage of LIME is, it can use a good explanatory model to explain the predictions even if it is not used for prediction and its ability to provide short descriptions of the decisions. (Interpretable Machine Learning, 2021). This is mainly helpful to health-practitioners to understand the results quickly.

LIME process can be explained as:

* Select the sample of interest for which you want to predict and find the explanation using black box model.
* Generate new dataset with perturbed samples.
* Get the predictions of black box for this new dataset.
* Assign weights to these new samples according to their proximity to the instance of interest.
* Train the interpretable model with the weighted new dataset with the variations.
* Explain the prediction by interpreting the local model.

Diagram

Description automatically generated

Figure 0.5 Steps of LIME [https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe]

1. DeepLIFT (Deep Learning Important FeaTures):

(Shrikumar et al., n.d.) DeepLIFT is a gradient-based attribution method for explaining the output prediction of neural networks, by assigning the scores to the input variables. DeepLIFT is implemented to overcome the saturation problem faced by LIME and other gradient based methods like Gradient \* Inputs, Integrated gradients, Saliency maps and Layerwise Relevance Propagation(LRP). Gradient based methods assigns scores by computing the gradients at each layer in the back propagation and establishing the relationship of the output and the neurons at layers below it as shown in Figure 3.6. This helps in finding which inputs are more sensitive to the output. This is the generic approach followed any other gradient based methods which leads to saturation issues(gradient is zero) as shown in Figure 3.7. DeepLIFT works around this problem by finding the difference in activation of each neuron to the “reference activation”, i.e. gradient \* Δinput. Reference activation are assigned during the forward propagation and the difference input is the average of all possible order of input. The feature of “reference” outperforms this model from other gradient-based methods, allowing the information to flow even when the gradient is zero.

DeepLIFT is pypi package, it can be easily installed using ‘pip install deeplift”, DeepLift provides auto conversion functions to convert models trained using Keras into DeepLIFT format. In this study "NonlinearMxtsMode.Rescale” is used to find the deeplift\_model. Then the contribution function for the deeplift model is obtained and the feature scores are computed using method from Keras.

Figure 0.6 Backpropagation approach, yellow=input, green=output

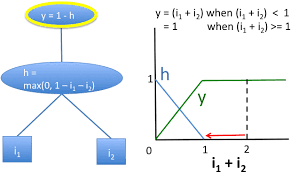


Figure 0.7 saturation problems faced by perturbations and gradient based methods

## **Proposed Method**

The purpose of this study is to build a model with high performance and explain the model showcasing the importance of features contributing towards the risk of readmission with diabetes patients. In this study we used the Diabetes dataset collected from 130 US hospitals for 10years. Dataset is pre-processed including class-balancing with SMOTE, Random forest for feature selection. This study selected two classifiers, (1) Random Forest, which provides better accuracy along with feature importance to explain the model. Random Forest with Gini is chosen due to binary classification task of this study. (2) Multilayer perceptron, with an input layer, one hidden layer and output layer. ReLU (Rectified Linear Unit) or PReLU (Parametric ReLU) activation function is used at the input layer, based on the performance to introduce non-linearity in the model. At the output layer, Sigmoid activation is used because of the binary classification problem that will be addressed in this study. The model is trained to reduce the loss and regularized by adding a Dropout after each of the input layers, avoiding overfitting of the models.

A model is split into training and test data with 80:20.

Random Forest is trained using Scikit-learn methods while MLP model is built using Tensflow API and DeepLIFT is used to explain the model predictions. LIME, SHAP are used with the machine learning model, Random Forest, to find the contribution of features towards prediction. Figure 3.7 shows the architecture of interpretable model, showing the pre-processing applied, model building and methods applied to explain the predictions.



start

Load the data from the csv file

Training data (80)

Test data (20)

Model Agnostic Interpretable methods

Split the dataset 80:20

Balance data

Data reduction

Modelling Random Forest /MLP

Feature Engineering

Pre-process data

Missing value imputation

Data Cleaning

Model Evaluation

Precision-recall

Accuracy

AUROC

Evaluate the features

Important features with explanation

Explainable visualization

Figure 0.8 Architecture of Explainable Model

## **Evaluation**

The prediction of the model is evaluated based on Model accuracy, precision, recall, AUC- ROC. In this study of medical science, recall is important because hospitals are penalized for the misclassification of patients who are likely to get readmitted, which incurs high cost to hospitals.

Confusion matrix is shown in Table 3.2, True Positive (TP) True Negative (TN) False Positive (FP) False Negative (FN).

Table 3.2: Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Predicted/Actual | Positive | Negative |
| Positive | True Positive | False Positive |
| Negative | False Negative | True Negative |

Accuracy: Accuracy is the basic and most commonly used evaluation metrics to measure performance of model. It is the ratio of number of correct predictions to the total number of input samples.

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

Recall: Recall is the ratio of true positives (items that are correctly labelled to positive class) to the total number of elements that actually belong to positive class.



Recall = (TP)/(TP+FN)

Precision: Precision is the ratio of true positives (items that are correctly labelled to positive class) to the total number of elements labelled as belonging to positive class.

Precision = (TP)/(TP+FP)

AUROC (Area Under the Receiver Operating Characteristics): The Area Under the Curve (AUC) specifies the ability of the model to distinguish between the classes (positive and negative). Higher the AUC score , better is the models prediction. AUC was the preferred metric of choice due to being used by the majority of the selected studies (75%) as shown literature review (Artetxe et al., 2018) and this study is with binary classification. AUC=1, says the model provides 100% separability and AUC=0, is the worst model with zero separability, which means it is completely misclassifying the results. In this study of readmission prediction, accuracy measurement can lead to biased results, a model with 90% accuracy can predict the majority class with 90% accuracy, but minority ones have high chances of misclassification. Since most of the time, minority are the samples of interest, so with this study, misclassification can incur high cost to patients and health providers. So, AUROC is considered as performance measurement method in this study.

(Doshi-Velez and Kim, 2017) has categorized three main factors for the evaluation of interpretable model.

* Application-ground level

Evaluating the explanations generated by the model as human-friendly explanations by letting the domain experts use the model and analyze the explanations. This involves high cost and time.

* Human-ground level

Human-ground evaluation involves humans but not experts, where lay humans evaluate the explanations without the end goal of the task. This only checks for the quality of the explanations, regardless of the correctness of associated prediction. This is only good to evaluate a generic explanation.

* Functional-ground level

This does not require a human to evaluate the performance or the explanations, considering the model is already tested with human-ground explanations, so only in cases where performance tuning is done, functional level evaluation is performed.

This study evaluates the explanations by comparing the factors and explanations provided by other researchers which are evaluated at Human level.

## **Summary**

Build a model using the diabetes dataset, with pre-processing data, applying feature engineering, class balancing and feature selection techniques. Use two models to Random Forest and MLP and derive the risk factors using model agnostic interpretable methods and compare the resulting feature importance.

# ANALYSIS

## **Introduction**

Model Building requires the data to be clean and consistent. Data Analysis is important part of model building to achieve better performance of the model. In this section, the details of each of the variables are explained to understand the data better. The implementation is performed in python using Tensorflow API.

## **Dataset Description**

The dataset consists of 101766 records with 50 attributes. It has mix of numeric and categorical columns. Figure 4.1 shows some of the features involved in the dataset. The target variable of the dataset is ‘readmitted’ and the aim of this study is to identify the patients at high risk of readmission. There are 3 classes, patients readmitted in less than 30days, greater than 30days and not readmitted. This study focuses on readmission less than 30days, so only 2 class is considered, greater than 30days are considered as not readmitted. With this, we get highly imbalanced dataset (0:90409, 1:11357).

|  |  |  |  |
| --- | --- | --- | --- |
| Feature name | Type | Description | %Missing |
| Encounter Id | Numeric | Unique identifier of an encounter | 0% |
| Patient number | Numeric | Unique identifier of a patient | 0% |
| Race | Nominal | Values: Caucasian, Asian, African American, Hispanic, and other | 2% |
| Gender | Nominal | Values: male, female, unknown/invalid | 0% |
| Age | Nominal | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) | 0% |
| Weight | Numeric | Weight in pounds | 97% |
| Admission type | Nominal | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available | 0% |
| Discharge disposition | Nominal | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available | 0% |
| Admission source | Nominal | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital | 0% |
| Time in hospital | Numeric | Integer number of days between admission and discharge | 0% |
| Payer code | Nominal | Integer identifier corresponding to 23 distinct values, for example, Blue Cross*/*Blue Shield, Medicare, and self-pay | 52% |
| Medical specialty | Nominal | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family*/*general practice, and surgeon | 53% |
| Number of Lab Procedures | Numeric | Number of lab tests performed during the encounter | 0% |
| Number of procedures | Numeric | Number of procedures (other than lab tests) performed during the encounter | 0% |
| Number of medications | Numeric | Number of distinct generic names administered during the encounter | 0% |
| Number of outpatient visits | Numeric | Number of outpatient visits of the patient in the year preceding the encounter | 0% |
| Number of emergency visits | Numeric | Number of emergency visits of the patient in the year preceding the encounter | 0% |
| Number of inpatient visits | Numeric | Number of inpatient visits of the patient in the year preceding the encounter | 0% |
| Diagnosis 1 | Nominal | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values | 0% |
| Diagnosis 2 | Nominal | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values | 0% |
| Diagnosis 3 | Nominal | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values | 1% |
| Number of Diagnosis | Numeric | Number of diagnoses entered to the system | 0% |
| Glucose serum test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured | 0% |
| A1c test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured. | 0% |
| Change of medical | Nominal | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” | 0% |
| Diabetes medication | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” | 0% |
| 24 features of medication | Nominal | For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage.  Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed | 0% |
| Readmitted | Nominal | Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission. | 0% |

Figure 4.1Dataset description with missing percentage

Graphical user interface, application

Description automatically generated

Figure 4.2 Snippet of the dataset

## **Exploratory Data Analysis**

There are two numeric variables, ‘patient\_nbr’ and ‘encounter\_id’ which are patients identity information and don’t contribute to the study, so drop these columns.

Figure 4.2 shows the distribution plot of the data, some skewness is observed for ‘number\_inpatient’, ‘number\_outpatient’ and ‘number\_emergency’.

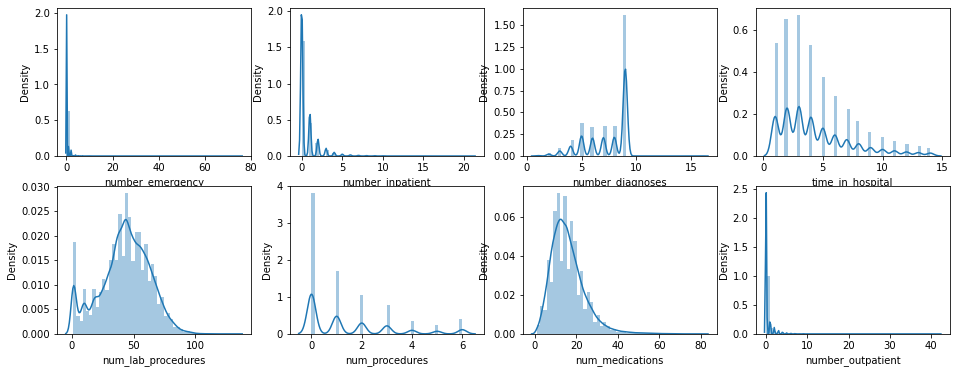


Figure 4.3 Data distribution to find the skewness

### **Univariate Analysis**

Analyse each of categorical variables influence on the target variable by plotting the ‘bar’ graphs. Medication Features like ‘troglitazone’, ‘metformin-pioglitazone’, ‘metformin-rosiglitazone’, ‘glimepiride-pioglitazone’, ‘citoglipton’, ‘examide’ shows all values as ‘No’, which indicates no impact of the target variable. So the entire columns can be safely removed. In the patients, where the ‘migliton’, ‘pioglitazone’, ‘glipizide’, ‘glyburide-metformin’ and ‘acarbose' medication was reduced, it showed high chances of readmission as shown in Figure 4.3.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Figure 4.4 High readmission risk when there is reduction in medication

For numerical variables, correlation matrix is plotted to check the variables highly co-related to target variable and if any pair of variables are highly co-related. There are no such variables in the dataset. Why should highly correlated variables be removed?

### **Missing value Imputation**

As shown in Figure 4.1, variables with high percentage of missing value are ‘weight’, ‘payer\_code’ and ‘medical specialty”. ‘weight’ is removed entirely and ‘medical\_specialty’ is replaced with ‘missing’. Other variables with low missing values, ‘diag\_1’, ‘diag\_2’, ‘diag\_3’ are imputed using imputation method, forward fill,‘ffill’. Any other imputation methods tried? Whats the result? What for payer\_code?

### **Class Imbalance**

The target variable of this study is ‘readmitted’. This column is categorical column with the data distributed among three labels, readmitted in less than 30 days (‘<30’), readmitted in greater than 30days (‘ >30’) and not readmitted (‘No’). For the purpose of this study, readmitted patient in greater than 30days are considered as Not readmitted. With this we have 2 class labels. Encode the variable with ‘0’ and ‘1’.

As shown in Figure 4.5 the data is highly class imbalanced. In case of medical data it is expected, because the chances of patients getting ‘readmitted’ (class 1) is always less compared to ‘not readmitted’ (class 0). Since the study focuses on identifying the readmitted patients, which are minority samples, balancing the data is important. Imbalanced data leads to bias in the model performance, that is, a model can give 95% accuracy in predicting class 0, but it fails to predict class 1.

Scatter chart, qr code

Description automatically generated

Figure 4.5 Class Imbalance with Counter({0: 90409, 1: 11357})

### **Feature Engineering**

Feature Engineering is required to enrich the dataset with relevant information, combining the data spread across multiple tables.

For a patient, the level of service utilized depends on ‘number\_inpatient’, ‘number\_outpatient’ and ‘number\_emergency’, sum of all these will give the total service utilized. This can be a good predictor of readmittance in patients. As shown in Figure 4.6, most of the patients have service\_utilization in the range of 5-10.

There are 24 medications, a patient is prescribed with one or more than one medication. The medication is either continued, changed, or stopped. So, each of the medication is categorised as ‘Steady’, ‘No’, ‘Up’, ‘Down’. A change in medication can be identified as one entity by summing the ‘Up’ and ‘Down’ as ‘medication\_change’. Similarly a new variable is derived by considering total medication given to the patient, as sum of ‘Steady’, ‘Up’ and ‘Down’. Figure 4.7 shows the newly derived variables and their distribution. ‘service\_utilization’ looks skewed.

A picture containing shape

Description automatically generated

(a)

Chart, scatter chart

Description automatically generated

(b)

Figure 4.6 (a) summary of the service\_utilization distribution (b) service\_utilzation w.r.t readmitted

Chart, histogram

Description automatically generated

Figure 4.7 Distribution of newly added features

Chart, histogram

Description automatically generated

Figure 4.8 Variables after log transformation

Log transformation is one of the most commonly used method to remove the skewness of the variables. Log(x) of number between 0 and 1 lead to NaN, so considered using log(x+1) and Figure 4.8 shows the variables after the log transformation. What happens with log transformation?

After the log transformation, standardize the data, where each of the numerical values are scaled to bring them on the same scale. After the standardization, the all the features are distributed with mean=0 and standard deviation as shown in Figure 4.9 shows Why?

Table

Description automatically generated

Figure 4.9 Data distribution with mean=0 and standard deviation 1

### **Outlier Removal**

### **Data Reduction**

Features like, ‘diag\_1’, ‘diag\_2’, ‘diag\_3’ have ICD9 Codes which are difficult to interpret and are clustered to 9 categories. 'admission\_type\_id' has 8 unique values. Based on the description, converged them to 4 categorize.

'medical\_specialty' has 73 unique values. Based on their semantic, surgery related were mapped to ‘Surgery’, pediatrics related were mapped to ‘Pediatrics’, replace ‘PhysicianNotFound as ‘Missing’, replace 'Family/GeneralPractice', 'Hospitalist', 'InfectiousDiseases' as 'InternalMedicine', replace 'Orthopedics-Reconstructive', 'SportsMedicine', 'Rheumatology', as 'Orthopedics', replace 'Perinatology', 'Obsterics&Gynecology-GynecologicOnco', 'ObstetricsandGynecology', 'Gynecology' as 'Obstetrics', replace 'Radiologist' as 'Radiology'. Other with less than 40% records are replaced as ‘Others’.

‘admission\_source\_id' has 26 distinct values. Replace ‘Not Avialable’, ‘NULL’, ‘Not mapped’, ‘Unknown/invalid’ to 9(‘Not available’). Replace 10, 25 with 7 (Emergency Room). Replace 5, 6, 22 with 4 (Transfer from Hospital). Replace 3 with 1 (Physicial Referral). Replace others as 27(Others).

'discharge\_disposition\_id' has 29 distinct values. Replace 3, 4, 5, 6, 8, 10, 15, 16, 17, 22, 23, 24, 27, 28, 29, 30, with 2(‘Discharged/transferred to another short term hospital’).

### **Feature Encoding**

Only few models like Decision Trees like the categorical variables, RandomForest and Neural Network expects the variables to be numeric, so encoding of each categorical variables is ra must. There are 37 categorical variables, of which 7 got dropped (‘weight’, ‘troglitazone’, ‘metformin-pioglitazone’, ‘metformin-rosiglitazone’, ‘glimepiride-pioglitazone’, ‘citoglipton’, ‘examide’).

‘Age’ is given as a range, it is encoded by considering the mid value. ‘gender’, ‘diabetesMed’ and ‘change’ are holding two unique values, so encoded with ‘0’ and ‘1’. For the rest of categorical variables, created dummy variables. With this all variables were converted to numeric. Why dummies? What about other encoding?

### **Feature selection**

Chart

Description automatically generated with low confidence

Figure 4.10 Top25 features contribution to prediction

## **Random Forest Hyper parameter tunning**

Random Forest: max\_features

**Chart, line chart

Description automatically generated**

Chart, line chart

Description automatically generated

N\_estimators

Chart, line chart

Description automatically generated

## **Neural Networks Hyper parameter tunning**

model4.fit(X\_train, y\_train, epochs=20, validation\_data=(X\_test,y\_test),

batch\_size=16)

## **Summary**

# RESULTS AND DISCUSSIONS

## **Class Imbalance**

Data before class imbalance

Data after class imbalance

## **Modelling**

### **Random Forest prediction**

Random Forest with Label Encoding – selecting 25 features

Accuracy\_score: 0.8811213448084957

AUC\_accu = 0.5274182178583818

Precision\_score = 0.3801452784503632

recall\_score = 0.06937693327441449

### **Multilayer perceptron prediction**

Table

Description automatically generated

## **Explain the Predictions of each instance**

### **Random Forest**

# Shap

**Chart

Description automatically generated**

A picture containing timeline

Description automatically generated

# Lime

# For instance = 15230

# Chart, waterfall chart Description automatically generated with medium confidence

### **Multilayer Perceptron**

# DeepExplainer

# DeepLIFT

# CONCLUSIONS AND RECOMMENDATIONS

# REFERENCES

Ahmed, M. and Kabir, M.M.J., (2019) Dimensionality reduction in predicting hospital readmissions of diabetics. *2019 5th International Conference on Advances in Electrical Engineering, ICAEE 2019*, September 2019, pp.155–159.

Alehegn, M., Joshi, R.R. and Mulay, P., (2019) Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach. *International Journal of Scientific and Technology Research*, 89, pp.1346–1354.

Alvarez-Melis, D. and Jaakkola, T.S., (n.d.) *On the Robustness of Interpretability Methods*.

Anon (2021) *Hospital Readmissions Reduction Program (HRRP) | CMS*. [online] Available at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program [Accessed 9 Jan. 2021].

Anon (2021) *Interpretable Machine Learning*. [online] Available at: https://christophm.github.io/interpretable-ml-book/ [Accessed 17 Jan. 2021].

Anon (2021) *National Diabetes Statistics Report, 2020 | CDC*. [online] Available at: https://www.cdc.gov/diabetes/data/statistics-report/index.html [Accessed 6 Mar. 2021].

Anon (2021) *Transitional Care Interventions Prevent Hospital Readmissions For Adults With Chronic Illnesses | Enhanced Reader*. [online] Available at: chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?pdf=https%3A%2F%2Fbrxt.mendeley.com%2Fdocument%2Fcontent%2F5f050702-9e3a-3050-b8cd-d78728cedcdd&doi=10.1377/hlthaff.2014.0160 [Accessed 21 Feb. 2021].

Artetxe, A., Beristain, A. and Graña, M., (2018) *Predictive models for hospital readmission risk: A systematic review of methods*. *Computer Methods and Programs in Biomedicine*, .

Barr Kumarakulasinghe, N., Blomberg, T., Liu, J., Saraiva Leao, A. and Papapetrou, P., (2020) Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In: *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. Institute of Electrical and Electronics Engineers Inc., pp.7–12.

Bhuvan, M.S., Kumar, A., Zafar, A. and Kishore, V., (2016) Identifying Diabetic Patients with High Risk of Readmission. [online] Available at: http://arxiv.org/abs/1602.04257.

Bologna, G. and Hayashi, Y., (2017) Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning. *Journal of Artificial Intelligence and Soft Computing Research*, 74, pp.265–286.

Canadian Institute for Health Information, (2013) Early Identification of People At-Risk of Hospitalization. [online] p.15. Available at: https://secure.cihi.ca/free\_products/HARP\_reportv\_En.pdf.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., (2015) Intelligible Models for HealthCare. pp.1721–1730.

Caughey, G.E., Pratt, N.L., Barratt, J.D., Shakib, S., Kemp-Casey, A.R. and Roughead, E.E., (2017) Understanding 30-day re-admission after hospitalisation of older patients for diabetes: Identifying those at greatest risk. *Medical Journal of Australia*, 2064, pp.170–175.

Chawla, N. V, Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2002) *SMOTE: Synthetic Minority Over-sampling Technique*. *Journal of Artificial Intelligence Research*, .

Chen, J.Y., Ma, Q., Chen, H. and Yermilov, I., (2012) New bundled world: Quality of care and readmission in diabetes patients. *Journal of Diabetes Science and Technology*, [online] 63, pp.563–571. Available at: /pmc/articles/PMC3440042/ [Accessed 14 Feb. 2021].

Chi, C.-Y., Winkler, A., Xu, J., Tomberg, A., Soltani, R. and Huang, C.-H., (2020) Explainable model of deep learning for outcomes prediction of in-hospital cardiac arrest patients. *Resuscitation*, [online] 155, p.S19. Available at: https://doi.org/10.1016/j.resuscitation.2020.08.063.

Chin, M.H., Zhang, J.X. and Merrell, K., (1998) Diabetes in the African-American medicare population: Morbidity, quality of care, and resource utilization. *Diabetes Care*, [online] 217, pp.1090–1095. Available at: https://pubmed.ncbi.nlm.nih.gov/9653601/ [Accessed 15 Feb. 2021].

Collins, J., Abbass, I.M., Harvey, R., Suehs, B., Uribe, C., Bouchard, J., Prewitt, T., DeLuzio, T. and Allen, E., (2017) Predictors of all-cause 30 day readmission among Medicare patients with type 2 diabetes. *Current Medical Research and Opinion*, [online] 338, pp.1517–1523. Available at: https://pubmed.ncbi.nlm.nih.gov/28498094/ [Accessed 15 Feb. 2021].

Doshi-Velez, F. and Kim, B., (2017) Towards A Rigorous Science of Interpretable Machine Learning. [online] Ml, pp.1–13. Available at: http://arxiv.org/abs/1702.08608.

Eby, E., Hardwick, C., Yu, M., Gelwicks, S., Deschamps, K., Xie, J. and George, T., (2015) Predictors of 30 day hospital readmission in patients with type 2 diabetes: A retrospective, case-control, database study. *Current Medical Research and Opinion*, [online] 311, pp.107–114. Available at: https://pubmed.ncbi.nlm.nih.gov/25369567/ [Accessed 16 Feb. 2021].

Emons, M.F., Bae, J.P., Hoogwerf, B.J., Kindermann, S.L., Taylor, R.J. and Nathanson, B.H., (2016) Risk factors for 30-day readmission following hypoglycemia-related emergency room and inpatient admissions. *BMJ Open Diabetes Research and Care*, 41.

Enomoto, L.M., Shrestha, D.P., Rosenthal, M.B., Hollenbeak, C.S. and Gabbay, R.A., (2017) Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. *Journal of Diabetes and its Complications*, 311, pp.122–127.

Friedman, B., Jiang, H.J. and Elixhauser, A., (2008) Costly hospital readmissions and complex chronic illness. *Inquiry : a journal of medical care organization, provision and financing*, [online] 454, pp.408–21. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19209836 [Accessed 6 Mar. 2021].

Futoma, J., Morris, J. and Lucas, J., (2015) A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56, pp.229–238.

Gek Sang Soh, J., Pong Wong, W., Mukhopadhyay, A., Chye Quek, S. and Choo Tai, B., (2020) Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis. *BMJ Open Diab Res Care*, [online] 8, p.1227. Available at: http://drc.bmj.com/.

Ghorbani, R. and Ghousi, R., (2020) Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques. *IEEE Access*, 8, pp.67899–67911.

Golmohammadi, D. and Radnia, N., (2016) Prediction modeling and pattern recognition for patient readmission. *International Journal of Production Economics*, 171, pp.151–161.

Goudjerkan, T. and Jayabalan, M., (2019) Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron. *International Journal of Advanced Computer Science and Applications*, 102, pp.268–275.

Graham, E., Saxena, A. and Kirby, H., (2019) Identifying High Risk Patients for Hospital Readmission. *SMU Data Science Review*, [online] 21. Available at: https://scholar.smu.edu/datasciencereview/vol2/iss1/22.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F. and Giannotti, F., (2018a) Local rule-based explanations of black box decision systems. *arXiv*, May.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D., (2018b) A survey of methods for explaining black box models. *ACM Computing Surveys*, 515.

Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., (n.d.) *Explainable AI for the Medical Domain What do we need to build explainable AI systems for the medical domain?* [online] Available at: https://www.computerworld.com.au/article/617359.

Hu, P., Li, S., Huang, Y.A. and Hu, L., (2019) Predicting hospital readmission of diabetics using deep forest. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, pp.2019–2020.

Hu, Y. and Sokolova, M., (n.d.) Explainable Multi-class Classification of Medical Data. pp.1–21.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M. and Kripalani, S., (2011) Risk prediction models for hospital readmission: A systematic review. *JAMA - Journal of the American Medical Association*, 30615, pp.1688–1698.

Kursa, M.B. and Rudnicki, W.R., (2010) Feature selection with the boruta package. *Journal of Statistical Software*, 3611, pp.1–13.

Lakkaraju, H., Bach, S.H. and Leskovec, J., (2016) Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, pp.1675–1684.

Lin, W.J. and Chen, J.J., (2013) *Class-imbalanced classifiers for high-dimensional data*. *Briefings in Bioinformatics*, Available at: https://pubmed.ncbi.nlm.nih.gov/22408190/ [Accessed 18 Feb. 2021].

Lou, Y., Caruana, R. and Gehrke, J., (2012) Intelligible models for classification and regression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.150–158.

Lou, Y., Caruana, R., Gehrke, J. and Hooker, G., (2013) Accurate intelligible models with pairwise interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1288, pp.623–631.

Loyola-González, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A. and García-Borroto, M., (2016) Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175, pp.935–947.

Lu, Y., Cheung, Y.M. and Tang, Y.Y., (2019) Bayes imbalance impact index: A measure of class imbalanced dataset for classification problem. *arXiv*, 319, pp.3525–3539.

Lundberg, S.M., Allen, P.G. and Lee, S.-I., (n.d.) *A Unified Approach to Interpreting Model Predictions*. [online] Available at: https://github.com/slundberg/shap.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J. and Lee, S.I., (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, [online] 210, pp.749–760. Available at: https://pubmed.ncbi.nlm.nih.gov/31001455/ [Accessed 20 Mar. 2021].

McCoy, R.G., Lipska, K.J., Herrin, J., Jeffery, M.M., Krumholz, H.M. and Shah, N.D., (2017) Hospital Readmissions among Commercially Insured and Medicare Advantage Beneficiaries with Diabetes and the Impact of Severe Hypoglycemic and Hyperglycemic Events. *Journal of General Internal Medicine*, [online] 3210, pp.1097–1105. Available at: https://pubmed.ncbi.nlm.nih.gov/28685482/ [Accessed 15 Feb. 2021].

Mingle, D., (2017) Predicting Diabetic Readmission Rates: Moving Beyond Hba1c. *Current Trends in Biomedical Engineering & Biosciences*, 73.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T., (2017) Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, [online] 196, pp.1236–1246. Available at: /pmc/articles/PMC6455466/ [Accessed 26 Feb. 2021].

Modzy, A.R. and Modzy, A.T., (n.d.) *An Adversarial Approach for Explaining the Predictions of Deep Neural Networks*. [online] Available at: www.aaai.org.

Napierala, K. and Stefanowski, J., (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 463, pp.563–597.

Ogedengbe, M.T. and Egbunu, C.O., (2020) CSE-DT Features selection technique for Diabetes classification. *Applications of Modelling and Simulation*, 4March, pp.101–109.

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., MacDonell, C. and Anvik, J., (2006) Visual explanation of evidence in additive classifiers. *Proceedings of the National Conference on Artificial Intelligence*, 2January, pp.1822–1829.

Reed, R.A., Morgan, A.S., Zeitlin, J., Jarreau, P.-H., Torchin, H., Pierrat, V., Ancel, P.-Y. and Khoshnood, B., (2021) Machine-Learning vs. Expert-Opinion Driven Logistic Regression Modelling for Predicting 30-Day Unplanned Rehospitalisation in Preterm Babies: A Prospective, Population-Based Study (EPIPAGE 2). *Frontiers in Pediatrics*, [online] 8. Available at: /pmc/articles/PMC7886676/ [Accessed 23 Feb. 2021].

Reid, C., (2019) Diabetes Diagnosis and Readmission Risks Predictive Modelling: USA MSc Research Project MSc in Data Analytics. [online] Available at: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html.

Ribeiro, M.T., Singh, S. and Guestrin, C., (2016) ‘Why should i trust you?’ Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, pp.1135–1144.

Robbins, J.M. and Webb, D.A., (2006) Diagnosing diabetes and preventing rehospitalizations: The urban diabetes study. *Medical Care*, 443, pp.292–296.

Robbins, T.D., Lim Choi Keung, S.N., Sankar, S., Randeva, H. and Arvanitis, T.N., (2019) *Risk factors for readmission of inpatients with diabetes: A systematic review*. *Journal of Diabetes and its Complications*, .

Rubin, D.J., Handorf, E.A., Golden, S.H., Nelson, D.B., McDonnell, M.E. and Zhao, H., (2016) Developement and validation of a novel tool to predict hospital readmission resk among patients with diabetes. *Endocrine Practice*, [online] 2210, pp.1204–1215. Available at: https://pubmed.ncbi.nlm.nih.gov/27732098/ [Accessed 14 Feb. 2021].

Rubin, D.J., Recco, D., Turchin, A., Zhao, H. and Golden, S.H., (2018a) External validation of the diabetes early re-admission risk indicator (derriTM). *Endocrine Practice*, 246, pp.527–541.

Rubin, D.J., Recco, D., Turchin, A., Zhao, H. and Golden, S.H., (2018b) External validation of the diabetes early re-admission risk indicator (derriTM). *Endocrine Practice*, 246, pp.527–541.

Rudin, C., (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 15, pp.206–215.

Sarthak, Shukla, S. and Prakash Tripathi, S., (2021) Embpred30: Assessing 30-days readmission for diabetic patients using categorical embeddings. *Advances in Intelligent Systems and Computing*, 1168February, pp.81–90.

Sidey-Gibbons, J.A.M. and Sidey-Gibbons, C.J., (2019) Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, [online] 191, p.64. Available at: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4 [Accessed 23 Feb. 2021].

Sonmez, H., Kambo, V., Avtanski, D., Lutsky, L. and Poretsky, L., (2017) The readmission rates in patients with versus those without diabetes mellitus at an urban teaching hospital. *Journal of Diabetes and its Complications*, 3112, pp.1681–1685.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K. and Cilar, L., (2020) Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 105, pp.1–13.

Štrumbelj, E. and Kononenko, I., (2010) An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, pp.1–18.

Swartout, W.R., (1983) XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*, 213, pp.285–325.

Tamin, F. and Iswari, N.M.S., (2017) Implementation of C4.5 algorithm to determine hospital readmission rate of diabetes patient. *Proceedings of 2017 4th International Conference on New Media Studies, CONMEDIA 2017*, 2018-Janua, pp.15–18.

Vellido, A., (2020) The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, [online] 3224, pp.18069–18083. Available at: https://link.springer.com/article/10.1007/s00521-019-04051-w [Accessed 20 Feb. 2021].

Visani, G., Chesani, F., Bagli, E., Capuzzo, D. and Poluzzi, A., (2020) Explanations of Machine Learning predictions: a mandatory step for its application to Operational Processes. [online] Available at: http://arxiv.org/abs/2012.15103 [Accessed 27 Feb. 2021].

Vucenovic, A., Ali-ozkan, O., Ekwempe, C. and Eren, O., (2020) Explainable AI in Decision Support Systems. pp.7–10.

Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A. Ben and Kronzer, A., (2018) Predicting Hospital Readmission via Cost-Sensitive Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 156, pp.1968–1978.

Yang, C., Delcher, C., Shenkman, E. and Ranka, S., (2016) Predicting 30-day all-cause readmissions from hospital inpatient discharge data. *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016*, pp.2–7.

Yu, K. and Xie, X., (2020) Predicting Hospital Readmission: A Joint Ensemble-Learning Model. *IEEE Journal of Biomedical and Health Informatics*, 242, pp.447–456.

Zhou, L., (2013) Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, 41, pp.16–25.

Zhou, Z.H. and Feng, J., (2017) Deep forest: Towards an alternative to deep neural networks. In: *IJCAI International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, pp.3553–3559.

# APPENDIX A: RESEARCH PLAN

# Chart Description automatically generated

# APPENDIX B: RESEARCH PROPOSAL

FEATURE ACCOUNTABILITY FOR 30-DAY HOSPITAL READMISSION PREDICTION FOR DIABETES PATIENTS

**Research Proposal**

Shridevi Bellatti

Liverpool John Moores University - Master’s in Data Science September, 2020

# Under the supervision of

Ankan Datta

# Abstract

A Hospital readmission is where a patient gets readmitted to the hospital post-discharge within a specified interval of time. Readmission is the key measure of the quality of treatment provided by the health centres to the patients during hospitalization and has proven to be immensely expensive. According to the Affordable care Act (ACT) 2010, hospitals are penalised with low reimbursement for the patients admitted within 30-days of discharge and the cost of readmission is estimated to be around $25 billion yearly in U.S hospitals. Studies have shown that patients with diabetes are more susceptible to readmission within 30-days of discharge. So, predicting the readmission is beneficial for patients and healthcare centres. Current practice of identifying diabetes patients for readmission is subjective, where the clinician will assess what should be the appropriate care given to the patients. But, research have shown that with the help of predictive modelling it is slightly better which learns the records of various patients. Considering this, the main goal of this study is to build a predictive classification model along with interpretation. This study benchmarks the existing machine learning models and proposes MultiLayer perceptron (MLP) with extensive pre-processing and transformation which includes feature selection using Random Forest feature selection, feature engineering, balancing the data using SMOTE to handle the noisy, inconsistent and imbalanced dataset of diabetes patients from 130 US Hospitals for the year 1999-2008. The proposed model will predict the readmission of diabetes patients within 30-days and derive meaningful insights of the attributes contributing towards the readmission by applying LIME, SHAP and DeepLIFT algorithms. Identifying the readmission of hospitalised diabetes patients, uncovering the features, helps healthcare centres to improvise the treatment and reduce the hospital expenditure. Model performance is evaluated using Accuracy, precision, recall, Area Under Receiver Operating characteristics (AUROC) and the interpretability of the model is evaluated, comparing the important features with the existing machine learning suggested features.

|  |  |
| --- | --- |
| [Abstract](#_bookmark0) | [2](#_bookmark0) |
| [1. Introduction](#_bookmark1) | [4](#_bookmark1) |
| [2. Background & Related research](#_bookmark2) | [5](#_bookmark2) |
| [3. Research Questions](#_bookmark3) | [7](#_bookmark3) |
| [4. Aim and Objectives](#_bookmark4) | [8](#_bookmark4) |
| [5. Significance of the Study](#_bookmark5) | [8](#_bookmark5) |
| [6. Scope of the Study](#_bookmark6) | [9](#_bookmark6) |
| [7. Research Methodology](#_bookmark7) | [9](#_bookmark7) |
| [8. Expected Outcomes](#_bookmark8) | [17](#_bookmark8) |
| [9. Requirements / resources](#_bookmark9) | [18](#_bookmark9) |
| [10. Research Plan](#_bookmark8) | [17](#_bookmark8) |
| [References](#_bookmark10) | [18](#_bookmark10) |

# Introduction

Diabetes has become an Epidemic in today’s world with moving towards Urbanisation and lifestyle of people around the world. The global diabetes prevalence has increased to 9.3% in 2019 and estimated to increase to 11% by 2045. This calls for a better hospital operational efficiency and better care of Diabetes patients. Readmission is one of measure of hospitals operational efficiency. So, predicting the 30-days readmission in Diabetes patients is of great importance. The literature talks about many models developed to increase the model’s prediction accuracy from various Machine Learning Models to more complex models and (Goudjerkan and Jayabalan, 2019) proved that Deep learning or Neural Networks can provide high accuracy. However, high accuracy is achieved using complex models like Deep Learning, ensembles model, trading off the model interpretability.

Model fairness and interpretability is required for data-scientists, researchers, developers to bring in value for their findings and support the accuracy. Interpretability is also important to debug machine learning models, understand the model bias and make informed decisions on how to improve them.

Model Interpretability is important in critical fields like medical, financial domains where not just the model prediction becomes important but also the need to understand why the model made this prediction is important. For example, by looking at the size of the cancer cell, if the models predicts that the patient if likely to have cancer, it’s a human tendency to understand what factors are contributing to the decision.

Some models like Logistic Regression, Linear Regression, decision trees, random forest exhibit the property of interpretability while other complex models like deep neural networks, so called black boxes requires explicit interpretability techniques to be used which are known as post- hoc interpretations.

In this paper, we will look at applying different methods of interpretability SHAP and DeepLIFT on the models using 130-US hospitals for years 1999-2008 Data Set, (UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set, 2021)

# Background & Related research

With high Risks involved with readmission of Diabetes patient, lot of research is done around predicting the readmission rate. At the same time, with more and more data available from Hospitals, many researchers have tried various models, but not all are successful. Let’s understand each of the research.

With the Initial study it was found the LACE index (Length of stay, acuity of admission, Charlson comorbidity index and Emergency visit) was considered to be the most preferrable metrics to predict the readmission in CHF, Diabetic patients as stated in (Wang et al., 2014; Mingle, 2017) because of its ease of use by health practitioners. But it was implemented using small dataset. So, there was a need for further research.

Recent study (Work, 2017) have used Diabetics Dataset from 130-US Hospital from 1999-2008 dataset which is 1,01,766 records with 50 attributes, which focuses on predicting Readmission of Diabetes patients and a model providing more information to doctors on the factors affecting the readmission. Random Forest provided an accuracy of 89% with promising feature weights that explained the model prediction. But, due to class Imbalance the model accuracy is lost to some extent. The same Random Forest model was tried by (Zhu et al., 2017) with small modification to Random Forest by adding randomness to each features and rank the features based on meanDecreaseAccuracy. The randomness gives high confidence by eliminating the bias in variables and selecting the high-risk factors influencing readmission. The researchers of (Goudjerkan and Jayabalan, 2019) also used Diabetes Dataset from 130-US Hospital from 1999-2008 dataset considered in-depth data preprocessing, performing, Missing value imputation using Hot-Deck imputation, Data reduction using clustering, handling data inconsistency, Outlier treatment, Class Imbalance using SMOTE along with Feature Engineering and achieved the model accuracy of 95% using Multilayer Perceptron to predict the 30-days Readmission for Diabetes patients. The study focuses more on accuracy and not interpretability of the model which is essential in health care as shown in (Yang et al., 2016). According to the study (Yang et al., 2016), High predictive power, interpretable results and prediction confidence constitutes a comprehensive framework to predict and understand hospital readmissions. LASSO and GBM were used in the study to predict and interpret the model to derive the risk factors. The model’s probabilities are used to construct 95% confidence interval and LASSO achieved a confidence interval of [0.77-0.95] with average readmission

probability of 0.875 and GBM with [0.696-0.845] confidence interval and average readmission probability of 0.781. This confidence interval adds Trust to the model prediction results.

CMS (Center of Medicare & Medicaid) started a program HRRP (Hospital Readmissions Reduction Program (HRRP) | CMS, 2021) to avoid excess readmission and improving the patient care with better discharge plans. According to section 3025 of Patients prediction and Affordable Act, Inpatient Perspective payment system hospitals will be reimbursed at lower rate for excessive readmissions. The article (Hospitals are avoiding admitting Medicare patients to dodge financial penalties, study suggests, 2021) stats that in FY-2019, 82% of the hospitals engaged in the program were penalized for readmission. (Maddipatla et al., 2015) aims to predict the 30-day readmission along the cost prediction model and the contributing risk factors associated with it. The study compared the performance of models Decision Tress, Gradient Boosting, Neural Network and Logistic Regression. Based on the AUC results, Decision Tree with 0.95 AUC was chosen to be the better model. And later, with the significant variables obtained, linear model was built to predict the revenue loss to the hospitals which shows financial impact on the hospitals.

(Bhuvan et al., 2016) conducted research on US 130 Hospital Dataset to predict 30-day readmission and identified the risk factors using black-box models, Naive Bayes, Bayesian Networks, Random Forest, Adaboost and MultiLayer Perceptron. MultiLayer Perceptron and Random Forest performs better when compared to other classifiers, with slight prevalence in Random Forest favor with AUC of 0.650. And feature analysis was performed to derive the risk factors influencing the readmission of diabetes patients and help medical practitioners gain more insights to understand why patients got readmitted within 30 days of the discharge. Ablation study of risk factors and Association Rule Mining was used for feature analysis and concluded with inpatient incidents, discharge disposition and admission type as most important features. While the initial study (Strack et al., 2014) on the same dataset tried to fit the relationship of HbA1c measurement on readmission of diabetes patients using multivariate logistic regression, controlling the covariates such as demographics, severity and type of the disease and type of admission. The statistical results suggested that the relationship between probability of readmission and HbA1c measurement depends on primary diagnosis.

Neural networks are receiving more demand compared to other Machine Learning classifiers in computer vision, image processing and medical classification. Well, machine learning

models are still of important because of its superior popularity of model interpretability and its ease of use and the main obstacle of neural networks in the black-box property that takes away the model interpretability as explained by the authors in (Dreiseitl and Ohno-Machado, 2002). Various research has been published (Hammoudeh et al., 2018; Goudjerkan and Jayabalan, 2019; Hu et al., 2019; Reddy et al., 2020) in predicting readmission of diabetes patients using neural networks with the main aim to improve the performance of classifier at the expense of interpretability of the model. (Hammoudeh et al., 2018) compared traditional artificial neural networks (ANN), convolution neural networks (CNN) and other machine learning models and resulted in CNN hitting ~92% accuracy, ~95% AUC. Neural networks performance depends on size of the data, so data engineering was handled appropriately, applying SMOTE to balance the data, feature creation by adding two new features from drug attribute, using one-hot encoding for feature transformation and removing the duplicate records. While other study (Sarthak et al., 2021) on the same dataset used deep neural network achieving an accuracy of 95.2%. Authors of (Reddy et al., 2020) compared the results of proposed algorithm DBFN (Deep Belief Network) with other machine learning models, Logistic regression, Ada-boost, Gradient Boosting, Random Forest, Decision Tree. DBFN resulted in 69% accuracy compared to other low accuracy machine learning models. Data Imbalance and reduction in data size might be the reason for low accuracy with models.

To add interpretability to neural networks, the book (Interpretable Machine Learning, 2021) shows the model interpretation methods and its importance. Various research has been done on the adapting of these methods. (Shrikumar et al., n.d.) used model agnostic method of interpretation, DeepLIFT (Deep Learning Important FeaTures) on MNIST dataset and simulated genomic data, LIME (Ribeiro et al., 2016) was introduced to explain the predictions using local models. Author (Lundberg et al., n.d.) compared LIME, SHAP, DeepLIFT to explain the model and presented SHAP to perform better in consideration with human explanations, while the recent study (Fidel et al., 2019) introduced a variation of SHAP, SHAP DeepExplainer to explain the DNN models.

# Research Questions

In this research project, there are two questions that are focused based on the literature survey conducted. Many researchers proved neural networks provide better performance with the exception of model interpretability.

1. Can the machine learning model, or a deep learning model be used by healthcare providers to predict the readmission of diabetes patient within 30-days of the discharge?
2. Can the model performance and interpretability (black box) be achieved using model agnostic interpretation methods (SHAP, LIME, DeepLIFT) to make the healthcare professionals trust the prediction of the model?

# Aim and Objectives

The aim of this study is to propose a model that predicts the readmission of Diabetes patients in less than 30-days of discharge and bring in interpretability of the models to detect the influencing factors which helps the health practitioner to better understand and trust the decision of the model and mitigate the risks with proper care and achieve transparency in treatment.

Following are the research objectives planned towards achieving the aim of the study:

* To visualize the features correlation.
* To suggest a suitable balancing technique to be applied on a class imbalance dataset.
* To compare between the decisions of predictive models to identify the influencing factors of readmission in diabetes patients for better care.
* To suggest a suitable interpretability method to explain the decision of the model.
* To evaluate the performance of predictive models based on the model interpretability.

# Significance of the Study

With the massive growth in digitization, large amount of clinical data is collected for the researchers to analyse and predict the readmission of hospitalized diabetes patients within a period of 30-days. Machine learning models and deep learning models are developed to make this prediction and achieve high accuracy models. But as we move towards high accuracy using the complex models, the model interpretation is lost. When there is a need to deploy the model, understanding why the model makes certain prediction becomes crucial in the field of medicine. The main purpose of this research is to develop a Neural network model with high accuracy and use model agnostic interpretable methods, SHAP and DeepLIFT to interpret the model. This study facilitates healthcare professionals to predict the early readmission of diabetes patients and improve the treatment of the patients by understanding the risk factors contributing to the predictions, establishing trust and confidence on the predictions of the model.

# Scope of the Study

The dataset used in this study is of US population, and the prediction of readmission of diabetes patients depends on certain demographic information of the patients, so the scope of this study is limited to diabetes patients within US. And, the dataset is highly class imbalanced, and not too large in size to reduce the size of the data with under-sampling, in order to leverage the information and retain data acceptable for deep learning models, SMOTE balancing technique is used.

With the recent growth in Artificial Intelligence, and understanding the body of literature, complex models are developed to gain high accuracy, compromising interpretability of the model. Along with it, research has grown towards developing new methods to interpret these complex models. DeepLIFT, SHAP has shown good explain-ability compared to LIME, Permutation feature importance and other interpretable methods. So, this study will perform analysis on how to leverage these methods with black-box models and make the models more acceptable and reliable.

# Research Methodology

Introduction**:**

This section will highlight the methods and algorithms used in this research to accomplish the objectives. Each of the following subsections explains Dataset used, data cleaning, data engineering techniques, SMOTE as a technique for balancing the data, feature selection using Random Forest, and the model classifier Random Forest, Multilayer Perceptron are trained and tested. Finally, the model interpretation methods SHAP, DeepLIFT are explained and used with models. The model performance is evaluated using accuracy, precision, recall, AUC and the decisions of the models are explained with the supporting features of prediction.

Dataset Description:

The Dataset used in this study is a de-identified abstract of the Health Facts database (Center Corporation, Kansas City, MO) (UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set, 2021). The dataset is collection of diabetes patients from 130 US Hospitals over a span of 10 years (1999-2008) which includes 1000000 records with 55 attributes. The attributes hold the following information,

1. It is a hospital admission.
2. It is diabetic encounter, that is one of 3 diagnosis is identified as diabetes.
3. The length of the stay between 1-14 days.
4. Medications offered at the time admission.
5. Laboratory tests conducted during the encounter.
6. Discharge deposition
7. Demographic information such as race, gender.
8. Target variable ‘readmitted’. (> 30 days, < 30 days, No)

Data Pre-processing:

The real world data is not gathered with respect to any specific purpose in mind. The data is noisy, inconsistent, heterogeneous, with large amount of missing values. Specially health care data includes lot of attributes, patient demographic information, medications, ways of discharge, laboratory tests, inpatient/outpatient details and so on. So, before starting to build the model, it is required to pre-process the data appropriately and have the relevant info to achieve better performance of the model (Khurana and Kumar, 2019). As seen in the literature survey (Mingle, 2017; Hammoudeh et al., 2018; Goudjerkan and Jayabalan, 2019; Hu et al., 2019) model performed better with data pre-processing applied, data transformation done, sampling techniques adapted, reducing dataset and optimizing it (Duggal et al., n.d.). In this study we will use the most relevant pre-processing identified by (Goudjerkan and Jayabalan, 2019). As such, Random Forest feature selection, ICD-9 code clustering, Hotdeck imputation are the key pre-processing tasks.

Missing values:

With this real world dataset of diabetes patients, there are a lot of missing values, ‘weight’ (97%), ‘payer code’ (40%) and ‘medical speciality’ (47%). As a general rule of thumb, a feature with more than 50% missing values should be removed. As such, ‘weight’ is dropped with highest percentage of missing value and does not influence this study. However, based on the previous research and understanding the background ‘payer code’ that shows the socio- economic condition of the patient is considered as important factor to predict readmission and ‘medical speciality’ is also highlighted to be the great influencer and its missing values are encoded as ‘missing’. In addition, few variables with low missing values and irrelevant to the study are removed from the dataset. Others with average missing value, imputation was performed in order to retain enough data to test and train the model. Imputation is a method to replace the missing value with the existing data in the dataset. There are multiple ways of

imputation, Mean Imputation, Logistic Regression Imputation, HotDeck Imputation. Imputation of each variable depends on its significance in the study, from the previous studies and recommendations, HotDeck imputation was considered to one of the reliable and commonly used imputation technique (Rithy, 2016). HotDeck imputation is a procedure of replacing missing value using values from one or more similar instances.

Data cleaning:

This second step of pre-processing refers to cleaning of duplicate or inconsistent data that alter the performance of the model. From the body of literature, this dataset has some unstable features that need to be taken care of. For instance, the dataset has multiple encounter of the same patient. So, only one record is retained while the others are deleted to avoid the bias in the model. From the literature study, the first encounter is selected (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021), while the last encounter would have lot of imbalanced data.

Discharge disposition refers to location or status, where the patient is moved after the admission. There are patients who died during their hospitalization and there is no probability of these patients to be readmitted, so exclude the records of the patients with ‘expired’ discharge disposition. Few others like ‘encounter\_id’, ‘patient\_nbr’ is removed. Features like ‘examide’ and ‘citogliption’ are listed as ‘No’ for all records, and will be deleted from further analysis.

Data reduction:

After the data is cleaned replacing the missing values and excluding the records leading to bias, improve the data by reducing the number of unique values in categorical variables. For instance, ICD-9 codes are used for ‘diag\_1’, ‘diag\_2’, ‘diag\_3’. The code ranges from 001-1000. These numerical values are difficult to interpret and make it unusable in the model. So, Clustering technique used in (Mingle, 2017) is applied to cluster the diagnosis code into nine categories, “Circulatory, Digestive, Diabetes, Injury, Genitourinary, Respiratory, Musculoskeletal, Neoplasms and others. Other features with high unique values, “medical\_speciality” , “admission\_type\_id”, “admission\_source\_id”, “discharge\_disposition\_id”, are converted to categorical values and will be clustered to enrich the dataset and interpret the model.

Class Imbalance:

The target variable considered for this study, “readmitted”, is found to be imbalanced with higher rate of “No” (~53%). In case of predictive modelling, Imbalanced classification results in poor performance, specially with respect to minority class. As the study focuses on predicting readmission within 30 days, which is a minority class, only holds ~11% of the data. This is a serious problem, because classification of minority class is more important than majority ones. Hence, treating class Imbalance is essential for the study. The oversampling method systematic minority over-sampling technique (SMOTE) is used to balance the data, generating new records of minority class, to balance the output classes. While to use under-sampling the data is not large enough to train and maintain the model performance.

Transformation:

Feature Engineering is the most important step for successful data-analysis. Specially with the medical data which contains lot of features with varied ranges. The study will perform feature extraction, feature encoding, feature selection to gain significant insight of features and Log transformation to standardise the data.

Feature Creation:

The services utilized by the patients over the last year are measured using, “number\_inpatient”, “number\_outpatient” and “number\_emergency”. Adding all these three variables, a new feature “service\_utilization” will be engineered.

Medication change was another feature of importance as predicted by (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021; Sarthak et al., 2021). There are 24 features of medication in the dataset, the feature indicates whether the drug was prescribed or was there a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no’ if the drug was not prescribed. A new feature, “medication\_change” is engineered by counting the change in medication during the patient encounter. And, to account for total number of medications used, “count\_of\_medication” will be added as new variable, which will be the sum of all medications used during the stay in hospital. Higher the number, higher will be the severity of the patient. Most of the researchers

tend to remove the original features and replacing it with new features, but this study considers all the features and statistically eliminate the features.

Feature Encoding:

The target variable “readmitted” is of three category, < 30-days, >30-days, “NO”. The study is focussed on readmission within 30-days, so above 30-days are considered as no readmission and it accounts for (~34%) and actual no-readmission is (~54%). Ideally the observations above 30-days is discarded, but it leads to reduction in the size of data and affects the model accuracy. Thus the output class is encoded as binary ‘1’ for a readmission and ‘0’ for “non-readmission”. A change in the medication represented by column, “change” will be encoded as ‘0’ for no change in medication and ‘1’ denoting a change in medication for values “No” (no change) and “Ch” (change) respectively.

A column describing if the diabetes medicine was prescribed or not, “diabetesMed” denoted with values ‘yes’ or ‘no’ are encoding to ‘1’ and ‘0’ respectively.

Feature selection: Feature selection is essential pre-requisite of data modelling (Ogedengbe and Egbunu, 2020) and has a great impact on model performance. This study will use Random Forest feature as the feature selection.

Log transformation:

Numerical variables such as “number\_inpatient”, “number\_outpatient”, “number\_emergency” are skewed and log transformed to ensure normal distribution (What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021).

After achieving the normal distribution of numerical variables from log transformation, standardise the data, using z-score scaling, where the difference in the mean of the variable and the variable itself is divided by standard deviation. This results in the dataset have equal mean(0) and standard deviation (1).

Outlier Removal:

Outliers are the observations or a set of observations which stands inconsistent from the rest of the observations of the same group. After the log transformation and standardisation, the outliers can be treated by deleting the data outside 3 standard deviation range (What are

Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information, 2021).

Modelling:

The main objective of this study is to get the model with high predictive performance along with explaining the features influencing the readmission. Based on the literature survey and the purpose of this research, the study will include two black box models, Random Forest with Gini and Multilayer perceptron (MLP). Then, use the model agnostic interpretation methods SHAP, DeepLIFT to derive the contribution of features.

A model is split into training and test data with 80:20.

Random Forest with Gini:

Random Forest classifier as the indicates is an ensemble of large number of individual trees. Each of these individual trees makes a prediction of the target class and the class with the maximum votes is elected for the final prediction of the model. And since this is a simple binary classification problem, random forest with Gini index is computed.

Gini = 1 - ∑𝑐 (pi)2

𝑖=1

Where, c is the number of class variables, pi is the probability of each class.

Multi-layer Perceptron:

Artificial Neural Networks (ANN) are one of the most powerful classifiers because of its ability to process large amount of training data and reduce the diagnosis time, resulting in high performance. The basic form of ANNs with one input layer and output layer is the single-layer perceptron (SLP), but it is proved that SLP cannot handle non-linearly separable parameters. To handle the non-linearity, hidden layers are added between input and output layer, with each layer comprising of one or more neurons and each layer fully connected to next layer. This multilayer network is the Multilayer perceptron.

MLP is a class of feedforward artificial neural networks, which accepts multidimensional input data, processes it through a weighted linear sum, followed by non-linear activation function and transforms the values to output layer to classify the output values. The performance of the model can be varied with number of layers added, adding more neurons and selecting the proper activation function.

The proposed model in this study will be MLP with one input layer, one hidden layer and one output layer. ReLU (Rectified Linear Unit) or PReLU (Parametric ReLU) activation function is used at the input layer, based on the performance. At the output layer, Sigmoid activation is used because of the binary classification problem that will be addressed in this study. The model is trained to reduce the loss and regularized by adding a Dropout after each of the input layers, avoiding overfitting of the models.

Model Agnostic interpretability Methods proposed for this study:

* 1. Shapley values :

Shapley values -- a method from coalitional game theory -- tells us how to fairly distribute the prediction among the features. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. As the number of features increases, number of possible coalitions increases leading to expensive computation. (Interpretable Machine Learning, 2021)

* 1. SHAP (Shapley Additive explanations):

SHAP is a solid theoretical foundation in game theory. The prediction is fairly distributed among the feature values. We get contrastive explanations that compare the prediction with the average prediction. SHAP is good for tree based models unlike Shapley values with high computational cost. With the fast computation of Shapley values, Global interpretation can be easily achieved, holding the three most important properties of Additive feature Attribution 1. Local accuracy 2. Missingness

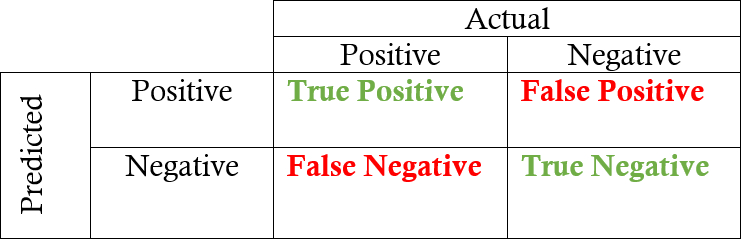
3. Consistency (Interpretable Machine Learning, 2021)

* 1. DeepLIFT (Deep Learning Important FeaTures):

DeepLIFT is a method for explaining the output prediction of neural networks, assigning the scores to the input variables. DeepLIFT compares activation of each neuron to the “reference activation” to assign the scores. The feature of “reference” outperforms this model from other gradient-based methods (Shrikumar et al., n.d.).

Evaluation Metrics:

The prediction of the model is evaluated based on Model accuracy, precision, recall, AUC- ROC. In this study of medical science, recall is important because hospitals are penalized for the patients readmission which incurs high cost to hospitals.



True Positive (TP) True Negative (TN) False Positive (FP) False Negative (FN)

Accuracy: Accuracy is the basic and most commonly used evaluation metrics to measure performance of model. It is the ratio of number of correct predictions to the total number of input samples.

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

Recall: Recall is the ratio of true positives (items that are correctly labelled to positive class) to the total number of elements that actually belong to positive class.

Recall = (TP)/(TP+FN)

Precision: Precision is the ratio of true positives (items that are correctly labelled to positive class) to the total number of elements labelled as belonging to positive class.

Precision = (TP)/(TP+FP)

The Area Under the Curve (AUC) specifies the ability of the model to distinguish between the classes (positive and negative). Higher the AUC score , better is the models prediction.

Basic flow chart of the Methodology:



start

Load the data from the csv file

Training data (80)

Test data (20)

Model Agnostic Interpretable methods

Split the dataset 80:20

Balance data

Data reduction

Modelling Random Forest Multilayer perceptron

Feature Engineering

Pre-process data

Missing value imputation

Data Cleaning

SHAP

DeepLIFT

Important features

Important features

END

Compare the features and conclude the important features of high performing model

# Expected Outcomes

Model Evaluation

Precision-recall

Accuracy

AUROC

A predictive classification model with high performance to predict the early readmission of hospitalised diabetes patients and explain the model’s decision, deriving the important features contributing to early readmission.

# Requirements / resources

Hardware Requirements:

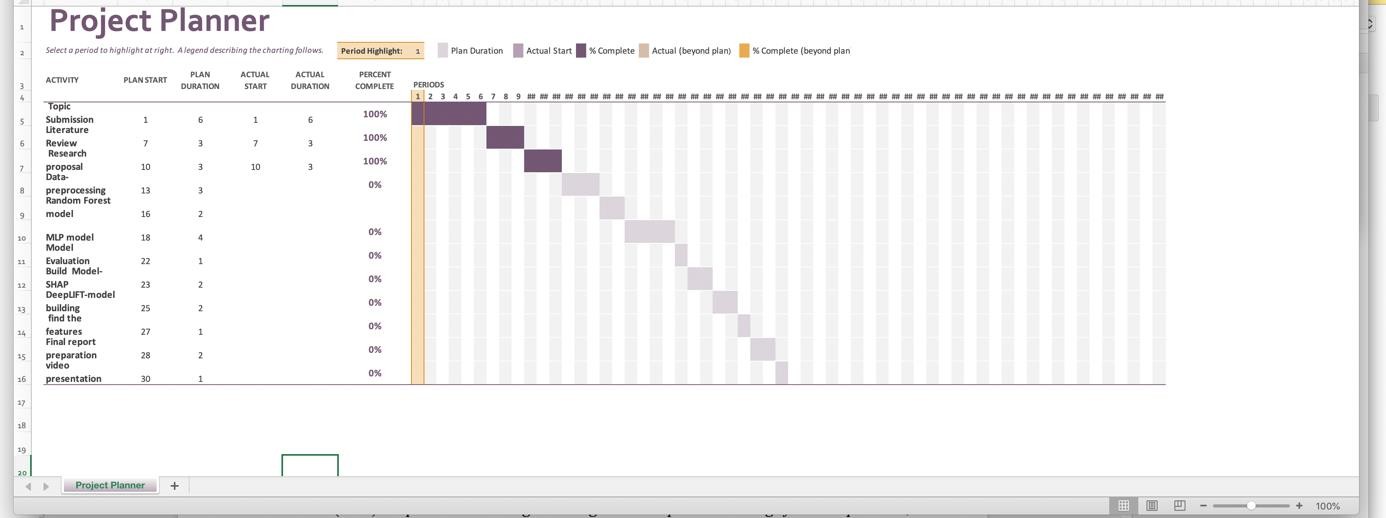
* 1. GPU with 16GB RAM

Software Requirements:

The predictive model will be built in Python

* 1. Anaconda Navigator with python 3
  2. Python libraries such as Pandas and NumPy for data pre-processing, Matplotlib, seaborn for data visualization, scikit-learn for model building and model evaluation, Shap python package for model interpretability.

# Research Plan



**References**

Anon (2021) *Hospital Readmissions Reduction Program (HRRP) | CMS*. [online] Available at: https:/[/www.cms.gov/Medica](http://www.cms.gov/Medicare/Medicare-Fee-for-Service-)r[e/Medicare-Fee-for-Service-](http://www.cms.gov/Medicare/Medicare-Fee-for-Service-) Payment/AcuteInpatientPPS/Readmissions-Reduction-Program [Accessed 9 Jan. 2021].

Anon (2021) *Hospitals are avoiding admitting Medicare patients to dodge financial penalties, study suggests*. [online] Available at: htt[ps://www.advisory.com/dail](http://www.advisory.com/daily-)y[-](http://www.advisory.com/daily-) briefing/2019/09/05/readmissions [Accessed 9 Jan. 2021].

Anon (2021) *Interpretable Machine Learning*. [online] Available at: https://christophm.github.io/interpretable-ml-book/ [Accessed 17 Jan. 2021].

Anon (2021) *UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999- 2008 Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Diabetes+130- US+hospitals+for+years+1999-2008 [Accessed 11 Jan. 2021].

Anon (2021) *What are Predictors of Medication Change and Hospital Readmission in Diabetic Patients? | UC Berkeley School of Information*. [online] Available at: htt[ps://www.ischool.berkeley.edu/proje](http://www.ischool.berkeley.edu/projects/2017/what-are-predictors-medication-change-and-)c[ts/2017/what-are-predictors-medication-change-and-](http://www.ischool.berkeley.edu/projects/2017/what-are-predictors-medication-change-and-) hospital-readmission-diabetic-patients [Accessed 13 Jan. 2021].

Bhuvan, M.S., Kumar, A., Zafar, A. and Kishore, V., (2016) Identifying Diabetic Patients with High Risk of Readmission. [online] Available a[t: http://arxiv.org/abs/1602.04257.](http://arxiv.org/abs/1602.04257)

Dreiseitl, S. and Ohno-Machado, L., (2002) Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 355–6, pp.352–359.

Duggal, R., Shukla, S., Chandra, S., Shukla, B., Sunil, & and Khatri, K., (n.d.) Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India. *International Journal of Diabetes in Developing Countries*.

Fidel, G., Bitton, R. and Shabtai, A., (2019) When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. *arXiv*, 1.

Goudjerkan, T. and Jayabalan, M., (2019) Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron. *International Journal of Advanced Computer Science and Applications*, 102, pp.268–275.

Hammoudeh, A., Al-Naymat, G., Ghannam, I. and Obied, N., (2018) Predicting hospital readmission among diabetics using deep learning. *Procedia Computer Science*, [online] 141, pp.484–489. Available at: https://doi.org/10.1016/j.procs.2018.10.138.

Hu, P., Li, S., Huang, Y.A. and Hu, L., (2019) Predicting hospital readmission of diabetics using deep forest. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, pp.2019–2020.

Khurana, G. and Kumar, P.A., (2019) Improving Accuracy for Diabetes Mellitus Prediction Using Data Pre-Processing and Various New Learning Models. *International Journal of Scientific Research in Science and Technology*, April, pp.502–515.

Lundberg, S.M., Allen, P.G. and Lee, S.-I., (n.d.) *A Unified Approach to Interpreting Model Predictions*. [online] Available at: https://github.com/slundberg/shap.

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

Mingle, D., (2017) Predicting Diabetic Readmission Rates: Moving Beyond Hba1c. *Current Trends in Biomedical Engineering & Biosciences*, 73.

Ogedengbe, M.T. and Egbunu, C.O., (2020) CSE-DT Features selection technique for

Diabetes classification. *Applications of Modelling and Simulation*, 4March, pp.101–109. Reddy, S.S., Sethi, N. and Rajender, R., (2020) Evaluation of Deep Belief Network to Predict Hospital Readmission of Diabetic Patients. *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020*, pp.5–9.

Ribeiro, M.T., Singh, S. and Guestrin, C., (2016) ‘Why should i trust you?’ Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, pp.1135–1144.

Rithy, D., (2016) *Simulation of Imputation Effects Under Different Assumptions A Senior Project presented to*.

Sarthak, Shukla, S. and Prakash Tripathi, S., (2021) Embpred30: Assessing 30-days readmission for diabetic patients using categorical embeddings. *Advances in Intelligent Systems and Computing*, 1168February, pp.81–90.

Shrikumar, A., Greenside, P. and Kundaje, A., (n.d.) *Learning Important Features Through Propagating Activation Differences*. [online] Available a[t: http://goo.gl/qKb7pL,.](http://goo.gl/qKb7pL)

Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J. and Clore, J.N., (2014) Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, [online] 2014, p.781670. Available at: https://doi.org/10.1155/2014/781670.

Wang, H., Robinson, R.D., Johnson, C., Zenarosa, N.R., Jayswal, R.D., Keithley, J. and Delaney, K.A., (2014) Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders*, 141, pp.1–8.

Work, R., (2017) Beating Diabetes : Predicting Early Diabetes Patient Hospital Readmittance to Help Optimize Patient. 2014May 2014, pp.1–7.

Yang, C., Delcher, C., Shenkman, E. and Ranka, S., (2016) Predicting 30-day all-cause readmissions from hospital inpatient discharge data. *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016*, pp.2–7. Zhu, Q., Akkati, A. and Hongwattanakul, P., (2017) Risk feature assessment of readmission for diabetes. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, pp.538–543.