National College of Ireland

Diabetes Diagnosis and Readmission Risks Predictive Modelling: USA

MSc Research Project
MSc in Data Analytics

Clodagh Reid
Student ID: X17161207

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Clodagh Reid
Student ID:	X17161207
Programme:	MSc in Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr Catherine Mulwa
Submission Due Date:	12/08/2019
Project Title:	Diabetes Diagnosis and Readmission Risks Predictive Modelling: USA
Word Count:	10,641
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 9/8/2019.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

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

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

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

Diabetes Diagnosis and Readmission Risks Predictive Modelling: USA

Clodagh Reid
X17161207

Abstract

One of the most critical healthcare problems today is diabetes. In the US, 30 million people are affected by diabetes with healthcare related costs at a sobering $327 billion annually. Long term effects if untreated can result in damage to the heart, blood vessels, eyes, kidneys, feet, nerves and even mortality from heart attack or stroke. As diabetic patients have increased, consequently the number of diabetic hospital readmissions have become greater. Early readmission can impact patient health, operational efficiency and cost burden. The aim of this research is to examine diabetes diagnosis and early readmission with the same dataset. A comprehensive methodology with pre-processing and transformation which comprised of permutation feature importance, feature engineering and SMOTE are some of the methods used to deal with noisy, inconsistent, imbalanced data. Predictive models include LR, BDT, SVM, NN, DF. The best performing model is selected to create a web service where users can input data and receive scored results connecting the user to the data. Metrics include accuracy, recall and AUC to measure the performance of the models where a Boosted Decision Tree achieved the highest results. Hospital readmission accuracy at 86% and diabetes diagnosis accuracy at 67%.

Keywords – Diabetes, Readmission, Predictive Modelling, Classification.

1 Introduction

The amount of people that have been diagnosed with diabetes has increased globally (WHO, 2017). Diabetes affects 30 million americans with related costs increasing from $245 billion in 2012 to a sobering $327 billion in 2017[1]. $98 billion of these costs are related to hospital inpatient care. We should be alarmed at the increasing and aggressive growth rate of diabetes related cases and the staggering diabetes related costs.

As a possible direct consequence of diabetes increasing, the number of hospital inpatients readmissions continues to rise. A hospital readmission is when a patient who has been discharged from hospital is readmitted again within a certain time period. Hospital readmissions are now a metric for hospital quality (CMS, 2019). Centers for Medicare & Medicaid Services created the Hospital Readmissions Reduction Program with an aim to improve quality of care and reduce healthcare spending. As hospital readmissions have increased inline with the prevalence of diabetes, it is likely that it may continue to do so compounding the problem (Rubin et al., 2014).

There are predominantly 3 types of diabetes. Type 1 is when there is a lack of insulin production and daily administration of insulin is required. Type 2 occurs when the body

[1] http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

ineffectively uses insulin. Type 3 Gestational diabetes occurs in pregnancy when women have raised blood sugar levels as the body isn't able to use the sugar in the blood as well as it should. Mother and baby have an increased risk of Type 2 diabetes in later life. Out of all 3 categories, Type 2 diabetes is the most common type of diabetes.

1.1 Motivation and Background

Early detection of Type 2 diabetes can reduce and delay diabetes. This can be achieved with exercising, healthy eating, not smoking and by maintaining a healthy body weight. The later the detection of the disease, the worse the diagnosis outcome. The contributing factors such as inactivity and obesity are non-genetic contributing factors. Diabetes can be a trade-off between healthy living comprising of healthy eating, exercising versus convenient, demanding and hectic lifestyles. Undiagnosed diabetes can overtime damage the heart, blood vessels, eyes, kidneys, feet and nerves and increase the risk of heart disease and stroke. Uncontrolled diabetes in pregnancy can have a detrimental effect on mother and baby, with increased chance of fatal loss, malformations, still birth, perinatal death and complications. Gestational diabetes increases the risks of complications before, during and after delivery.

The beneficiaries of this project are twofold, the patient themselves who will benefit in terms of disease management, overall health and early detection. The health service providers will gain, they will have a better understanding of the data where action can be taken to reduce early readmissions associated with the patient diagnosis. Early detection and treatment are essential in order to provide bettter treatment to patients and potentially saving lives and reducing readmitted patients treatment healthcare costs.

The Diabetes 130-US hospitals for years 1999-2008 datasest (Clore et al., 2014) (Strack et al., 2014) which has been selected is more contemporary with several cultures and age profiles ranging from 0-100. This research focuses on datamining techniques to develop predictive models for classifying diabetes patients by predicting diabetes readmission, short term (within 30 days) or long term (after 30 days) and predicting diabetes diagnosis. This project seeks to incorporate a higher recall (sensitivity) as this is suited to the healthcare industry. In healthcare it is important to predict a result but more so to have the correct patient result when a patient is suffering from diabetes (true positives). This will ensure no patient is left untreated. Thus, the research metrics include recall and accuracy.

1.2 Research Question

In this research project, there are two different research questions which are applied to one dataset. The same techniques, methods and models are carried out for each research question. Two different research questions are used to gain valuable information and insights for the health service provider. A new feature 'Diabetes' is created for the Sub RQ.

RQ: "Can hospital diabetic patient readmissions (i.e. within 30 days) be predicted using predictive modelling techniques (Logistic Regression, Boosted Decision Tree, Decision Forest, Neural Network, Support Vector Machine) to allow health service providers to better address unplanned readmissions while improving operational efficiency and cost efficiency."

Sub RQ: "Can patients with diabetes be predicted using predictive modelling techniques (Logistic Regression, Boosted Decision Tree, Decision Forest, Neural Network, Support Vector Machine) to allow health service providers to improve patient diagnoses and quality of care."

To solve the research questions the objectives in Table 1 are specified and implemented.

1.3 Research Objectives

TABLE 1. Research Objectives

Objectives	Description	Evaluation Metrics
Objective 1	A critical review of literature on diabetes diagnosis and hospital readmission risk classification and regression predictive models	
Objective 2	Incorporate improved feature selection to learning classifiers and investigate the contribution factors	
Objective 3(a):	Implementation, evaluate and results	Logistic Regression
Objective 3(b):	Implementation, evaluate and results	Boosted Decision Tree
Objective 3(c):	Implementation, evaluate and results	Decision Tree Forest
Objective 3(d):	Implementation, evaluate and results	Neural Network
Objective 3(e):	Implementation, evaluate and results	Support Vector Machine
Objective 4:	Comparison of developed models	
Objective 5:	Comparison of developed models with existing models	
Objective 6:	Design, implementation and evaluation of web services for visualization of results	

The rest of the technical report is structured as follows. Chapter 2 presents an investigation into diabetes diagnosis and hospital readmission existing literature. Chapter 3 presents the research methodology. Chapter 4 presents the design specification. Chapter 5 presents the implementation, evaluation and results of predictive models to predict diabetes diagnosis and hospital readmission. Chapter 6 presents the discussion. Chapter 7 presents the conclusion based on the results and recommendation of future work.

2 Related Work

2.1 Introduction

Medical databases present exciting opportunities learning from patient data to make predictions. There is a variety of hospital readmission and diabetes diagnosis research literature available which lead to diverse insights and approaches. Section 2.2 presents a critical review of diabetes diagnosis methods, techniques, algorithms and identified gaps. Section 2.3 presents a critical review of hospital readmission methods, techniques, algorithms and identified gaps.

2.2 A Critical Review of Diabetes Diagnosis Methods, Techniques, Algorithms and Identified Gaps

A considerable amount of research has been carried out in the diabetes prediction with the implementation of different machine learning algorithms. The Pima Indian Diabetes dataset (Kaggle, 2016) has been used extensively during diabetes diagnosis research. There are limitations to the data set as it contains 768 females, no males and all patients are over 21 years

old and from the same ethnic background. The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and is useful for comparitive purposes.

In a survey carried out by (Vijayan and Anjali, 2015) 200 samples of data were collected for a diabetes local data set. The Pima Indian diabetes dataset is used to train data while testing was carried out on a local dataset from Kerala. There is a fundamental flaw with this concept as the data is trained and tested on different datasets. The data is from the same demographic and socioeconomic background, but a better solution would be to combine the two datasets, shuffle the data and then split for training and testing.

This paper (Wu *et al.*, 2018), seeks to improve accuracy by utilizing K-means to cluster the data. The incorrectly classified data are removed followed by Logistic Regression to classify the data. The data had to be 75% of the total amount to move onto the Logistic Regression stage. 10-fold cross validation was carried out with an aim to reduce bias. The model was trained and tested 10 times. The kappa statistic was high which confirms the model pertains great consistency. In comparison (Patil, Joshi and Toshniwal, 2010) who eliminated missing data and removed incorrectly classified data at the K-means clustering stage had 433 instances after K-means and C4.5 Decision Trees achieving accuracy of 92.38%. (Wu et al., 2018) with 589 instances correctly classified with K-means and Logistic Regression achieved high accuracy results than (Patil, Joshi and Toshniwal, 2010) at 95.42%.

(Kandhasamy and Balamurali, 2015) proposes to use classifiers to predict diabetes with noisy and non-noisy data to compare accuracy results. J48 achieves the highest performing results for accuracy, sensitivity results and specificity with noisy data. Random Forest achieved the lowest accuracy result due to its inability to deal with missing data. After classifying without the noisy data Random Forest and KNN achieve the highest results at 100% accuracy, sensitivity and specificity 100%. The pre-processing techniques to remove the noisy data are crucial to machine learning.

(Chen et al., 2017) research paper highlights data mining techniques which are vital in medical diagnosis. The approach includes data reduction with K-means followed by J48 Decision Tree with 10-fold validation method. K-means has proved to be popular as it is easy to use and understand (Jain, 2010). 236 instances were incorrectly classified and removed resulting in a smaller sample size at 532. The results for predicting Type 2 diabetes are high with accuracy at 90.04%, sensitivity at 87.27% and specificity at 91.28% in comparison to other works. In the paper, (Al Jarullah, 2011) researchers implement pre-processing with feature selection, handling missing values and numerical discretization, followed by 10-fold cross validation to reduce bias. Random sampling was implemented to improve data quality and increase accuracy. The J48 Decision Tree algorithm model accuracy result at 78.2% is a lower result than (Chen et al., 2017).

(Guo, Yang; Bai, Guohua; Hu, 2012) utilizes the Pima Indians Diabetes dataset (Kaggle, 2016) and performs pre-processing techniques such as data cleaning, discretization and data normalization. The algorithms carried out comprise of Bayes Network to predict Type 2 diabetes. It is over complicated for the accuracy results achieved with accuracy at 72.3%. (Komi et al., 2017) researchers use five predictive models to predict diabetes with Gaussian mixed model, Extreme Learning Machine, Support Vector Machine, Logistic Regression and Artificial Neural Network. The Artificial Neural Network produces the highest result at 89%

with 2 hidden layers and 5 hidden neurons. Artificial Neural Networks outperforms the other algorithms but interpretability of the relationships between the variable is harder to understand.

(Meng et al., 2013) presented a research paper comparing three predictive models to predict diabetes using common risk factors with patients from Guangzhou China. The members consisted of two groups pertaining 735 patients with diabetes and 753 volunteers without diabetes. Models were created utilizing 12 input variables and ranked by importance. Age is ranked number 1 for all 3 predictive models. Gender, age, marital status, educational level, history of diabetes, BMI, physical activity, duration of sleep, work stress all showed a statistically significant at .001 when predicting diabetes with the Pearson Chi-square Test. Results are evaluated for accuracy, sensitivity and specificity with a dataset. The C5.0 Decision Tree achieved the highest accuracy at 77.87% while the Artificial Neural Network achieved the best sensitivity at 82.18%. This dataset is balanced due to the number of members in each class and the sample size is larger than the Pima Indian dataset at 1,488 members.

When predicting diabetes, the most common models used are Logistic Regression, Neural Network, Support Vector Machine, Decision Tree and Naïve Bayes with the most common metrics accuracy, sensitivity and specificity.

2.3 A Critical Review of Hospital Readmission Methods, Techniques, Algorithms and Identified Gaps

Much of the existing work regarding hospital readmission has been exploratory and statistical. There have been varied uses of hospital data that will be covered in this section. (Rubin, 2018) research paper examines how diabetes patients are at higher risk from hospitalization than those patients without diabetes. Potential ways to reduce that risk is through education, specialty care, improved discharge instructions, coordination of care and discharge support while focusing on the patients which are at the highest risk. 20% of all hospitalizations are diabetes related (Jencks, Williams and Coleman, 2009). Thus, if the number of readmissions reduced, the total cost burden would greatly reduce improving medical care.

Recent work[2] to predict hospital readmission with Logistic Regression, Decision Tree and Random Forest models used the 130 US hospital dataset. Comprehensive pre-processing techniques are carried out achieving high accuracy results at 94%. However, while this paper although achieved high results, it hasn't been published or peer reviewed.

Datamining is crucial in machine learning as data maybe noisy, inconsistent and contain missing values. (Goudjerkan and Jayabalan, 2019) researchers used the 130-US hospitals for 1999-2008 dataset and performed in-depth pre-processing incorporating approximate Bayesian Bootstrap, clustering and Random Forest feature selection, missing values, inconsistencies, data reduction and feature engineering steps to improve dataset imbalance and inconsistent data. The Neural Network and Multilayer Perceptron model accuracy at 95%, precision, recall and AUC results are successful in predicting 30-day readmission for patients with diabetes. While (Negi and Jaiswal, 2016) research use the same Diabetes 130-US hospitals for years 1999-2008 dataset, they combine the dataset with the Pima Indians diabetes dataset increasing its size. Pre-processing, data fusion and normalization of data is carried out and a Support Vector Machine model is implemented. The combined datasets contain 102,538 instances.

[2]https://www.ischool.berkeley.edu/projects/2017/what-are-predictors-medication-change-and-hospital-readmission-diabetic-patients

There is a fundamental issue with this concept, the patients are from different demographic and socioeconomic backgrounds. The class imbalance hasn't been addressed as 102,528 instances equates to the datasets combined total instances while if SMOTE or other techniques were performed to transform the data, the number of instances would move.

Other researchers have implemented popular supervised models such as Decision Trees, Neural Networks, Support Vector Machine. (Turgeman and May, 2016) researchers built a boosted C5.0 tree as the base classifier and a Support Vector Machine as a secondary classifier in an ensemble with a dataset encompassing 4,840 patients and 20,321 inpatient admissions. High results are achieved with total accuracy ranging from 81%-85%. Logistic Regression models are used to compare results. Logistic Regression model assumption is that observations should be statistically independent. As patients that were readmitted on the same day were merged into one record this assumption has been violated. The researchers confirm the data is highly imbalanced but the data pre-processing techniques don't mention how to deal with this issue. (Baskaran *et al.*, 2011) researchers predict breast cancer by creating a new algorithm that includes back propagation and radial basis function Neural Networks for prediction. Models achieved high results with accuracy at 80% and positive predictive value at 88%, the results for negative predictive value were less successful and deemed further work.

(Duggal *et al.*, 2016a) research proposed to pre-process the data effectively and predict 30-day readmission risk with diabetic patients in India. Challenges include understanding and identifying the relevant attributes as real world data is noisy and inconsistent. Class labels are encoded with two values readmitted <30 days and other (no readmission and readmission >30days). Pre-processing include feature selection, missing value imputation and class imbalance resolution. The classifiers which include pre-processing outperform the baseline methods which do not include pre-processing. Pre-processing is an integral part of machine learning.

In addition to supervised classification modelling, regression analysis is another technique that has been used to gain insights. (Goodney *et al.*, 2003) researchers use regression techniques to examine the relationship between length of stay, 30-day readmission and hospital volume. The length of stay ranged from 3.4 days to 19.6 days with no consistent relationship between volume and mean length of stay. Linear regression is used to examine the relationship between hospital volume and length of stay and Logistic Regression for the hospital volume and readmission rate analysis. Patients with a long length of stay were outliers and skewed the distribution which violated the normal distribution assumption of regression. Thus, Logarithmic Transformation is carried out. The length of stay varied but patients with cancer in general had the longest stay and the highest 30-day readmission included patients undergoing mitral valve replacement. (Strack *et al.*, 2014) uses multivariable Logistic Regression to fit the relationship between HbA1c and early readmission controlling for covariates such as demographics, severity and type of disease. Evidence has shown that the relationship between HbA1c and early readmission depends on the primary diagnosis.

The Affordable Care Act of 2010 section 3025 states hospitals maybe reimbursed at a lower rate for patients readmitted within 30 days[3]. (Maddipatla *et al.*, 2015) aims to predict 30-day readmission, the cost implication associated with the readmissions and the contributing factors.

[3] https://www.dpc.senate.gov/healthreformbill/healthbill05.pdf

The models built to compare results include Decision Trees, Gradient Boosting, Logistic Regression and Neural Networks with Decision Trees achieving the best AUC results. Linear Regression is utilized to build a cost prediction model successfully illustrating an estimate of the associated financial impact. (Duggal *et al.*, 2016b) cost analysis highlights the actual cost of readmissions. The research includes 5 popular datamining classifiers suitable for binary classification with the best results achieved by a Random Forest model with accuracy at 87.61%. The AUC is low for all classifiers due to the dataset class imbalance. The classifier is biased towards the majority class resulting in a high accuracy.

(Zheng *et al.*, 2015) research paper includes a map of papers in Table 1 detailing sample size, attributes, methodology and readmission length. 30 days readmission length accounts for over half of the papers, also known as early readmission. This supports the continued focus on cost savings during this time. Support Vector Machine modelling with radial basis function achieved the best results at 78.4% for accuracy and 97.3% for sensitivity in comparison to other Support Vector Machine, Logistic Regression and Random Forest classifiers. Random over-sampling is used during pre-processing to adjust the classes resulting in a balanced dataset.

Other researchers that demonstrate varied analysis (Zhao and Yoo, 2017) proposed readmission prediction modelling with the best results achieved by Naïve Bayes at an average AUC = 0.655 ± 0.078. The results are greatly impacted by the imbalanced data streams.

The research referred to in this section demonstrates exploratory and statistical analysis to understand and identify the relevant attributes as real world data is noisy and inconsistent. The pre-processing, transformation and feature selection steps are crucial to the predictive modelling results. Some researchers (Zhao and Yoo, 2017)(Maddipatla *et al.*, 2015) use one metric AUC, while a variety of metrics can give additional insights.

2.4 Conclusion

Research has focused on exploratory analysis and statistical analysis when predicting diabetes diagnosis and hospital readmission, so health service providers can better understand and address patient diagnosis and unplanned readmission to improve quality of care and reduce costs. The quality of the data to a large extent affects the prediction of the model, thus pre-processing and transformation are vital in medical diagnosis.

Based on the literature review, gaps were clearly identified such as class imbalance and the delivery of predictive models to the health service providers. In order to address these gaps, there is a clear need for detailed pre-processing and classification prediction models to answer research question and research objectives in Chapter 1. Furthermore, the delivery of these models to the health service providers will be addressed in this research by implementing Azure Web Services. Many authors have done research in the diabetes and hospital readmission field but the link to the health service provider is missing. To the best of my knowledge there is no research which uses the Diabetes 130 hospitals for years 1999-2008 dataset for predicting hospital readmission and diabetes diagnosis and linking this information back to the health service provider using a web service.

In the papers I have investigated for this project, researchers did not use Azure Machine Learning Studio. A unique approach to this project will be to create and deploy predictive models as web services in Azure Machine Learning. The next chapter presents the research

methodology approach used to develop the prediction models to support the key stakeholders in diagnosing diabetes and hospital readmission risk.

3 Research Methodology Approach Used, Design and Data Pre-Processing

3.1 Introduction

The purpose of this research is to develop a method of predicting diabetes diagnosis and diabetes hospital readmission risk. On review of the CRISP DM, SEMMA and Knowledge Discovery in Databases (KDD) methodologies, KDD provides feature selection which is significant to this research (Azevedo and Santos, 2008). The KDD modified approach was selected to guide this research through the stages comprising of selection, pre-processing, transformation, data mining and interpretation/evaluation (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Each stage is described in more detail in Section 3.2. The dataset represents 10 years (1999-2008) at 130 US hospitals (Clore et al., 2014)(Strack et al., 2014) with 50 attributes and 101,766 instances.

3.2 Diabetes Diagnosis and Readmission Risk Methodology Approach

The KDD methodology (Feyyad, 1996) has been modified as per Fig. 1 for diabetes diagnosis and readmission risk with the following stages (i) data selection with the data from 130 US hospitals is available in a zipped format (ii) All the data in .CSV format are extracted and loaded into Azure Machine Learning. (iii) Data understanding, data cleaning and pre-processing are performed (iv) The features are selected, feature engineered, outlier removal, SMOTE and normalization transform the data (v) Logistic Regression, Boosted Decision Tree, Decision Forest, Support Vector Machine and Neural Network models are trained and tested. (vi) Models are evaluated and interpreted with accuracy, recall, precision, F1 and AUC.

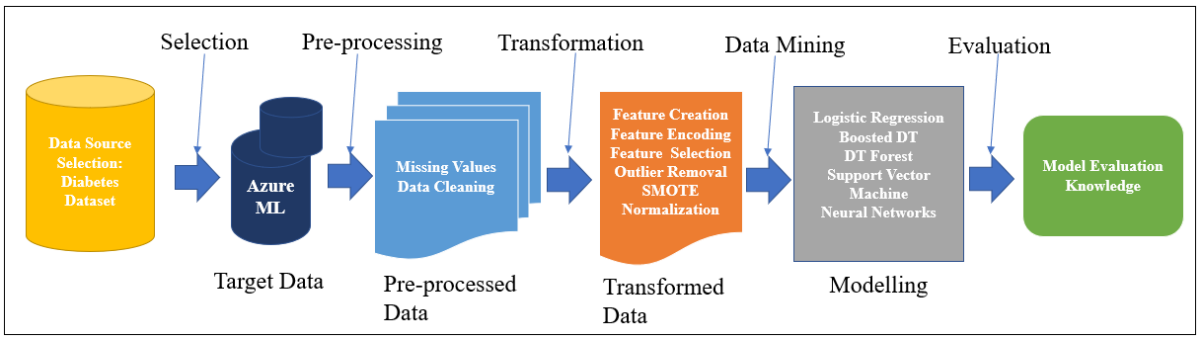

Fig. 1. Diabetes Diagnosis and Readmission Risk Methodology

Selection of Target Data: The Diabetes 130-US hospitals for year 1999-2008 dataset (Clore et al., 2014)(Strack et al., 2014) consists of a diabetic dataset and a supporting IDs mapping.csv file. The dataset has been selected primarily due to its diversity of more complex data with detailed inpatient diabetic encounter records. The IDs mapping file contains a mapping of unique identifiers for ‘admission type’ and ‘discharge disposition’ within the diabetes 130 US hospitals for year 1999-2008 dataset. This data is challenging with incomplete and noisy data.

Pre-processing: Datasets used in data mining are not necessarily gathered with a specific purpose in mind. Data maybe lacking in quality, contain errors, missing data and outliers. In order to use those datasets in the data mining process the data must undergo pre-processing which can account for 60% of the time and effort spent on the data. The purpose of pre-processing is to gain consistent data which will benefit the data mining stage and yield higher results. The pre-processing steps in this section include the most appropriate steps relevant to this project (Goudjerkan and Jayabalan, 2019)(Strack *et al.*, 2014)(Duggal *et al.*, 2016a).

Missing Values: There are several features with a high percentage of missing values. These features were 'weight' (97% missing), 'payer code' (40% missing) and 'medical specialty' (47%). The 'weight' feature is removed. The 'payer code' is removed due to the lack of data and it is not applicable for this research. The rule of thumb that data should be removed if less than 50% was performed. While this is normally the case, researchers have highlighted the importance of 'medical specialty' and the missing values have been encoded 'missing'. 'Race' (6% missing) are replaced with 'Unknown' by using the 'Clean Missing Data' pill.

Data Cleaning: The preliminary dataset included unique patients with one or more encounters. A Logistic Regresion model assumption is that the observations should be statistically independent. For this reason, the use of each unique patient with the first encounter is selected. Additionally, discharge disposition contains the patients status or location after admission. As per other researchers patients who have died or who are in a hospice will not to be readmitted. Thus, patients with 'discharge disposition' codes 11, 13, 14, 19, 20, 21 which are related to a hospice or expired have been removed during this research.

Patient number, discharge disposition id and glucose serum test results are updated to more manageable and consistant feature names such as 'patient_id', 'discharge_id' and 'glucose' with the 'Edit Metadata' pill. The change of name will make it easier during coding.

The 'examide' and 'citoglipton' predictor variables are removed as these variables pertained 'No' for all instances. Missing values are analysed and replaced with the value '0' with the 'Convert to Datset' pill. Data types are explored to confirm each data type matches the expected one. The 'Edit Metadata' pill is used to change the columns 'Admission Type', 'Discharge id' and 'Admission source id' from numeric to categorical data types.

Class Imbalance: The quality of the predictive model results are challenged with class imbalance. Class imbalance is an uneven distribution between the majority and minority classes which leads to bias in the majority class. SMOTE is a technique which takes the entire dataset as an input and creates new instances from the minority class, increasing the minority class portion in relation to the total classes (Sáez *et al.*, 2015).

Transformation: The purpose of transformation is to convert the data with methods such as feature selection and engineering, SMOTE and normalization into a more consistent dataset. Azure Machine Learning and SQL transformation code are implemented[4]. The split at 70% and 30% is used to train and evaluate models. Cross validation accuracy and recall performance results achieved lower results, hence the split and train method are selected.

Feature Creation: In Table 2 three new features are created, 'Diabetes', 'Service Utilization' and count of 'Medication Change' with the 'SQL Transformation' code in Azure

[4] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/apply-sql-transformation

Machine Learning. New features are created to enrich the original dataset. Some researchers would drop the original columns and maintain the new feature. In this research, all features have been kept and will be used in feature selection process. The dataset contained 23 medication features with categorical data types, 'no', 'steady', 'up' or 'down'. A count of medication change has been engineered to count all the changes.

TABLE 2. Newly derived and calculated features

Field Name	Data Type	Notes
Diabetes	Binary	All Diabetes patients are classed. A binary '1' for diabetes patients diagnosed with diabetes ICD9 code 250 for diagnosis 1, 2 or 3 or binary '0' for no diabetes.
Service Utilization	Numeric	Sum of 'number of outpatient', 'number of emergency' and 'number of inpatient'.
Count Medication Change	Numeric	Count of medication changes across 23 medications.

Feature Encoding: The research is designed to predict diabetes hospital readmission risk. Thus, encoding the target variable 'readmitted' is required. The original dataset 'Readmitted' contained the values 'NO' for no readmission, '<30' for less than 30 days and '>30' for greater than 30 days. With the 'Convert Dataset' pill, the 'NO' and '>30' are transformed to a binary '0'. Readmitted '<30' is transformed to binary '1'. With the 'Edit Metadata' pill, a label is assigned to 'readmitted' and the data type is made categorical.

The original dataset contained a predictor variable 'change' indicating if there was a change in diabetes medication or not. With the 'SQL Transformation' pill the 'medication change' predictor variable is updated and encoded with a binary '1' for yes there is a medication change or '0' for no medication change.

'Diagnosis 1', 'Diagnosis 2', 'Diagnosis 3' contain multiple ICD9 codes related to diagnosis groups e.g. circulatory, digestive, diabetes. ICD9 codes related to 250.xx are diabetes. Diagnosis 1, 2 and 3 features are converted from a string to a 3 digit integer value with the 'Edit Metadata' pill in Azure Machine Learning (Microsoft Azure, 2019c). Diagnosis 1, 2 and 3 are then encoded with the 'SQL Transformation' pill code to binary '1' yes diabetes for every ICD9 code with 250 or '0' for no diabetes ICD9 code.

Feature selection: Permutation feature importance is a statistical metric which provides a scored value for each column in the dataset, refer to Fig. 2 and Fig. 3. Predictor variables are randomly selected column by column and the performance of the model is measured before and after. The scores represent the change in the performance of the trained model after permutation. The important predictor variables result in higher importance scores as they are more sensitive to this process. In total 10 predictor variables are selected for the permutation feature importance from the target set for 'Readmitted'. 11 variables are selected for the permutation feature importance for 'Diabetes'. Descriptive values such as 'encounter id' were removed as a patient specific label when predicting 'readmitted' and 'diabetes'. The biggest contributor for 'Readmitted' is 'discharge id'. The biggest contributor to 'Diabetes' is 'Age'. The permutation importance scores the data giving an insight to which feature columns contribute the most in relation to the accuracy of the model. For example, in this research how well the model can predict the readmission of a patient, given demographic features. The metric for measuring performance is 'Classification Accuracy' as it is a binary classification model. The permutation feature importance is selected over Chi squared as there can be limitations to

p-values. If a p-value is < 0.05, the null hypothesis is rejected where there is no statistical significance between the variables and the alternative is accepted where there is a statistically significant relationship between two variables. There is a risk of less than 5% that the hypothesis selected is incorrect (Verhagen, Ostelo and Rademaker, 2004). For this research, permutation feature importance is selected.

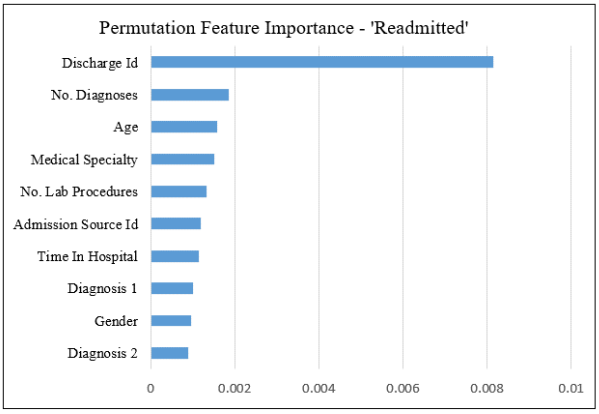

Fig. 2. Permutation Feature 'Readmitted'

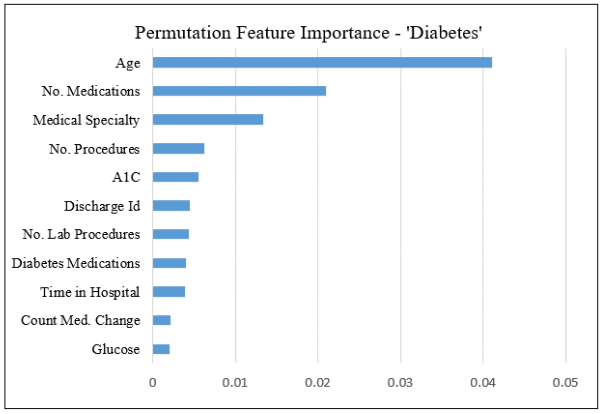

Fig. 3. Permutation Feature 'Diabetes'

In this research, the number of predictor variables selected is less than other researchers as the business level decision would be to simplify the web service app while still achieving strong results from the highest performing model.

Outlier Removal: Each predictor variable has been reviewed focusing on the max and threshold values. The 'Clip values' pill has been applied to deal with the outliers where the peaks are clipped by customising a maximum threshold. The number of emergency patients per year for example is clipped reducing the maximum threshold from 42 to 5 reducing sample skewness and sample kurtosis. Kurtosis measures the shape of the distributed values in the dataset. A high kurtosis value usually have heavy tails or outliers. Positive sample skewness means the distribution is skewed to the right.

SMOTE: SMOTE (Sáez et al., 2015)(Duggal *et al.*, 2016a) is used to deal with class imbalance within datasets. The original dataset pertains no readmission at 54%, readmitted after 30 days at 35% and 11% of the patients are readmitted. This is an imbalanced dataset. Class imbalance means that different categories are not represented evenly. The SMOTE percentage was selected effectively generating the same number of minority classes increasing the number of nearest neighbours at 1. Table 3 illustrates the classes pre and post SMOTE.

Normalization: Normalization is used to scale the numerical values in the datasets. MinMax is selected to rescale each feature to the [0,1] interval. Depending on the algorithm selected, normalization is carried out.

TABLE 3. Final Target Datasets

Dataset name	Dependant Variable	Total Features	Total Rows	Class	Pre Smote	Post Smote
UCI Diabetic Data	Readmitted	11	82,157	0	38,218 (86%)	38,218 (47%)
				1	6,277 (14%)	43,939 (53%)
UCI Diabetic Data	Diabetes	12	83,761	0	42,396 (61%)	42,396 (51%)
				1	27,577 (39%)	41,365 (49%)

Data Mining: The data mining section comprises of exploratory data analysis performed to visualise data, descriptive statistics, linear correlation and predictive model creation for classification problems.

Exploratory Data Analysis: Data mining was initiated by applying Exploratory Data Analysis (EDA) techniques. EDA originated when John Tukey in the 1960's concentrated on easy to draw pictures and arithmetic (Mueller and Tukey, 1980). This research utilises Azure Machine Learning (Microsoft Azure, 2019c), Azure SQL Database[5] and Tableau[6] to explore and visualize the data.

Descriptive Statistics: Table 4 provides significant numbers such as standard deviation, mean, median, max and min for each numeric variable. Examining the statistics helps to identify data types that need to be modified. During this analysis encounter id and patient id are transformed from string to a numeric discrete variable. Diagnosis 1, Diagnosis 2 and 3 are transformed from string to categorical variable. Based on this insight, it was decided that a new feature would be created 'diabetes' which accounted for 39% of primary, secondary and tertiary diabetes diagnoses with the remaining 61% groups related to circulatory, respiratory, digestive, injury, musculoskeletal, genitourinary, neoplasms and other. This led to in-depth analysis in diabetes readmission.

TABLE 4. Descriptive Statistics

Attribute	Description and Values
Age	Mean age at 65, 25%[68-77], 22%[59-68], 16%[77-86], 18%[50-59], 9.8%[41-50], 3.8%[32-42], Other 5.4%.
Gender	Females at 53% and Males at 47%, Unknown less than 1%.
Race	Caucasian at 75%, African American at 18%, Hispanic at 2.1% and Other at 4.9%.
Glucose	95% of patients are not tested, 2.4% abnormal, 2.4% normal.
A1C	82% of patients are not tested, 13% abnormal, 5.0% normal.
Time in Hospital	Mean stay 4.3 days. 32%[1-2] days, 18%[2-4] days, 13%[4-5] days, 17%[5-6] days, 5.7%[6-8] days, Other 14.30%.
No. Medications	Mean medications at 16. 29%[11-18], 22%[6-11], 21%[16-21], 11%[21-26], 6%[1-6], 2.5%[30-35] Other 8.5%.
Medication Change	55% of patients have no medication change, 45% of patients have a medication change.
Diabetes Medication	76% of patients take diabetes medication. 24% of patients do not take diabetes medication.
No. Procedures	Mean procedures is 1. 44% of patients have [0-1] procedures, 20%[1], 13%[2], 10%[3-4], Other 13%.
No. Lab Procedures	Mean lab procedures is 43. 27%[40-53], 21%[27-40], 20%[53-67], 11%[1-14], Other 21%.
No. Diagnoses	Mean diagnoses is 7 with 44% of diagnoses between [9-10], 21%[7-9], 19%[4-6], 11%[6-7], Other 5%.
Medical Specialty	Unknown medical specialty is the largest at 48%. Internal Medicine at 15%.
Diagnosis 1	Circulatory 29.9%, Respiratory 14.2%, Digestive 9.3%, Diabetes 8.2%, Injury 6.9%, Musculoskeletal 4.9%, Other 26.6%.
Discharge ID	29 distinct values e.g. discharge to home, hospice, expired. Discharge to home at 90.5%* and Other(Null) at 9.5%.
Admission Type	9 distinct values. Emergency at 55%, Urgent at 19%, Elective 16%, Null(NA) at 9.5% and Newborn at 0.5%.
Admission Source	21 distinct values. Emergency room at 57%, Physician referral at 27%, Transfers at 8.0%, Null(Other) at 8.0%.
*90.5% include discharge to home 54%, SNF 14%, home health at 13%, short term hospital 5.3%, ICF at 2.3% and inpatient institution 1.9%.	

Linear Correlation Matrix: Linear Correlation is beneficial to measure the relationship between the variables. Azure Machine Learning is used to compute a set of Pearson correlation coefficients for the variables with the dataset for this research. Plotting the data can help visualize the relationship between the variables. The Pearson correlation scale ranges from -1 to 1 where 1 has a strong positive correlation and -1 has a strong negative correlation. 0 signifies there is no linear relationship between the variables. Pearson correlation is sensitive to outliers and performs best with clean normally distributed numeric data. Spearman can be more appropriate if the data is non-linear. Linear correlation is used to identify potential feature columns.

[5] https://azure.microsoft.com/en-in/services/sql-database/
[6] www.tableau.com

Predictive Model Builds: This research uses predictive modelling to identify patterns and make predictions with Azure Machine Learning (Microsoft Azure, 2019c) (Jordan, Kleinberg and Schölkopf, 2006). Azure Machine Learning predictive classification models selected for the purpose of this research include Logistic Regression, Support Vector Machine, Boosted Decision Tree, Decision Forest and Neural Network Models (Microsoft Azure, 2019b). The metrics include Accuracy, Precision, Recall, F1 and AUC values. Accuracy is the number of correctly classified instances. Class imbalance can have a negative effect on accuracy. Thus, steps have been carried out in section 3 to address the class imbalance issue. Precision is the proportion of positives which are correctly classified. Recall is the portion of actual positives that are correctly identified. For example, the percentage of patients with diabetes who are correctly identified with the condition. F1 is a summary metric which takes into account precision and recall. The area under the curve is inspected for the relationship between the true positive rate and false positive rate. Curves nearest the upper left corner have the best classifier performance rate.

3.3 Conclusion

The KDD modified methodology is carried out from selection stage through to evaluation. Pre-processing and transformation stages are crucial for the quality of the algorithm results. Each algorithm is tested on the target data where the metrics to evaluate the predictive model include Accuracy, Precision, Recall, F1 and AUC. The predictive model with the highest accuracy, recall and AUC results are selected and deployed as web applications. The methodology is successfully implemented during this research and is applied to the architectural design specification comprising of 3 tiers in section 4.

4 Design Specification

4.1 Introduction

The architectural design specification for this research is detailed in Fig. 4.

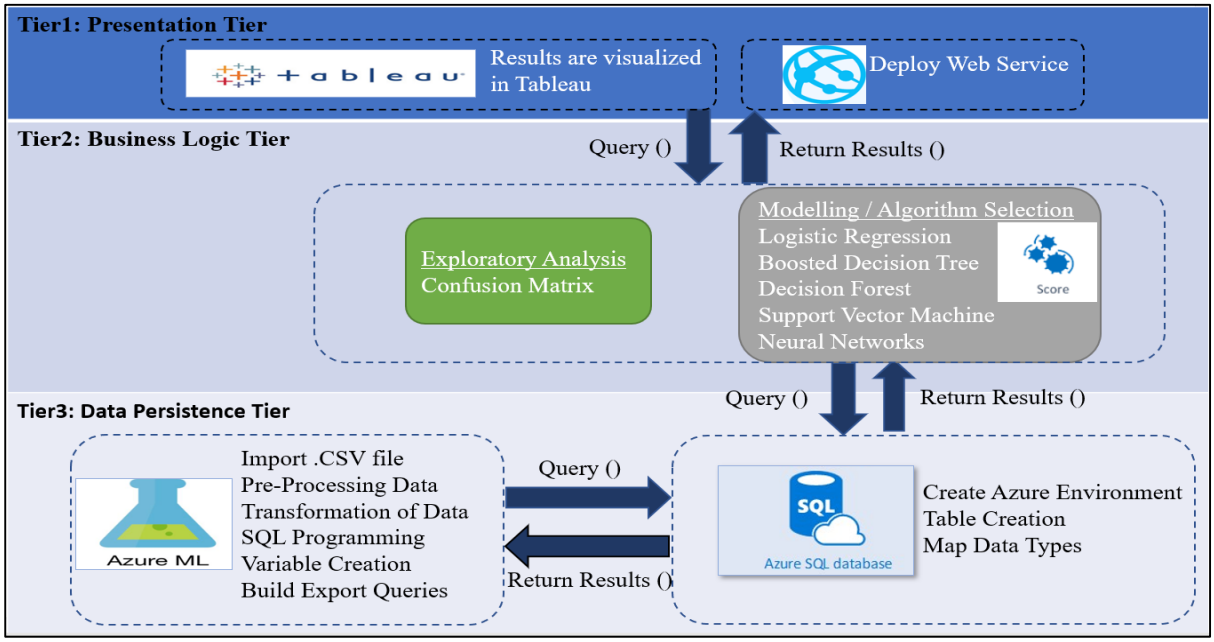

Fig. 4. Architectural Design of the Predictive System

It comprises of 3 tiers (i) Data Persistence Tier with Azure Machine Learning, Azure SQL (ii) Business Logic Tier encompassing exploratory analysis, algorithm selection/modelling and evaluation of results with the confusion matrix as detailed later in this section. (iii) Presentation Tier where the results are visualized in Tableau and the deployment of a web service.

4.2 Predictive Models with Azure Machine Learning Studio

Azure Machine Learning software uses predictive modelling to discover patterns in historical data. It learns from these patterns, so it can automatically make predictions when new data is evident. Azure Machine Learning is a cloud-based platform which integrates seamlessly with Azure SQL.

Logistic Regression is utilized to predict the outcome of a categorical dependant variable based on several predictor variables in this research project. It is a statistical model that tries to fit a line to the data. This technique is used extensively in medical diagnosis. A parameter range is used to tune the model during this research (Hosmer, Lemeshow and Sturdivant, 2013) (Microsoft Azure, 2019b).

Boosted Decision Tree is an ensemble of Decision Trees where the second Decision Tree corrects the errors of the first Decision Tree and so on. They can achieve good accuracy results, but they are memory intensive so may not be suitable to very large datasets. Decision Tree configuration comprises of a parameter range to tune the model (Microsoft Azure, 2019b).

Support Vector Machine is a supervised machine learning technique widely used with classification problems (Yu *et al.*, 2010). Support Vector Machines are classifiers that divide data instances with a linear boundary and a maximum clear gap. Kernels which are nonlinear are used to transform input space into a multidimensional space. The algorithm is suited to two class categorical target variable. For example, if a patient has diabetes or not. The Support Vector Machine configuration selected comprised of a single parameter with the number of iterations at 1, Lambda at .001 and unknown categories allowed. (Microsoft Azure, 2019b).

Decision Forest is a fast supervised ensemble model. Decision Forests are suitable to this research as the aim is to predict two outcomes with both target variables ‘readmitted’ and ‘diabetes’. An ensemble creates and combines multiple trees achieving better results in comparison to a single tree. A bagging resampling method is used during this research (Microsoft Azure, 2019b).

Neural Network is a set of interconnected layers which are used to predict a target with two values. Between the input and output layer there are hidden layers, the number of input nodes for this research matches the number of features in the dataset. The parameter range was used to tune the model. Normalisation is carried out in the Neural Network configuration (Microsoft Azure, 2019b).

4.3 Conclusion

The diabetes diagnosis and readmission risk methodology are carried out during this research project. The 3-tier architecture in Fig. 4 are used during this research with the KDD modified methodology spanning over the 3-tiers. The implementation of diabetes diagnosis and hospital readmission is carried out in the next section.

5 Implementation, Evaluation and Results of diabetes diagnosis and hospital readmission

5.1 Introduction

This section contains the software technologies used during this research, the linear correlation analysis plus the implementation, evaluation and results of individual models used for diabetes diagnosis and readmission predictive modelling. The implementation and configuration are the same for predicting hospital readmission and predicting diabetes. Each model is explained in this section. The implemented models are compared against developed models and existing models, the best performing model is then selected for 'readmitted' and 'diabetes' to create a web service.

5.2 Software Technologies

Microsoft Azure Machine Learning is a cloud based platform which is used to create predictive models for this research. The reason it was selected was due to its visual drag and drop interface, the wide range of well known algorithorms available, web service facilities and the seamless integration with Azure SQL. The software is user friendly and it is easy for a business to understand the workflows in comparison to interrupting code. Tableau was selected due to its powerful interactive visualizations.

5.2.1 Correlation Analysis Results

Linear correlation is carried out to identify potential variables which have a relationship. Fig. 5 shows a total of 7 numeric variables that are selected to compute linear correlation. Non-numeric or descriptive variables are removed such as the 24 generic medication features for example 'Metformin' and 'Repaglinide'. R^2 is denoted in the list of each pair of variables. The results show that variables +1 have a positive linear relationship, -1 have a strong negative linear correlation and 0 denotes a no linear relationship between the two variables. The highest correlation coefficient at .47 and the lowest is -.04 (Microsoft Azure, 2019a). The 'number of procedures' and 'service utilization' have a moderate negative linear correlation at -.04 where the number of 'service utilization' increases while the 'number of procedures' decreases. 'Number of procedures' and 'age' have a moderate negative linear correlation at -.02. As age increases, the number of procedures decrease from the age of 75. 'Time in hospital' and the 'number of lab procedures' have a moderate positive linear correlation at .33. The number of lab procedures increases as the time in hospital increase. 'Time in hospital' and 'number of medications' have a moderate positive linear correlation at .47. The number of medications increase as the time in hospital increases. The moderate result can indicate that some points are close to the line while other points are far from the line resulting in a moderate linear relationship between the variables. Logistic Regression is suited to binary classification problems over Linear Regression. Thus, as the target variable is binary, Logistic Regression has been selected. The Pearson correlation matrix have the same features for 'Readmitted' and 'Diabetes' as the numeric features are selected, all others are omitted.

Pearson Correlation Matrix

	Service Utilization	Age	Time in Hospital	No. Lab Procedures	No. Procedures	No. Medications	No. Diagnoses	
Service Utilization	1.00	0.01	0.01	0.03	-0.04	0.04	0.10	0.50
Age	0.01	1.00	0.13	0.03	-0.02	0.06	0.25	0.40
Time in Hospital	0.01	0.13	1.00	0.33	0.19	0.47	0.23	0.30
No. Lab Procedures	0.03	0.03	0.33	1.00	0.04	0.26	0.15	0.20
No. Procedures	-0.04	-0.02	0.19	0.04	1.00	0.40	0.09	0.10
No. Medications	0.04	0.06	0.47	0.26	0.40	1.00	0.26	0.00
No. Diagnoses	0.10	0.25	0.23	0.15	0.09	0.26	1.00	-0.10

Fig. 5. The Pearson's coefficient score for each of the 7 variables selected.

5.2.2 Evaluation and results of Classification Algorithms

In this section, a comparison of classification models is carried out based on the main metrics Accuracy, Recall and AUC as per Fig. 8 and Fig. 9. Precision and F1 values are supporting metrics during this research. The metrics are calculated with the confusion matrix components[7]. For example, True Positive (TP) is when a patient has diabetes and they are correctly identified by the classifier. True Negative (TN) is when a patient does not have diabetes and they are correctly identified by the classifier. False Positive (FP) is when a patient does not have diabetes, but the diagnosis is positive. False Negative (FN), is when a patient has diabetes, but the diagnosis is negative. It is of critical importance to identify false negatives in healthcare.

In this research, the following equations are used to measure Accuracy, Recall, Precision and F1. Accuracy Eq. (1) is the goodness of a classification model and is the proportion of correctly identified instances. Recall Eq. (2) is the fraction of all correct results returned by the model. Precision Eq. (3) is the portion of true results over all positive results. F1-score Eq. (4) is computed as the weighted average of precision and recall between 0 and 1, where the ideal F-score value is 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1 = 2.(Precision.Recall)/(Precision + Recall) \quad (4)$$

AUC measures the area under the curve plotted with true positives on the Y axis and false positives on the X axis. This metric is useful when comparing different models.

[7] https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance

5.3 Implementation, Evaluation and Results

5.3.1 Implementation, Evaluation and Results of Logistic Regression

The Logistic Regression model is a supervised learning method used to predict one of two outcomes and is a popular choice within healthcare.

Implementation: The Logistic Regression model is built with a parameter range while tuning the modelling hyperparameter to reach optimal performance. The model is trained on the training data. The 'Tune Model Hyperparameters' pill random sweep has been selected with a maximum of 5 runs increasing model performance and to conserve computing resources. The data is split into training and testing datasets in a 70:30 ratio i.e. 70% train data and 30% test data. The metric accuracy, recall and AUC are used to evaluate the model.

Evaluation and Results:

Readmitted - The model performance was lower than the Decision Tree model with accuracy at 76% and AUC at 84%. Recall is the lowest result at 72% as per Fig. 8.

Diabetes – The model performance came in third when compared to the other classifiers with accuracy at 64% and AUC at 70% as per Fig. 9. Recall at 62% had a lower performance when compared other classifiers. Overall the Logistic Regression classifier performance is good.

5.3.2 Implementation, Evaluation and Results of Boosted Decision Tree

Boosted Decision Trees are an ensemble of Decision Trees where the second tree corrects the error of the first, the third corrects the error of the first two trees and so on. The results are based on the ensemble of the trees. The trees are visualized by clicking the train model output.

Implementation: The Boosted Decision Tree parameter range is selected to run a parameter sweep. A range of values for the maximum number of leaves per tree, minimum number of samples per leaf node and learning rate are selected to repeat and reiterate. The 'Tune Model Hyperparameters' is then selected to do a random sweep by performing a parameter sweep over the range of parameters with a maximum of 5 runs, learning the optimal set of hyperparameters increasing model performance and conserving computing resources. Implementation is memory intensive, but it is suitable to this dataset. Normalization is not required. The Boosted Decision Tree is trained on the training data. The data is split into training and testing datasets in a 70:30 ratio i.e. 70% train data and 30% test data. The metric accuracy, recall and AUC are used to evaluate the model. The Boosted Decision Tree algorithm is important as it is easy for the business to understand and interrupt.

Evaluation and Results:

Readmitted - The entropy and information gain split on 'discharge id' 2 which relates to patients which are discharged to another short-term hospital. By following the logic of the tree, we can see the node which has the highest probability of being readmitted in less than 30 days are patients 'discharge_id.2' where the patient is not discharged to another short-term hospital and the patient has circulatory related primary diagnosis (ICD9 code 410). The Boosted Decision Trees classifier achieved top performance results with accuracy at 86%, Recall at 87% and AUC at 92% as per Fig. 8.

Diabetes – The Decision Tree split on age due to entropy and information gain. Following the logic of the tree we can see that the node with the highest probability of being diagnosed with diabetes are patients aged between 40 and 60 with the number of medications taken greater

than 12.5, the count of medication changes is greater than .22 and the number of procedures is greater than 3.5. The Boosted Decision Tree model performance is high with accuracy at 67%, Recall at 66% and AUC at 74% as per Fig. 9. Fig. 6 and Fig. 7 illustrates Boosted Decision Tree and Logistic Regression AUC results compared.

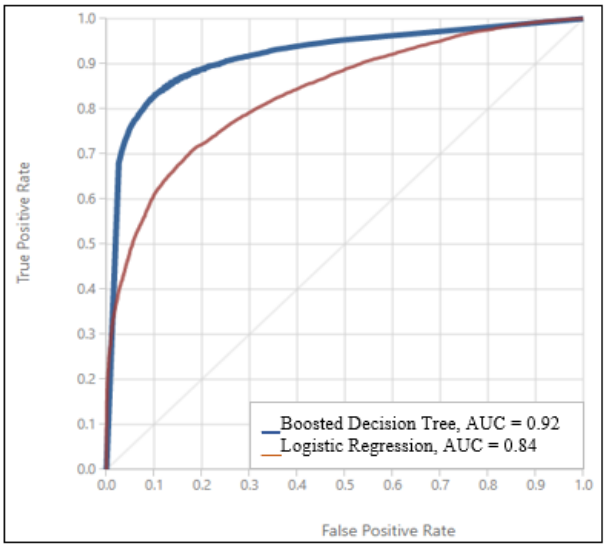

Fig. 6. 'Readmitted' BDT and LR compared

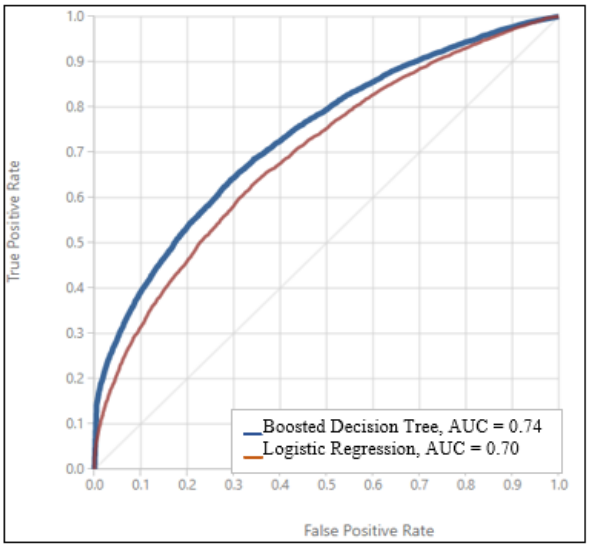

Fig. 7. 'Diabetes' BDT and LR compared

5.3.3 Implementation, Evaluation and Results of Decision Forest

The supervised ensemble learning algorithm are fast and have many advantages such as capturing non linear boundaries. They are efficient in terms of memory usage, they are not sensitive to noisy data and they can handle data with varied distributions.

Implementation: For this research the Decision Forest is built from multiple Decision Trees and then voting is carried out on the most popular output class. Configuration of the two-class Decision Forest with bagging is selected. The 'single parameter' is selected to determine the maximum values for the number of Decision Trees at 8, the maximum depth of the Decision Trees at 50, number of random splits per node at 150 and the minimum number of samples per leaf nodes is 1. The Decision Forest is trained on the training data. The data is split into training and testing datasets in a 70:30 ratio i.e. 70% train data and 30% test data. The 'Train Model' pill is selected as the 'Tune Model Hyperparameters' was computationally intensive. The metric accuracy, recall and AUC are used to evaluate the model.

Evaluation and Results:

Readmitted – The Decision Forest model was the second-best performing classifier with accuracy at 77%, recall at 70% and AUC at 85% as per Fig. 8.

Diabetes – The Decision Forest model performance is average when predicting diabetes with accuracy at 66%, recall at 62% and AUC at 72% as per Fig. 9.

Overall the Decision Forest classifiers are in the top 3 performing models.

5.3.4 Implementation, Evaluation and Results of Neural Networks

A classification method using Neural Networks require a label column which is suitable for this research as the target variables are binary. A Neural Network includes input nodes, a set of hidden layers and output nodes.

Implementation: The Neural Network model is built with a 'Parameter Range' which is fully connected case where the input layer is fully connected to the hidden layers and the hidden layers are fully connected to the output layer and the number of hidden layers is the same as the number of features in the training data. The 'use range builder' is selected for the 'learning rate' which is the step taken by each iteration at a maximum of 450. The 'min-max normalizer' has been selected to rescale the data to the [0,1] range. The Neural Network is trained on the training data. The 'Tune Model Hyperparameters' pill random sweep are selected with a maximum of 5 runs increasing model performance while conserving computing resources. The data is split into training and testing datasets in a 70:30 ratio i.e. 70% train data and 30% test data. The Neural Network is trained on the training data. The metric accuracy, recall and AUC are used to evaluate the model.

Evaluation and Results:

'Readmitted' The 'Tune Model Hyperparameters' pill achieved higher results in comparison to the 'Train Model' method. The model performance is fair in comparison to other classifiers with accuracy at 76%, recall at 78% and AUC at 84% as per Fig. 8.

'Diabetes' The Neural Network model achieved the high accuracy results at 66% with the Boosted Decision Tree model accuracy 1% more at 67%. The Neural Network has the highest achieved recall results at 68% and AUC at 73% as per Fig. 9.

Overall the Neural Network classifiers are high performing models.

5.3.5 Implementation, Evaluation and Results of Support Vector Machine

SVM are classifiers that divide data instances which are from different classes with a linear boundary with a clear gap in between (Cortes and Vapnik, 1995).

Implementation: The SVM model is built with a single parameter, 1 iteration is used when building the model and Lambda at .001 is used to tune the model. Normalize features are selected during the SVM build. The single parameter achieved higher results than the parameter range after numerous iterations. The data is split into training and testing datasets in a 70:30 ratio i.e. 70% train data and 30% test data. The SVM is trained on the training data. The 'Tune Model Hyperparameters' pill random sweep has been selected with a maximum of 5 runs increasing model performance and conserve computing resources. The metric accuracy, recall and AUC are used to evaluate the model.

Evaluation and Results:

Readmitted - The 'Tune Model Hyperparameters' pill achieved higher results in comparison to the 'train model' method. Metrics are low in comparison to other classifiers with accuracy at 74%, Recall at 69% which is the lowest recall result of all classifiers and AUC at 82% as per Fig. 8.

Diabetes – The SVM accuracy results are 64%, recall of 61% and AUC 70% which are the lowest performing results in comparison to other classifiers as per Fig. 9.

Overall the SVM classifiers are the lowest performing classifiers.

5.4 Comparison of Developed Models

The developed models are compared in Fig. 8 and Fig. 9. The Boosted Decision Tree model achieves the highest results when predicting hospital readmission for accuracy, recall and AUC. The Boosted Decision Tree model also achieves the highest result for predicting diabetes

diagnosis with accuracy and AUC. The Neural Network model achieves the highest result for recall when predicting diabetes diagnosis.

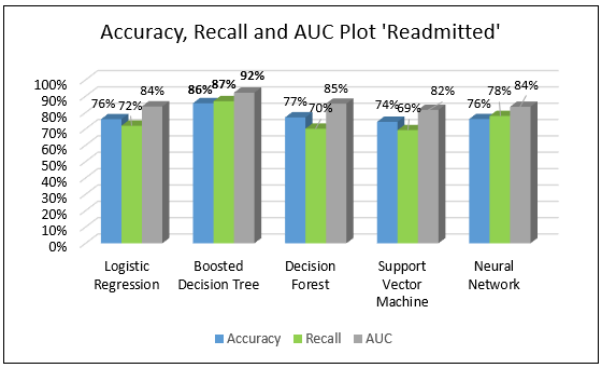

Fig. 8. 'Readmitted' Accuracy and Recall Plot

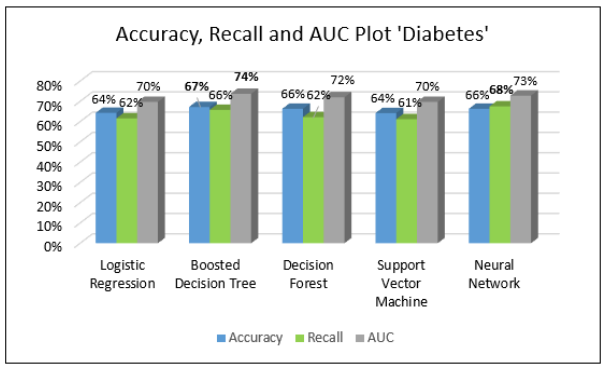

Fig. 9. 'Diabetes' Accuracy and Recall Plot

Table 5 and Table 6 demonstrate the results of all metrics. The highest achieving classifiers are the Boosted Decision Trees. The Boosted Decision Tree models are selected for the web service and for export to Tableau for visualization.

TABLE 5. Comparison of results 'Readmitted'

Predicting Readmission	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	76%	81%	72%	76%	84%
Boosted Decision Tree	**86%**	**87%**	**87%**	**87%**	**92%**
Decision Forest	77%	84%	70%	77%	85%
Support Vector Machine	74%	80%	69%	74%	82%
Neural Network	76%	77%	78%	78%	84%

TABLE 6. Comparison of Results 'Diabetes'

Predicting Diabetes Diagnosis	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	64%	65%	62%	63%	70%
Boosted Decision Tree	**67%**	**67%**	66%	**67%**	**74%**
Decision Forest	66%	**67%**	62%	65%	72%
Support Vector Machine	64%	65%	61%	63%	70%
Neural Network	66%	65%	**68%**	66%	73%

5.5 Comparison of Developed Models with Existing Models

The comparison Table 7 and Table 8 contain literature review methods which are compared to my research.

TABLE 7. Comparison of existing models/literature review 'Readmitted'

Method	Accuracy	Reference
Multilayer Perceptron	95%	Goudjerkan and Jayabalan, 2019
Boosted Decision Tree	**86%**	This Study
C5.0, Support Vector Machine	81%-85%	Turgeman and May, 2016
Neural Networks	80%	Baskaran *et al.*, 2011
Support Vector Machine	78%	Zheng *et al.*, 2015

TABLE 8. Comparison of existing models/literature review 'Diabetes'

Method	Accuracy	Reference
J48 Decision Trees	100%	Kandhasamy and Balamurali, 2015
K-means, Logistic Regression	95.42%	Wu *et al.* , 2018
K-means, Decision Trees	92.38%	Patil, Joshi and Toshniwal, 2010
J48 Decision Tree	90.04%	Chen et al., 2017
Artifical Neural Network	89%	Komi et al., 2017
C5.0 Decision Tree	77.87%	Meng et al., 2013
Proposed Bayes Network	72.3%	Guo, Yang; Bai, Guohua; Hu, 2012
Boosted Decision Tree	**67%**	This Study

5.6 Design, Implementation and Evaluation of Web Services

Azure Machine Learning has the capabilities to develop, run, test and iterate a predictive model. Once the predictive models are created in Azure Machine Learning, the one model with the maximum accuracy and recall is selected. The aim of the web service in this research project is to deploy it as a classic Azure Machine Learning web service so that potential users can input new data and receive results instantly. The build is a simple, user friendly web service consisting of simple questions. The answers are either an input of a number or an input is made available via a drop-down menu. The web service is created in Microsoft Azure Machine Learning and the request response web application.

Design and Implementation: The Boosted Decision Tree training experiment is created, and optimal results are obtained. The input parameters, for example age, medical speciality, time in hospital are fed into the model which returns a prediction value and label. For this research a prediction value is returned if a patient has diabetes or not and if a patient is readmitted or not. Once the prediction function has been established, the trained model is converted to a predictive experiment. By doing so, the trained model is ready to be deployed as a scoring web service. Users of the 'readmitted' web service can input data and scored results output data will be returned whether the patient will be readmitted in 30 days or not. Users of the 'diabetes' web service can input data and scored results output data are returned if the person has diabetes or not based on the algorithm learning from the dataset.

Evaluation: The target variable is excluded as only the required input variables in the web application and the scored results are selected. The training experiment is converted to predictive model and deployed as a web service. The test web service request response page displays the inputs and outputs[8] as per Fig. 10 and Fig. 11. The web service is then called directly from excel where new data can be entered as the input, returning output results.

[8] https://docs.microsoft.com/en-us/azure/machine-learning/studio/publish-a-machine-learning-web-service

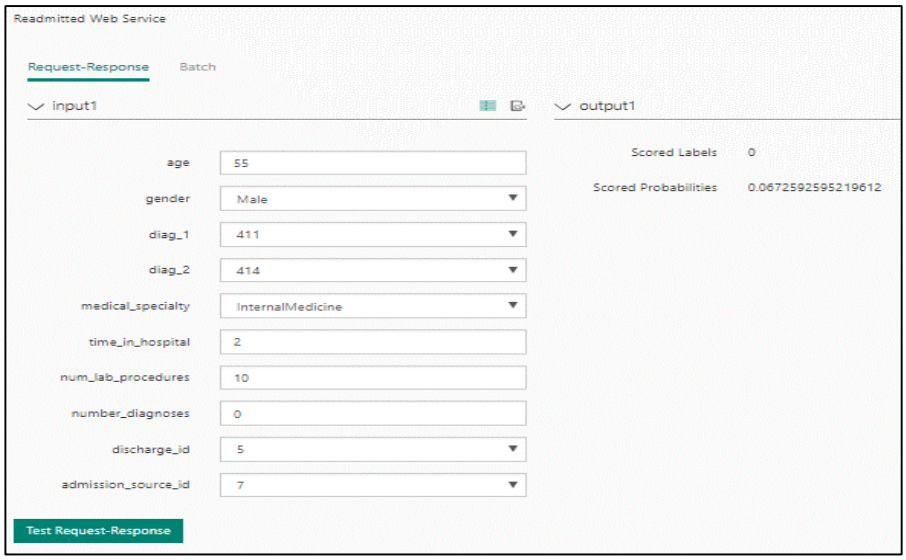

Fig. 10. 'Readmitted' Web Service

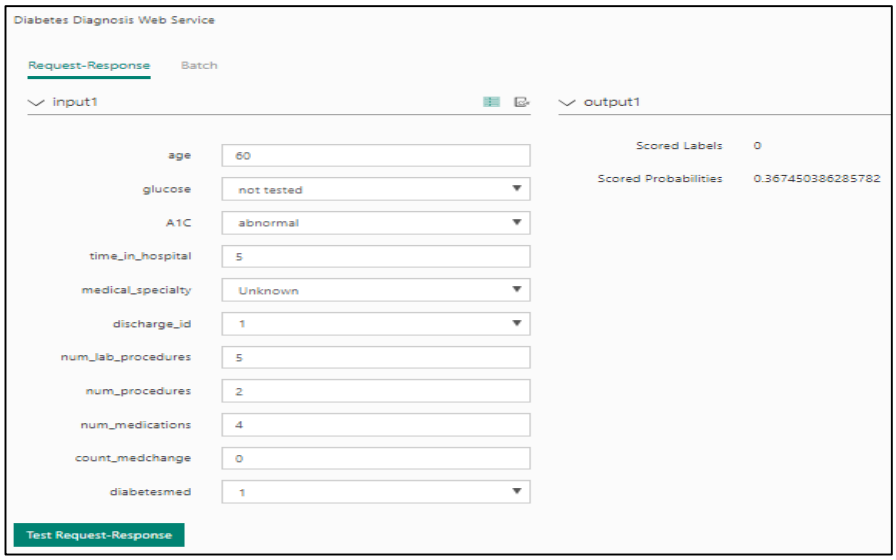

Fig. 11. 'Diabetes' Web Service

5.7 Prediction Model Results

After extracting, cleaning and transforming the data, the final scored data from the highest performing algorithm is exported to Azure SQL and analysed using Tableau. When reviewing the Boosted Decision Tree AUC at 92% as per Fig. 6, the point which has the largest distance from the curve should be examined. The scored probabilities at this point should be investigated, the data can be sorted and the encounters with the highest probabilities for diabetes diagnosis or hospital readmission can be identified. Thus, the healthcare service providers can make informed and improved decisions based on the results.

5.8 Implementation, Evaluation and Results Conclusion

The predictive models used for diabetes diagnosis and readmission risks predictive modelling: USA are implemented successfully. They are tuned to achieve optimal results. All the models are tested on the target data. Accuracy, recall and AUC are the main metrics used to evaluate the performance of the model. The best performing models, Boosted Decision Trees achieve

the highest scores and are selected and deployed as a web service in Azure. The data is also exported to Azure SQL and Tableau for analysis. The model appears to be a balanced dataset with a high AUC of 92% for 'readmitted' and AUC of 74% for 'diabetes'.

6 Discussion

During this research, the number of predictor variables are lower in comparison to other researchers. High model performance results can be achieved with less features. The business level decision would be to select the lower number of features ensuring a simple, user-friendly web service. This research also highlights that data can be used for more than one purpose to gain insights into related issues. Age, medical specialty, number of procedures and time in hospital are important features for both predicting diabetes diagnosis and hospital readmission.

7 Conclusion and Future Work

To cope with the challenge of diabetes diagnosis and unplanned 30-day hospital readmissions, this study implemented a detailed pre-processing and transformation framework to improve data quality producing high accuracy, recall and AUC results. The framework includes feature importance and selection, data cleaning, feature engineering, SMOTE and normalisation to optimize features for diabetes diagnosis and hospital readmission.

The proposed Boosted Decision Tree with tuned hyperparameters obtains the highest results when predicting diabetes diagnosis and readmission risk outperforming other machine learning algorithms. The model was found to be balanced against all metrics including accuracy, recall, precision, F1 and AUC. This research demonstrates that predictive modelling can be used to predict diabetes diagnosis and hospital readmission risk. Diabetes diagnosis and 30-day hospital readmission is of critical importance to health service providers and to patient quality of care. Pre-processing and transformation is the key to success and this research selected a comprehensive methodical approach to ensure accurate predictions. Features are selected through statistical analysis in addition to domain knowledge. All the predictive models were tested for accuracy, recall and AUC. The top performing algorithm Boosted Decision Trees are selected to produce two working web services in Azure Machine Learning. Health service providers can visualize data in two forms. In Tableau where the highest scored patients can be targeted or in an Azure Web Service where new data inputs will result in scored outputs.

The objective of this research was to add to the body of knowledge of diabetes diagnosis and hospital readmission. There are predictive modelling opportunities when working with diabetes hospital data. Azure Machine Learning improves the timeliness of information bringing the data to the health service providers where intelligent decisions can be made.

ACKNOWLEDGEMENT

The author would like to thank:

My supervisor Dr Catherine Mulwa.

My husband, Brian for the unlimited support.

My twin boys, Aaron and Fionn for the hugs when it's past their bedtime.

My classmates for their encouragement and who made it a more enjoyable experience.

References

Azevedo, A. and Santos, M. F. (2008) *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. Available at: https://pdfs.semanticscholar.org/7dfe/3bc6035da527deaa72007a27cef94047a7f9.pdf (Accessed: 21 June 2019).

Baskaran, V. *et al.* (2011) 'Predicting Breast Screening Attendance Using Machine Learning Techniques', *IEEE Transactions on Information Technology in Biomedicine*, 15(2), pp. 251–259. doi: 10.1109/TITB.2010.2103954.

Chen, W. *et al.* (2017) 'A hybrid prediction model for type 2 diabetes using K-means and decision tree', in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, pp. 386–390. doi: 10.1109/ICSESS.2017.8342938.

Clore, J. N. *et al.* (2014) *UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 (Accessed: 17 May 2019).

CMS (2019) 'Readmissions-Reduction-Program'. Available at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html (Accessed: 30 May 2019).

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*. Kluwer Academic Publishers, 20(3), pp. 273–297. doi: 10.1007/BF00994018.

Duggal, R. *et al.* (2016a) '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*, 36(4), pp. 469–476. doi: 10.1007/s13410-016-0495-4.

Duggal, R. *et al.* (2016b) 'Predictive risk modelling for early hospital readmission of patients with diabetes in India', *International Journal of Diabetes in Developing Countries*, 36(4), pp. 519–528. doi: 10.1007/s13410-016-0511-8.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine*, 17(3), pp. 37–37. doi: 10.1609/AIMAG.V17I3.1230.

Feyyad, U. M. (1996) 'Data mining and knowledge discovery: making sense out of data', *IEEE Expert*, 11(5), pp. 20–25. doi: 10.1109/64.539013.

Goodney, P. P. *et al.* (2003) 'Hospital volume, length of stay, and readmission rates in high-risk surgery.', *Annals of surgery*. Lippincott, Williams, and Wilkins, 238(2), pp. 161–7. doi: 10.1097/01.SLA.0000081094.66659.c3.

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*, 10(2). doi: 10.14569/IJACSA.2019.0100236.

Guo, Yang; Bai, Guohua; Hu, Y. (2012) 'Using Bayes Network for Prediction of Type-2 Diabetes', *The 7th conference for internet technology and secured transactions (ICITST-2012)*. IEEE. Available at: https://ieeexplore.ieee.org/document/6470852 (Accessed: 3 May 2019).

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013) *Applied Logistic Regression*. Available at: https://learning.oreilly.com/library/view/applied-logistic-regression/9781118548356/ (Accessed: 30 July 2019).

Jain, A. K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*. North-Holland, 31(8), pp. 651–666. doi: 10.1016/J.PATREC.2009.09.011.

Al Jarullah, A. A. (2011) 'Decision tree discovery for the diagnosis of type II diabetes', in *2011 International Conference on Innovations in Information Technology*. IEEE, pp. 303–307. doi: 10.1109/INNOVATIONS.2011.5893838.

Jencks, S. F., Williams, M. V. and Coleman, E. A. (2009) 'Rehospitalizations among Patients in the Medicare Fee-for-Service Program', *New England Journal of Medicine*. Massachusetts Medical Society , 360(14), pp. 1418–1428. doi: 10.1056/NEJMsa0803563.

Jordan, M., Kleinberg, J. and Schölkopf, B. (2006) *Pattern Recognition and Machine Learning*. Available at: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf (Accessed: 30 July 2019).

Kaggle (2016) *Pima Indians Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database*. Available at: https://www.kaggle.com/uciml/pima-indians-diabetes-database (Accessed: 2 December 2018).

Kandhasamy, J. P. and Balamurali, S. (2015) 'Performance Analysis of Classifier Models to Predict Diabetes Mellitus', *Procedia Computer Science*. Elsevier, 47, pp. 45–51. doi: 10.1016/J.PROCS.2015.03.182.

Komi, M. *et al.* (2017) 'Application of data mining methods in diabetes prediction', in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, pp. 1006–1010. doi: 10.1109/ICIVC.2017.7984706.

Maddipatla, R. M. *et al.* (2015) '30 Day hospital readmission analysis', in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 2922–2924. doi: 10.1109/BigData.2015.7364123.

Meng, X.-H. *et al.* (2013) 'Comparison of three data mining models for predicting diabetes or prediabetes by risk factors', *The Kaohsiung Journal of Medical Sciences*. Elsevier, 29(2), pp. 93–99. doi: 10.1016/J.KJMS.2012.08.016.

Microsoft Azure (2019a) *Compute Linear Correlation - Azure Machine Learning Studio | Microsoft Docs*. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation (Accessed: 23 July 2019).

Microsoft Azure (2019b) *Machine Learning - Initialize Model - Classification - Azure Machine Learning Studio | Microsoft Docs*. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification (Accessed: 23 July 2019).

Microsoft Azure (2019c) *Microsoft Azure Machine Learning Studio*. Available at: https://studio.azureml.net/ (Accessed: 1 July 2019).

Mueller, C. W. and Tukey, J. W. (1980) 'Exploratory Data Analysis.', *Administrative Science Quarterly*, 25(4), p. 700. doi: 10.2307/2392291.

Negi, A. and Jaiswal, V. (2016) 'A first attempt to develop a diabetes prediction method based on different global datasets', in *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE, pp. 237–241. doi: 10.1109/PDGC.2016.7913152.

Patil, B. M. M., Joshi, R. C. C. and Toshniwal, D. (2010) 'Hybrid prediction model for Type-2 diabetic patients'. Pergamon, 37(12), pp. 8102–8108. Available at: https://www.sciencedirect.com/science/article/pii/S0957417410004896 (Accessed: 2 December 2018).

Rubin, D. J. *et al.* (2014) 'Early readmission among patients with diabetes: A qualitative assessment of contributing factors', *Journal of Diabetes and its Complications*. Elsevier, 28(6), pp. 869–873. doi: 10.1016/J.JDIACOMP.2014.06.013.

Rubin, D. J. (2018) ‘Correction to: Hospital Readmission of Patients with Diabetes’, *Current Diabetes Reports*. Springer US, 18(4), p. 21. doi: 10.1007/s11892-018-0989-1.

Sáez, J. A. *et al.* (2015) ‘SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering’, *Information Sciences*, 291, pp. 184–203. doi: 10.1016/j.ins.2014.08.051.

Strack, B. *et al.* (2014) ‘Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records’, *BioMed Research International*. Hindawi, 2014. doi: 10.1155/2014/781670.

Turgeman, L. and May, J. H. (2016) ‘A mixed-ensemble model for hospital readmission’, *Artificial Intelligence in Medicine*. Elsevier, 72, pp. 72–82. doi: 10.1016/J.ARTMED.2016.08.005.

Verhagen, A. P., Ostelo, R. W. and Rademaker, A. (2004) ‘Is the p value really so significant?*’, *Australian Journal of Physiotherapy*. Elsevier, 50(4), pp. 261–262. doi: 10.1016/S0004-9514(14)60122-7.

Vijayan, V. V. and Anjali, C. (2015) ‘Prediction and diagnosis of diabetes mellitus — A machine learning approach’, in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, pp. 122–127. doi: 10.1109/RAICS.2015.7488400.

WHO (2017) ‘WHO | Global report on diabetes’, *WHO*. World Health Organization. Available at: http://www.who.int/diabetes/global-report/en/ (Accessed: 1 December 2018).

Wu, H. *et al.* (2018) ‘Type 2 diabetes mellitus prediction model based on data mining’, *Informatics in Medicine Unlocked*. Elsevier, 10, pp. 100–107. doi: 10.1016/J.IMU.2017.12.006.

Yu, W. *et al.* (2010) ‘Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes’, *BMC Medical Informatics and Decision Making*. BioMed Central, 10(1), p. 16. doi: 10.1186/1472-6947-10-16.

Zhao, P. and Yoo, I. (2017) ‘A self-adaptive 30-day diabetic readmission prediction model based on incremental learning’, in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 895–898. doi: 10.1109/BIBM.2017.8217775.

Zheng, B. *et al.* (2015) ‘Predictive modeling of hospital readmissions using metaheuristics and data mining’, *Expert Systems with Applications*, 42(20), pp. 7110–7120. doi: 10.1016/j.eswa.2015.04.066.