National College of Ireland

Prediction of Length of Stay and Hospital Readmission for Diabetic Patients

MSc Research Project
Data Analytics

Silky Jain
Student ID: x19213590

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Silky Jain
Student ID:	x19213590
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	23/09/2021
Project Title:	Prediction of Length of Stay and Hospital Readmission for Diabetic Patients
Word Count:	8208
Page Count:	22

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:	Silky Jain
Date:	23rd September 2021

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):	

Prediction of Length of Stay and Hospital Readmission for Diabetic Patients

Silky Jain
x19213590

Abstract

Unplanned readmission and unexpected long stay of diabetic patients is one of the biggest problems of the hospitals. This impacts the management and patient care services provided at hospitals. For efficient management and optimal utilization of hospital resources such as bed allocation, diagnoses test labs etc., it becomes necessary to predict the risk of readmission and length of stay of patients. This research proposed a stacked ensemble learning model with a comprehensive methodological approach including data cleaning and transformation techniques like feature encoding, selection, normalization, feature importance etc., to predict readmission and length of stay of patients. The motive is to ensure the quality of data to produce a highly effective and efficient predictive model that helps in saving high healthcare expenses. The models like Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR), k Nearest Neighbors (kNN), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB) are used as base models to build the stacked model based on the performance and parameter tuning of the models. The final stacked ensemble model not only outperforms all the base classifiers but also the existing work in the same field. For hospital readmission prediction stacked model gives the accuracy, precision, specificity, F1Score, and AUC of 94%, 90%, 95%, 90% and 98% respectively and for the prediction of length of stay gives 90% for each evaluation metric. However, the research could include Deep learning for better models and results which is addressed in future works.

Keywords— *Healthcare, Diabetes, Predictive Analytics, Hospital Readmission, Length of Stay, Stacked Ensemble Learning Model, Data Pre-processing*

Contents

1 Introduction

Diabetes is one of the chronic diseases which usually leads to one or more organs failure and it happens because the person's body is not producing insulin properly. Insulin is required to regulate blood sugar levels, and if the sugar level increases significantly in blood then it causes diabetes. Diabetes is classified into two categories - Type 1 and Type 2. Among adults Type 2 diabetes is becoming common and the percentage of people having diabetes is increasing at an alarming rate. Thus, it is becoming one of the major areas of concern and from past decades it is considered as a global threat[1]. As diabetes is a prolonged and recurring disease so the patients suffering from them have a higher chance of getting readmitted to the hospital and might stay for a longer duration in hospitals to get tested. To predict such scenarios machine learning algorithms can be used.

The rate of readmission cases seems to increase significantly over the past decades thus leading to the failure of efficient working of hospital staff and management. Therefore, healthcare service providers and centres are now taking more interest in predicting and knowing the factors causing the readmissions and longer stay of patients at the hospital to improve the patient care services and potential savings by cutting irrelevant costs (Bhuvan et al.; 2016).

The readmission rate can be explained as the number of patients who get readmitted to the same or different hospitals within a specific period of discharge after the first diagnosis. Length of stay (LOS) can be explained as the duration of a patient's stay at the hospital, i.e. the time at which the patient was admitted to the patient's discharge time. Both the factors, readmission and LOS are significant while assessing the quality and efficiency of hospital services. Hospitals often face limitations in terms of staff and resources such as wards, beds, labs etc. in case of medical emergencies or crises. Due to the unavailability of test labs, the patient has to stay at the hospital for a longer period which eventually increases hospitalization expenses. Thus, there should be more research in this area to manage hospital expenses and improve patient care services.

1.1 Background and Motivation

Diabetes is a well-known global threat where many researchers have investigated and applied various machine learning approaches and data mining techniques to open large datasets consisting of the details of the patient record. However, there still need to be research done concentrating more on the prediction of hospital readmission rate and length of stay of patients. It has been stated that diabetic patients are more prone to the risk of readmission and stays for longer duration at hospitals due to the number of tests, subsequently increasing the operation time and cost (Comino et al.; 2015; Sharma et al.; 2013).

There has been some research in the same domain using different techniques and algorithms like Decision Tree (Zheng et al.; 2015), Support Vector Machine (SVM) (Cui et al.; 2018), Logistic Regression (Rosen et al.; 2017) and k-Nearest Neighbors. These approaches might have provided good results but they lack in generalization of application and coping with the uncertainty of the readmission case and length of stay of patients. This happens because of the inconsistency in the way of storing data by the health care systems and the lack of understanding of the data. Medical data is often interrelated and high in dimensionality thus it becomes really difficult to deal with it. High dimension increases the computation complexity and might result in a decrease in the performance of the model.

There are a few challenges and limitations which has been observed while building a model for predicting using the medical dataset. One of the challenges is the usage of the medical codes and high dimensionality which causes problems while mining the data. This was also the case with the dataset used for this research where there were diagnostic codes were used for test

[1] `https://www.who.int/news-room/fact-sheets/detail/diabetes`

results variables. Another challenge is the problem of class imbalance due to less number of cases for readmission cases in the dataset, which affects the predictive ability of the model. In addition to that, in the past various ML algorithms were applied for the same problem, but the comparison of different classifiers didn't result in any best or suitable model conquering all the challenges. It is not justified to compare one model over the other while every algorithm has its strength and weakness (Alves et al.; 2018). The ensemble learning approach helps in improving the predictive ability of single classifiers. While selecting models interpretability, reliability and generalizability are considered, which are then stacked together averaging out each other's limitations and providing a better model. The whole framework of the proposed model including data cleaning, sampling, feature engineering, encoding and the selection of base models resulted in a final stacked ensemble predictive model.

A stacked ensemble model with all the comprehensive data pre-processing and transformation techniques is considered a novel approach as this methodology was used with the motive of building a good predictive model which overcomes all the challenges. Also, to the best of my knowledge stacked ensemble learning was not applied for predicting both the hospital readmission and length of stay of patients, so this approach was used and implemented in this project.

1.2 Research Question

RQ: "How well can stacked ensemble learning method predict the hospital readmission rate and length of stay of diabetic patients? "

1.3 Research Objectives

- To identify the factors which are influencing the readmission rate and length of stay of diabetic patients in the hospital.
- To improve the quality of data by applying comprehensive data cleaning and pre-processing techniques such that it becomes easier for the model to process the data and provide better results.
- To examine data and observe the useful patterns by performing exploratory data analysis on the dataset.
- To implement the base models and build the stacked ensemble model with a motive to get accurate predictions.
- To assess the performance of the models implemented using the metrics – Accuracy, recall, AUC, specificity and F1 Score.
- To examine the shortcomings of the proposed model and discuss the future work which can help in enhancing the performance of the model.

1.4 Document Structure

This report illustrates the research project in the following sections: Section 2 gives an overview of the existing studies in the same domain which is further divided into subsections based on the type of problem addressed and methodology used - hospital readmission, length of stay and application of stacked ensemble models. Section 3 provides the methodology with a process flow to develop a model for this research to meet research objectives. Section 4 provides the architecture for the proposed model and Section 5 shows the implementation of the model and discusses the results of each model. Section 6 gives the evaluation of model based on metrics. Section 7 concludes the research by discussing the results and limitations of the work.

2 Related Work

There has been recent developments and exploration in predicting hospital readmission and length of stay of patients in the hospital in detail. There were some useful insights obtained on the origin and evolution of the problem over the years and the application of ML in solving the issue from the literature review. The ideology of this research is minimizing the rate of hospital readmission and having a shorter duration of patient stay by predicting them to have optimal hospital resource utilization. The state of art and methodologies have been studied in this section to learn from the existing solutions and build an optimal framework for the prediction system.

Medical datasets are mostly complex and it's challenging working on them, as it brings opportunities to explore and identification of patterns helping in making predictions using patients records. These are discussed in different sections: Section 2.1 gives a review of the algorithms and methods for predicting length of stay. Section 2.2 gives a review of the algorithms and methods for predicting hospital readmission. Section 2.3 presents the application of the stacked ensemble model and techniques applied to the prediction problems and Section 2.4 gives an overview and summary of identified gaps in the literature and concluding them.

2.1 Length of Stay Prediction Methods and Algorithm Review

LOS prediction is usually given by the duration of the length of the patient's stay at the hospital, that is, the time when the patient is admitted and then discharged from the hospital. Decision tree methods: Adaboost, Bagging, and Random Forest using two tests – bisection and periodic tests for assessing the performance of the prediction methods (Ma et al.; 2020). The dataset used for the study was retrieved from one of the largest hospitals in Xiamen, having 11,206 records with the admission details of respiratory disease patients between the years 2014 and 2016. Insurance claim data was used in some studies to predict the length of stay, but there was very little research using hospital records for predicting LOS. The results obtained were also good, bagging outperforming other decision tree algorithms in the bisection test with RMSE of 0.296 and 0.831 R2, accuracy 0.723 which is the better result from other methods using insurance data. Overall, all the methods were sufficient to predict LOS but Bagging and Adaboost were better in bisection and periodic tests respectively.

The novel approach for predicting LOS is not always obtained from algorithms or methodologies, some researchers provided good results by using the novel R library tidy models for machine learning that enabled successful implementation of the different algorithms for the patients with respiratory diseases using the MIMIC – III dataset (Williams Batista and Sanchez-Arias; 2020). Categorization of LOS: A(LOS less than or equal to 3 days), B (LOS between 3 and 5 days), and C (LOS more than 5 days) helped in classification and it was observed that the C5.0 classification model, Random Forest, Support Vector Machine performed better in the case of ICU LOS prediction, but the results vary based on the dataset selected for the research. In another similar study, researchers compared regression algorithms – Elastic Net, Lasso, Linear, and Ridge for building a system for predicting LOS using the MIMIC – II dataset of ICU inpatients. Linear regressor was not able to perform well as usually, the medical dataset does not have a linear relationship between the features, rather Lasso, Elastic Net and Ridge performed satisfactorily. As all the three regressors were performing equally well, so the average was taken of them to predict LOS. Additionally, it was concluded that the reason for the patients staying for a longer duration is the number of the tests being performed, the patient having a greater number of tests tends to stay longer, but the reason for the more number tests was not discussed.

As the medical dataset has features that have non-linear features relation, so the machine learning algorithm considering the linear relationship does not work well with such datasets. So,

a non-linear feature selection method based on the Artificial Neural Network was implemented in one of the research papers (Kabir and Farrokhvar; 2019). LOS was binary classified as short or long-term stay, a preliminary step to using the model to identify the most significant features. SVM, LR and ANN were then used to fit on various subsets which were chosen for LOS classification. The results concluded that the ANN performed better than the SVM and LR, further confirming the relation between the LOS and its predictor as non-linear which can be more accurately classified by using non-linear classifiers such as ANN.

The results verified that better predictive results were obtained as compared to the individual or base learners' performance, and in addition to that it was observed that if the least performing model is removed then it degrades the performance of the stacked model. So, every model has its importance and supports in building a system. Area Under Curve (AUC) is chosen as the evaluation metric because of its wide usage in biostatistics. Zolbanin et al. (2020) employed the approach where patients historic data was considered to predict the LOS with COPD and pneumonia, by engineering new variables using the existing features helped in building an accurate model, this highlights the fact that the medical data is interrelated, as an example, the number of prior visits to the hospital can be potentially useful in determining the LOS of the patient for the coming visit.

2.2 Hospital Readmission Prediction Methods and Algorithms Review

Hospital readmission just after getting discharge threatens hospital management as it affects inpatient care. To avoid that, various methodologies were proposed, but most of the algorithms applied to a particular domain lacking generalizability. Thus, Im et al. (2020) introduced the idea of a discriminative-pattern based feature to classify the readmission of the patient. The unique pattern is identified by observing the common features in readmitted patients. Then using these features a prediction model is built and validated on three different datasets. The results showed that the predictions were better in terms of AUC, which was increased by 12% and in medical even the slightest improvement is significant. There was another study that used the MIMIC – III dataset to convert into binary features using the International Classification of Diseases (ICD) codes which helped in the diagnosis of disease in the patient (Assaf and Jayousi; 2020). These features were then used to train different classifiers - LR, SVM, MLP, and RF and among them RF performed best achieving 65% of accuracy and 0.66 of AUC. But researchers believed that the neural network such as Gated Recurrent Unit and LSTM would learn time-series patterns quickly, as MIMIC data is rich in temporal data which was not covered in the current research.

Most of the algorithm takes long computation time for processing the data, so the group of researchers build a system to predict hospital readmission wherein first stage the significant factors are identified and fed to the different machine learning algorithms - Decision Trees classifiers, k Nearest Neighbours, Random Forest method, deep neural networks and Support Vector Machine giving 35.7%, 53%, 35%, 50%, and 11.6% accuracy respectively (Al-Rubaei and Alhanjouri; 2020). In the second stage, to reduce the computation time a spark cluster cloud-based framework was built which enabled parallel working of algorithm thus decreasing the training time of the methods significantly and was able to reduce up to 32% of time when ML algorithms used Apache Spark. Apart from traditional models, Reddy et al. (2020) was inspired to build a prediction system using a Deep Belief Network (DBFN) for the risk of readmission. DBFN is compared to other conventional models like LR, RF, DT, Adaboost and Gradient boost using pre-processed diabetes dataset, the results proved that DBFN thrived on all the performance metrics – NPV, accuracy, precision, specificity except F1 score which was highest for Gradient boosting. Thus, DBFN can be used in building future prediction systems for other datasets as well.

Feature selection is found out to be a major part of most of the research based on the prediction of readmission of the patient and it becomes very important than choosing the ML algorithm which is easy to interpret. Alturki et al. (2019) focused majorly on extracting important features and did not include deep learning algorithms as a part of research due to the lack of interpretability of these algorithms, although it is acknowledged that deep learning has a good potential to enhance model accuracy. Recursive feature elimination wrapper method was used where attributes were ranked based on importance and then selected based on it. Small training sets are used to fit LR, RF, SVM, XGBoost, and k-NN on the 15 selected feature sets. Among all the models, RF performed best with an accuracy of 94.8% and SVM had the best AUC of 0.97. It was also concluded that the ensemble model such as RF and XGBoost perform well for this type of problem. On the same dataset, another study was done by Pujianto et al. (2019) which compared the C4.5 Decision Tree and Naïve Bayes for classifying the readmission of diabetic patients who had taken the HbA1c test. Stepwise feature selection wrapper method based on greedy approach is used and then Naïve Bayes is fit using 25 important attributes and C4.5 DT using only 9 features. The conclusion showed that the feature selection method combined with the sampling technique (SMOTE) gave better results where the C4.5 Dt model outperformed Naïve Bayes.

2.3 Review of Stacked Ensemble-Learning Model and Techniques for Prediction problem

Ensemble learning models are believed to advance the predictive ability of the weak or base learners and it is very difficult to choose one algorithm over the other for the classification problem as every algorithm has its strength and weakness. This inspired the researchers to build predictive models by using ensemble learning. Yu and Xie (2020) used a weight boosting algorithm combined with a stacked ensemble learning model to predict hospital readmission. Based on the Pearson's Chi-Square test of independence and correlation matrix some features are identified and feature selection methods such as ICD- Based clustering, wrapper method etc. are applied to get the final set of features to train weight boosting algorithm. The model showed an improvement of 22.7% accuracy and increased recall when compared to base models LR, DT, NB, SVM, ANN, and RF. A similar set of models were selected by El-Rashidy et al. (2020) for predicting the death of Intensive Care Unit (ICU) patients by using a stacked ensemble classifier. In the medical field, dealing with the dataset to predict some critical results such as mortality of ICU patients within the 24 hours of admission is challenging. Thus, it becomes important to have a prediction system and ensemble learning gave pretty good results with accuracy = 0.944, F1-score = 0.937, and AUC = 0.933. In future work, researchers emphasised that LSTM and CNN might give a better model with temporal and longitudinal data.

Often it is observed that the feature selection method combined with ensemble learning lead to great outcomes. Nguyen et al. (2019) did similar where used ensemble learning with voting for predicting breast cancer. Principal component analysis (PCA) and scaling were used to reduce the dimension of the dataset, as pre-processing makes it easier for the algorithm to process data. Various models are applied on the pre-processed data and voting is applied where each classifier's vote is given equal weightage and then the final prediction is made based on the majority. The evaluation revealed that ensemble learning has great potential for predicting breast cancer when compared with the base model. The machine learning algorithm can also be optimized using k fold cross-validation, by exposing the model to unseen data in different folds so that it does not overfit the sample data. This approach was used by the Sundaramurthy et al. (2020) where ensemble learning was used to predict and classify Rheumatoid Arthritis. The dataset is sampled using 10-fold cross-validation and used to recognize the better outcome of the model based on accuracy, precision, and ROC as metrics. SVM ensemble performed best with base classifier as RF and kNN giving a drastic improvement in the accuracy as compared

to bagging ensemble learning technique.

Another ensemble learning technique was developed by the Pham et al. (2019) for predicting hospital readmission for diabetic patients. The implemented model has two categories – the supervised and unsupervised ensemble model. Supervised ensemble model was built using five base classifiers – Chi-Squared Automatic Interaction Detection, NN with bagging, CHAID with boosting, CART with boosting networks, Augmented Tree with Naïve Bayes Network from the large pool of models. These models were finalized after many validations and then were used to develop an ensemble model. The implemented model was having an improved sensitivity of about 56% when compared to the existing models while having the same level of accuracy. An unsupervised model was built using k-means clustering and PCA was applied to reduce the dimension of the dataset so that it is convenient to form clusters. The results showed that patients data were classified into 4 clusters using Cubic Cluster Criterion (CCC) and one important was derived that the patients with a history of more inpatient visits have more chances of getting readmitted.

2.4 Summary of the Related work and final discussion

The most important conclusion drawn from reviewing all the literature is that the pre-processing and feature selection methods have helped in enhancing the performance of the models whether it is used for predicting LOS or hospital readmission. The results vary based on the quality and type of dataset used for building the prediction model, but data pre-processing and transformation plays a vital role. Misclassification or misidentification of LOS or hospital readmission can lead to serious mishaps or mismanagement of the hospital, so it is very important to focus on getting the least false negatives so that the purpose of this research is justified.

Table 1: Literature Review Summary

Author	**Method**	**Strength (Advantage)**	**Limitation**
Alahmar et al. (2019)	Length of stay prediction uing stacked Ensemble Model	Better result than the constituent learning algorithm	Usage of only supervised algorithm and LOS not predicted as a numeric value.
Ma et al. (2020)	Length of stay prediction using decision tree methods: Bagging, Adaboost, and Random forest	Bagging method on the testing set of the whole data set test (RMSE, 0.296; R2 = 0.831) on both metrics.	Less features used and generalization of model
Kabir and Farrokhvar (2019)	Length of stay prediction using artificial Neural Network	ANN as a non-linear classifier outperforms SVM and LR in LOS prediction	Complex architecture and less interpretability
Alturki et al. (2019)	Length of stay and hospital readmission prediction using feature Elimination wrapper method with algorithms - LR, RF, SVM, XGBoost, and k-NN	RF performed best with an accuracy of 94.8% and SVM had the best AUC of 0.97	Did not use Deep learning algorithm
Assaf and Jayousi (2020)	Hospital readmission prediction using ICD code with multiple classifiers RF, SVM, LR and MLP	Best model, Random Forest, achieved 0.65 accuracy and 0.66 Area Under the Curve (AUC)	Experimented only classical ML algorithm, LSTM and GRU can be explored
Yu and Xie (2020)	Hospital readmission prediction using weight boosted algorithm with ensemble learning	Improvement of 22.7% accuracy and increased recall when compared to base models	Explore Bayesian networks to further refine the classification model and Markov decision process (MDP) to determine optimal timing of certain interventions such as follow-ups
Nguyen et al. (2019)	Breast cancer prediction using ensemble learning with voting	Precision, recall, ROC-AUC, F1-measure, and computation time is best for the proposed model	Multi-layer perceptron can enhance the overall accuracy of the ensemble model

There are some gaps identified in LR because of the imbalanced nature of the dataset where the patients having readmission history are far less than the ones not getting readmitted and the most critical problem is the high dimensionality of the data which makes data pre-processing difficult. SMOTE or other sampling techniques needs to be applied to the dataset to avoid biased results and meet the research objective. Most of the research papers were either focused on predicting hospital readmission or LOS and the popular ML models used for building predictive models were – RF, SVM, DT, and deep neural network. Stacked ensemble learning models have promising results in the healthcare domain with complex datasets, as this excludes the possibility of overfitting and biased results as a combination of single or base classifiers cancel each other weakness and gives overall a strong predictive model, thus ensemble learning model is used to address the research problem. Table 1 explains the significant research papers with their results and limitations.

3 Methodology

The research objective is to build a model for predicting LOS and risk of hospital readmission for the patients. Having prior knowledge of the time it might take in the recovery of the patient and if there is a risk of readmission helps hospitals in organizing and managing their resources optimally and cut irrelevant costs. The model built while meeting the research objective helps in improving maintaining staff requirements in case of medical crisis and efficient operation of hospital facilities. Knowledge Discovery in Databases (KDD) is the most suitable methodology for this research as the most important step of this work is feature selection and it aligns best with KDD steps (Beniwal and Arora; 2012). KDD guided research to move along in a phase-wise manner - data selection, data pre-processing, data transformation (including feature engineering and selection), data mining, and model evaluation metrics. This is further explained in Section 3.1.

3.1 Methodology Steps

The steps followed while building the model as per KDD are shown in Figure 1: (i) Section 3.1.1 - Data extraction and selection from the dataset from 130 US hospitals as CSV files with metadata (ii) Section 3.1.2 - Data cleaning and removal of unnecessary columns and handling missing data (iii) Section 3.1.3 - Data transformation by using feature encoding, engineering and selection method, and handling outliers and imbalance by using sampling techniques (iv) Section 3.1.4 - Data Mining by performing exploratory data analysis and applying the proposed model to the dataset (v) Section 3.1.5 - Evaluation of the model by using performance metrics.

3.1.1 Data Selection

The data selected[2] for this research is obtained from the UCI ML data repository which is having diabetic patient records from 130 US hospitals from the year 1999 to 2008 and has about 50 features which are further described in Strack et al. (2014). The file obtained is in CSV format and some attributes store the data in the form of codes and a mapping file is also provided with the dataset. This dataset is considered for the research because of the diverse features and high dimension providing flexibility to explore data with many objectives and it has attributes storing information about the LOS and hospital readmission which is required to meet the objectives of this research. There are few challenges as well with this dataset, imbalanced data and missing values but this allows applying pre-processing and transformation techniques on the data to make it easy for the algorithms to process it.

[2] `https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008`

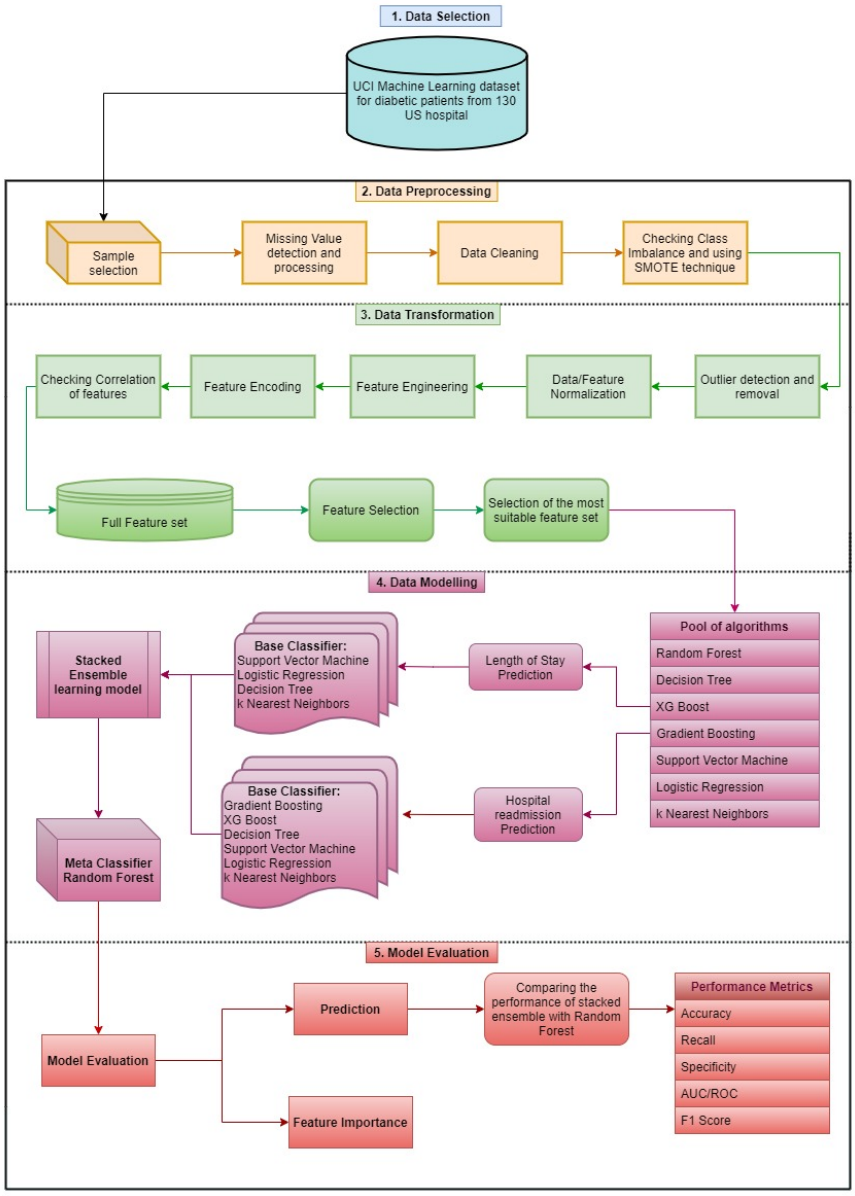

Figure 1: Methodology Process Flow

3.1.2 Data Pre-processing

The dataset is usually not gathered or built with a mindset of specific purpose or need, so there are high chances that the dataset might lack in quality or have missing values, so it becomes necessary to clean or process data before using it to train the model. Thus, pre-processing and transformation techniques become essential and usually take up to more than 50% of the time and effort in the building model.

As this is a medical dataset, there are some attribute names which does not make much sense like 'patient_id', 'discharge_id', 'glucose' renamed as the patient number, disposition discharge, glucose serum test respectively to make the dataset look more consistent. Few predictors such as 'examide' and 'citoglipton' (diabetes medication stores the values as Yes/No whether the medicine was prescribed or not to the patient) were having only 'No' value for every record, so it was better to not consider these features for building model. Also, the attribute 'discharge disposition' have the values in the form codes and a few of the codes denote the death of a patient in the medical facility, home, or hospice so these records were excluded from the dataset

for predicting hospital readmission it seems irrelevant. Also, there are missing values in the dataset as shown in Figure 2, the features like 'weight', 'payer code', 'medical speciality' have around 97%, 43% and 53% of missing data respectively. The fields 'weight' and 'payer code' does not hold much significance thus they are removed and 'medical speciality', 'race' has very few missing values so they are replaced as 'Unknown'.

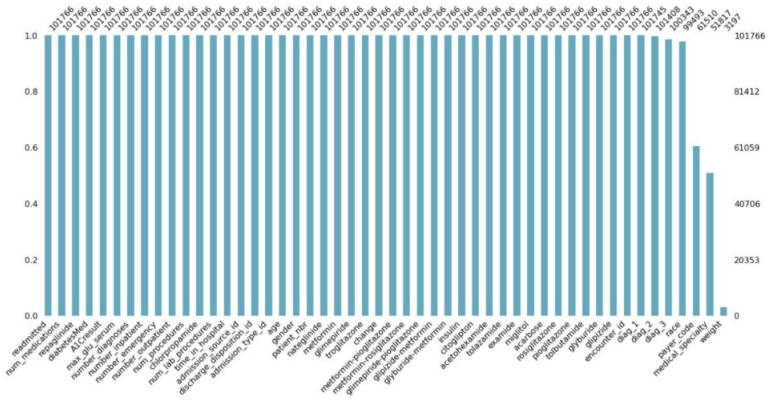

Figure 2: Missing Values in dataset

3.1.3 Data Transformation

The objective of transforming data is to restructure it and enhance the quality to make it more consistent by applying methods such as feature encoding, feature selection and engineering, Sampling, normalization, and handling outliers to make it easier for the predictive model to process data.

The relationship between the features can be utilized to create a new feature that can potentially act as a significant predictor while building a predictive model. Using this dataset, the variables 'number_inpatient', 'number_outpatient', 'number_emergency' which gives the count of the inpatient, outpatient and emergency visits of the patient after the first encounter of diabetes is combined to engineer a new variable by summing the number of all the types of patient visits and named as 'service_utilization' which denotes the total number of visits of the patients telling the how well the services are utilized by the patient. Similarly, 23 different features are giving the information whether there was medication change for that patient, the sum of which gives a new feature 'medication_change_count' denoting the number of times the medication was changed for that patient.

The features are encoded such that similar categories are combined into one category thus simplifying the categorization of features. The variable 'Readmitted' have values like '<30' and '>30' is encoded as '1' and '0' denoting that there was readmission and no readmission respectively. The 23 features having the information on the medication change have values such as 'No' and 'Ch' which are coded as '0' and '1' denoting that there is no change and change in medication respectively. Similarly, the variables 'A1c laboratory test', 'Glucose serum test', 'Discharge disposition' etc. are also encoded.

Features are selected using the Random Forest classifier which gives the topmost influencing factors which are then used to train the base classifier. Few outliers are also detected using the local Outlier Factor method in the attributes like 'number of medications', 'time in hospital', 'number of lab procedures' which were decided to remove as it was affecting the results of prediction. For handling the imbalance of the data as there are only 11% of patients who were readmitted and 89% were not readmitted shown in Figure 3, Synthetic Minority Oversampling Technique (SMOTE) technique was used to duplicate the minority class records so that there are then an equal number of readmission and no readmission cases. Feature scaling is done

using the z-score normalization method such that after standardization all the variables have a similar influence on the model.

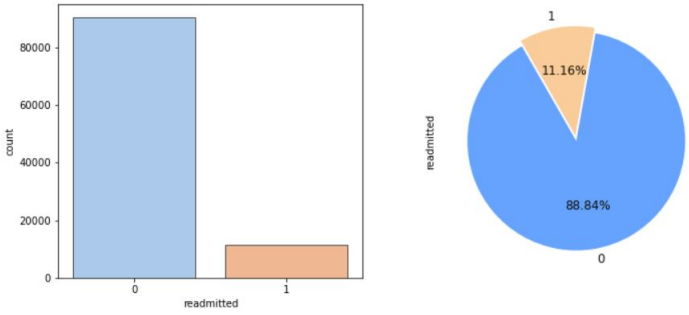

Figure 3: Imbalance class of readmission in data

3.1.4 Data Modelling

In the past, many different predictive models were built for predicting hospital readmission and length of stay. From the literature, it is observed that the stacked ensemble learning with the right selection of base classifier improves the performance of the model resulting in better prediction. So, for this project stacked ensemble model is built with a comprehensive approach of data cleaning and pre-processing which ensures the high performance of the model. This is a further explanation provided in Section 4 and Section 5 about the architecture and implementation of the proposed model.

3.1.5 Model Evaluation

The evaluation of the model predicting the readmission and length of stay of patients is evaluated using the metrics - Accuracy, recall, AUC, specificity and F1 Score. The base models are compared among each other and the stacking of the best performing models are done to build a stacked ensemble model. The final stacked ensemble model is then compared to existing models as well based on the results as discussed in Section 6.

4 Design Specification

The proposed model for this research is Stacked ensemble learning where using the subset of significant features a model is built for prediction LOS and hospital readmission of the patients. The framework explaining the step by step process of building the model is shown in Figure 1 and Figure 4.

1. For designing the model, the first step is extracting the data from the UCI machine learning data repository and processing the data using the data processing and transformation techniques which are also explained in the Sections 3.1.2 and Section 3.1.3 respectively.

2. The cleaned and transformed feature set is obtained from the previous step. As analysed from the Exploratory Data Analysis (EDA) that the features have high correlation so new features were engineered by combining existing features and then used for creating the feature set which is used for training models.

3. The base classifiers are then stacked together and are trained with 5-fold cross-validation.

4. The results obtained from the base models is then augmented to the dataset and the augmented dataset is used to train the meta classifier, Random Forest.

5. The performance of the stacked ensemble learning model is then compared to the Random Forest model based on the evaluation metrics.

4.1 Ensemble Learning Model

Most of the existing research shows that the ML algorithms or models were built for specific problems or diseases using one or two classifiers and some of them provided good results as well, but the same model fails to address a different problem or dataset. The results and performance of the algorithms vary, as no one algorithm or solution outperforms all other algorithms. Every algorithm has its strength and weakness and if a model is formed using multiple ML algorithms averaging out each other's weaknesses then a strong model can be obtained. Following this ideology, ensemble learning was proposed in this research, as it somehow imitates the human nature of getting different perspectives or views before taking any decision. It is thus believed that ensemble learning might provide a better result than single classifiers. Therefore, single classifiers selected for this study are different and diverse.

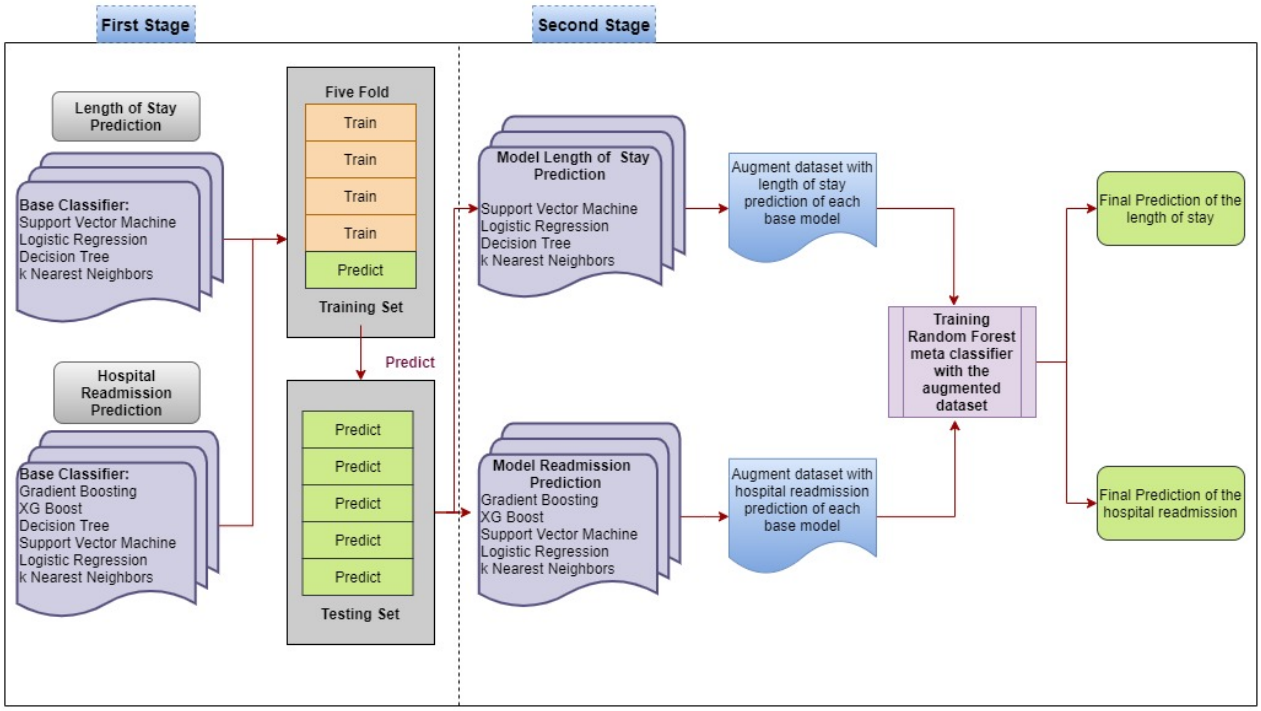

Figure 4: Stacked Ensemble Learning Model

4.2 Stacked Ensemble Model

The stack ensemble model is having two stages as shown in Figure 4, in the first stage single or base classifiers are trained using 5-fold cross-validation to avoid overfitting the dataset and then the prediction of the algorithm is augmented to the dataset. The augmented data with prediction is then used to train the meta learner RF in the second stage which is used to get the final prediction of the model. The prediction results from each of the base classifiers are compared to the Stacked ensemble, as the motive is to get better results when compared to

single classifiers so the possibility of getting a better outcome heavily depends on the selection of base classifiers.

The reason for the selection of RF is as one of the classifiers is that it's the most flexible algorithm and without much hyperparameter tuning also provide good results. In the past for similar studies, RF has proved to be a good model (Im et al.; 2020; Alturki et al.; 2019). Also, one of the objectives of the research is to find the most influential or significant features while predicting LOS and readmission, RF decides to split a node based on the importance of feature thus searching for the best feature among the subset of features, due to this it results in wide diversity and building generally a good model. The dataset selected for this research is having huge feature set and have noisy data and SVM is known to deal with such kind of problems, also it is less prone to over fitting which can help in overcoming this problem that might arise in case of RF and provided good results (Alturki et al.; 2019).

kNN is chosen to be a part of ensemble learning as it is rather a simple classifier and does not have any assumptions on data thus making it a good choice for the data with irregular boundaries. But it lacks in providing the important features as equal weightage is given to each feature in their contribution to the distance function. kNN lacking in providing the feature importance is covered by RF. Most of the algorithms are sensitive to noise, thus requires data cleaning, processing, and feature extraction. Logistic Regression (LR) is considered as a base model when there is a possibility to have a relationship between the feature sets and outcome in the form of log-likelihood. LR has always been a standard tool for binary classification and its wide usage and acceptance in the healthcare field where model interpretability is essential (Alturki et al.; 2019).

Extreme Gradient Boosting (XGBoost) is an algorithm that is aware of the sparse data and used a weighted quantile approach for forming a tree. The model implemented using gradient boosted decision trees are known for its speed and performance (Ma et al.; 2020). This algorithm gives good performance if the data is tabular or structured. As the data used in this research is tabular, this algorithm is apt.

5 Implementation

This section presents the software and technologies which were used to implement the project. The implementation of the project is done using the python language and the environment used for executing code in Jupyter Notebook on the Anaconda platform. Anaconda is preferred as it is easy to use and take care of the library management and other dependencies to run the code. Jupyter notebook runs on the web which is an open-source application that allows the creation and sharing of the executed code, visualization, and processing etc. For loading data from CSV files, pandas library is used and MatPlotLib, seaborn, and plotly libraries for Exploratory data analysis. Scikit-learn machine learning library is probably one of the most useful which includes efficient tools for ML and statistical modelling, all the data processing, model building and evaluation metrics is done using it.

5.1 Exploratory Data Analysis

The exploratory data analysis is done to have a better understanding of the data and metadata of the dataset. During the analysis, various conclusions are derived which further helps in refining and pre-processing the data. Figure 5 shows that the number of readmission cases increases with an increase in the lab procedures. The patients with a shorter duration of stay during their first time of visit have higher chances of getting readmitted as shown in Figure 6. According to Figure 7, The patients who are older with the age of 65 to 85 years are more prone to the risk of readmission (where '0' states the no readmission and '1' as readmission).

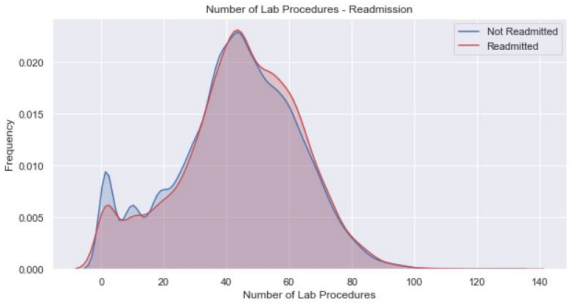

Figure 5: Number of lab procedures Vs Readmission

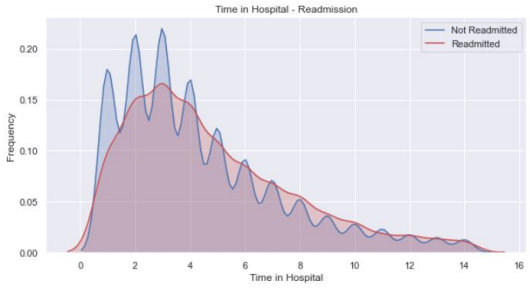

Figure 6: Time in Hospital Vs Readmission

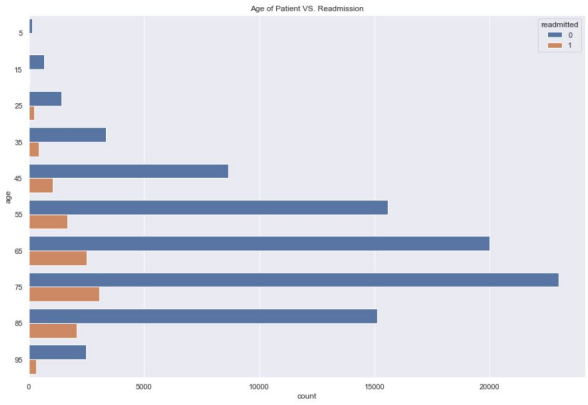

Figure 7: Age Vs Readmission

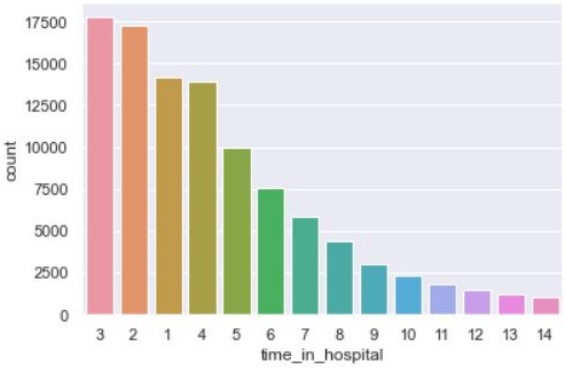

Figure 8: Time in Hospital

This data shows that the patients stays usually for around 1 to 3 days at the hospital given by Figure 8.

5.2 Implementation of Logistic Regression

The logistic regression model is implemented with a range of parameters, where the most important one is the penalty. The initial model built does not have a penalty, but then the second model is built using 'L1' penalty with the data split into training and testing datasets in an 80:20 ratio i.e., 80% training data and 20% testing data for predicting length of stay and hospital readmission.

5.3 Implementation of Support Vector Machines

A support vector machine is implemented using the Scikit-learn library and the kernel used for building the classifier is linear, polynomial, and sigmoid, out of which linear and polynomial performed better with the kappa optimised for the value of C of 1. Models were also built for polynomial and sigmoid with the value of C of 100, where the polynomial kernel showed an improved performance.

5.4 Implementation of k Nearest Neighbors

k Nearest Neighbors is implemented using Scikit-learn library where range of 1 to 15 values of 'k' were used to train the model and validated against the test data and plotted on a graph as Figure 9 and the best value was observed as '2', thus final model was built with this parameter.

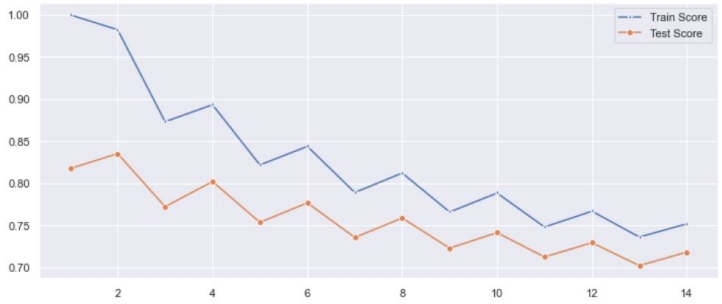

Figure 9: Plot for the value of 'k' Vs Accuracy for kNN model

5.5 Implementation of Decision Tree

The decision tree model was developed with some parameters like a criterion for splitting the tree is taken as 'entropy' as it will help in gaining information on important features on which splitting takes place, depth as '28', and minimum samples split as '10'. These parameters are selected after the 20 trials for best fitting the model.

5.6 Implementation of Gradient Boosting and Extreme Gradient Boosting:

Gradient Boosting and Extreme Gradient Boosting is implemented using the Scikit-learn library and the parameters selected for building gradient boost are random state as '42' and for building Extreme Gradient Boosted are random state as '42', max depth as '3' and to determine the best model by performing 100 iterations keeping the maximum depth to learn when the performance of model starts degrading. The final agreement was on the maximum accuracy with a depth of 3.

5.7 Implementation of Random Forest

The random forest was built using the Scikit-learn library where it was modelled with 500 trees and the optimised accuracy was obtained at the maximum sample split of 10 and maximum depth as 25, a split was based on the Gini and it was identified that the minimum estimator as 10 was best fit for the model.

5.8 Implementation of Stacked Ensemble Model

The stacked Ensemble model was built using the 'vecstack' package was used for stacking the models and then training the stack ensemble model with various parameters using the 'stacking' method. The major parameters for building this model were the mode as 'oof_pred_bag' which states that the training of the model is done in many folds, the optimal folds for training were decided as '5' and the metric selected was 'accuracy'. In 5-cross validation training, 4 folds are used for the training model and the 5th fold is used for the testing model. Thus, the predictions obtained from the base models are augmented in the training dataset, which is used to train the Random Forest meta classifier as explained in Section 4.2.

5.9 Result Discussion

For each model, the best combination of the hyperparameters was identified and then the models were scored based on the prediction results. All the evaluation metrics are chosen after

considering the algorithms and parameters applied. As observed in Table 2, overall Random Forest has shown the best results in terms of accuracy, AUC, Specificity and F1-Score with the respective values as 0.89, 0.88, 0.87, 0.89 but Decision tree and k Nearest Neighbors has also performed equivalently well. For hospital management, the biggest problem is to predict whether the patient has a risk of readmission and determine if the intervention can be put in place to reduce the variance for effective hospital operation. For this purpose, the specificity metric is most relevant with accuracy so that it can appropriately address the research problem.

Table 3 shows the comparison of models predicting length of stay whether short-term or long-term at the hospital. Random Forest performed better in this case as well with the accuracy of 0.89, recall of 0.90, AUC as 0.89, specificity as 0.88 and F1-score as 0.89, decision tree and decision tree and k Nearest Neighbors also did well.

Table 2: Hospital readmission prediction model results

Model	**Hyperparameter**	**Accuracy**	**Recall**	**AUC**	**Specificity**	**F1Score**
Logistic Regression	without hyperparameter	0.91	0.67	0.50	0.69	0.76
	L1' penalty, solver 'liblinear'	0.75	0.75	0.75	0.74	0.74
Support Vector Machine	without hyperparameter	0.76	0.78	0.76	0.75	0.75
	Polynomial kernel and C=1.0	0.75	0.77	0.75	0.72	0.76
	Polynomial kernel and C=100.0	0.78	0.82	0.75	0.74	0.79
k Nearest Neighbor	k=2	0.83	0.91	0.84	0.76	0.85
Decision Tree	criterion='entropy', max_depth=28, min_samples_split=10	0.89	0.86	0.82	0.78	0.82
Random Forest	max_depth=25, min_samples_split=10, n_estimators=10	0.89	0.89	0.88	0.87	0.89
Gradient Boosting	random_state=42	0.62	0.59	0.66	0.66	0.61
Extreme Gradient Boosting	random_state=42, n_jobs=-1,max_depth=3	0.70	0.60	0.62	0.67	0.62

Table 3: Length of stay prediction model results

Model	**Hyperparameter**	**Accuracy**	**Recall**	**AUC**	**Specificity**	**F1Score**
Logistic Regression	without hyperparameter	0.88	0.67	0.59	0.68	0.70
	L1' penalty, solver 'liblinear'	0.81	0.80	0.81	0.80	0.81
Support Vector Machine	without hyperparameter	0.82	0.83	0.80	0.71	0.82
	Polynomial kernel and C=1.0	0.81	0.82	0.81	0.81	0.80
	Polynomial kernel and C=100.0	0.83	0.86	0.83	0.80	0.75
k Nearest Neighbor	k=2	0.85	0.89	0.85	0.81	0.86
Decision Tree	criterion='entropy', max_depth=28, min_samples_split=10	0.89	0.82	0.81	0.78	0.82
Random Forest	max_depth=25, min_samples_split=10, n_estimators=10	0.89	0.90	0.89	0.88	0.89

Figure 10 and Figure 11 shows the results of applying a stacked ensemble model for predicting hospital readmission and length of stay respectively. The stacked ensemble performance results in terms of accuracy, ROC, specificity, F1 score outperforms each of the base models. However, there is a chance that the stacked model does not perform better than the single model. Compared to Random Forest, the stacked ensemble was a little better, but in the medical research field, even the slightest improvement in the performance is relevant. In addition, utilizing the high computational resource and parallel architecture with 5-cross validation folds helped in optimizing the algorithm and provide good results in the form of the stacked ensemble technique. For hospital readmission prediction, the stacked ensemble gave an accuracy of 0.94 which is the highest among all the base models and specificity of 0.99 which is best among all

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	STacked Classifier2	0.943973	0.989357	0.898059	0.99026	0.9415	0.94416	1.935111	0.89182

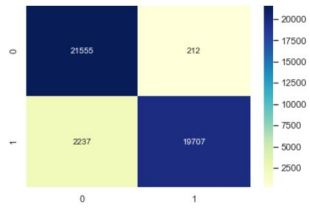

Figure 10: Hospital Readmission prediction

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	STacked Classifier2	0.901444	0.900973	0.902794	0.900084	0.901883	0.901439	3.404057	0.802885

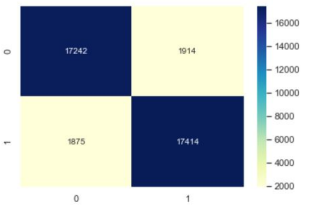

Figure 11: Hospital Stay prediction

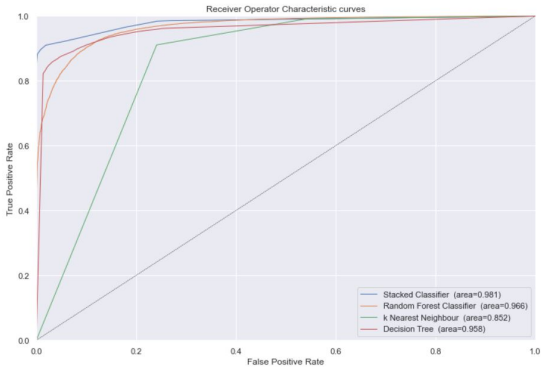

Figure 12: Hospital Readmission prediction ROC

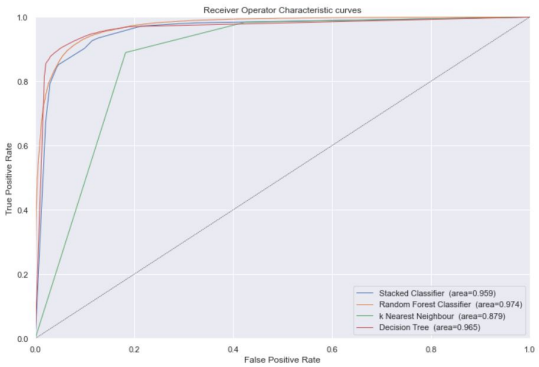

Figure 13: Hospital Stay prediction ROC

models. Similarly, for predicting length of stay at the hospital, the stacked ensemble gave an accuracy of 0.90 and specificity of 0.90 which is slightly better than Random Forest.

6 Evaluation

The predictive models and approach proposed in this research are implemented successfully. The classifiers – Random Forest (RF), k Nearest Neighbors (kNN), Decision Tree (DT), Logistic Regression (LR), Support vector machines (SVM), Gradient boosting (GB) and extreme gradient boosting (XGB) were built. Stacked Ensemble model was built for predicting hospital readmission using the base classifiers as DT, SVM, LR, kNN, GB, XGB and it was compared with the RF, as it performed best among the other classifiers. Similarly, a stacked ensemble model was built for predicting length of stay using the base classifiers as DT, SVM, LR, kNN and compared with RF. The stacked ensemble in each case outperformed the base classifiers, proving that if the base models are selected carefully then combining those models can give better results in the form of a stacked ensemble. Table 4 shows the comparison of the final stacked ensemble model implemented in this research with other existing models. The stacked ensemble with the diabetes data in this research has given better results as compared to the existing models. The hyperparameters and choice of base classifiers were the most important factors in the success of the model. The implemented stacked ensemble has the best AUC and specificity among all the existing models which are 0.99 in case of readmission and 0.90 in case of Length of stay prediction, this states that this is not only accurate, but it identifies the true negatives very well and has good AUC for false positives which is essential in the medical research. It can be observed from the Figure 12 and Figure 13 that the Stacked ensemble in both the cases has the highest area covered in Receiver Operator Characteristics (ROC) curve which is 0.98 and 0.96 respectively.

There are two main reasons for the increase in the performance of the Stacked ensemble

model:

1. The dataset for this research was quite complex, had noisy and missing data. The methodological approach used for building the predictive model included comprehensive data pre-processing and transformation technique which acted as a key component for the high performance of the model.

2. It is very important to select the base models which are diverse and complement each other to build a strong predictive model. Based on this approach, the diverse single classifier or base models which performed well when trained using the processed data were selected to build a stacked ensemble model. The models were also tuned to get their better version which further enhanced the results.

The pre-processing was done to enhance the performance of base models so that when the top-performing models from the pool of base models are selected then there is a higher chance of getting a strong predictive model.

Table 4: Comparison of models with existing work

Research	Method	Accuracy	Recall	Precision	Sensitivity	Specificity	F1Score	AUC/ROC
Proposed Model	**Stacked Ensemble Model for length of stay prediction**	**0.9**		**0.9**	**0.9**	**0.9**	**0.9**	**0.9**
	Stacked Ensemble Model for hospital readmission prediction	**0.94**		**0.98**	**0.89**	**0.99**	**0.94**	**0.94**
Alturki et al. (2019)	RF, LR, XG Boost, SVM, kNN for hospital readmission prediction (Best Model RF with evaluation metrics)	0.94	0.94	0.95			0.94	
	For Length of Stay prediction (Best model SVM with evaluation metrics)	0.87	0.87	0.88			0.87	
Assaf and Jayousi (2020)	Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Multi-Layer Perceptron (MLP) for predicting hospital readmission (Best model - Random Forest)	0.65						0.66
Alahmar et al. (2019)	Stacked Ensemble model for predicting length of stay for diabetic patients (Deep Learning , Distributed Random Forest, Generalized Linear Model, Gradient Boosting Machine, and Naïve Bayes Classifier)							0.81
Yu and Xie (2020)	Joint Ensemble learning model the modified weight boosting algorithm with stacking algorithm for predicting hospital readmission	0.88			0.89	0.87		0.88

In addition, another aspect of evaluation is to meet the research objective of identifying the most influential features for predicting the hospital readmission and length of stay prediction. Random Forest is popular for providing the most important features affecting the model, thus after applying the hyperparameters to the RF to get an optimized model. The important features are extracted which are given in Figure 14 and Figure 15. The most significant feature for predicting hospital readmission is 'time_in_hospital' which is the length of stay, 'number_diagnoses', 'num_procedures', 'age', 'num_medications'. This shows that if the patient stays for a longer duration in hospital, have a high number of diagnoses, lab procedures, medications prescribed and are older then the probability of readmission in the future increases. On the other hand, the most significant features for predicting length of stay are 'num_medications', 'num_procedures', 'number_diagnoses' which states that if the patient has a high number of medications prescribed, number of lab procedures, and number of diagnoses test then person tend to stay for longer duration in hospitals.

7 Conclusion and Future Work

To cope with the challenges of predicting the risk of readmission and length of stay for diabetic patients, this research with detailed pre-processing and transformation techniques improved

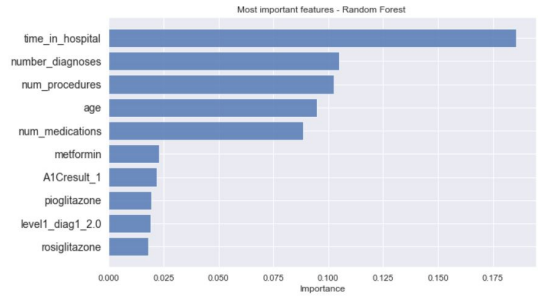

Figure 14: Important Features for readmission prediction

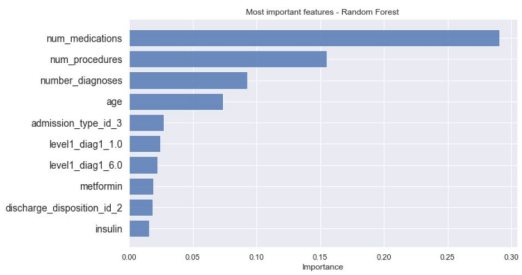

Figure 15: Important Features for LOS prediction

data quality to produce a highly effective and efficient model. The methodology of the research includes the methods like data cleaning, feature selection and detecting feature importance, feature encoding, feature engineering, sampling using SMOTE, outlier detection, and normalization to extract the optimal subset features to train the models.

The proposed Stacked Ensemble learning model is built using the different and diverse base classifiers selected based on the performance, which is also tuned using the hyperparameters which outperform other machine learning algorithms. The model is found to have a good performance for all the evaluation metrics – Accuracy, recall, specificity, F1Score, and AUC. Diabetes patient's readmission and length of stay prediction is a critical subject for health care service providers and to maintain the quality of patient care. The data pre-processing and transformation techniques acted as the key component for the success of the model, as this research is more comprehensive methodological oriented, so these methods ensured the and accurate predictions. The top-performing models are selected from the pool of classifiers based on the tuning results and then stacked together to get better results compared to any other base classifier. The stacked ensemble model not only outperformed the base classifiers but also the existing models by attaining the accuracy, precision, specificity, F1Score, AUC of about 90% each in case of length of stay and 94%, 90%, 95%, 90% and 98% respectively in case of readmission prediction.

The meta classifier of the stacked ensemble model is Random Forest (RF), but when the RF is trained using the augmented data with stacked prediction results of the base classifier then it performed better than the single RF classifier without stacking. So, stacking can improve the result of the model significantly if trained by the right set of base models. In addition to that, some important features like 'time_in_hospital' which is the length of stay, 'number_diagnoses', 'num_procedures', 'age', 'num_medications' were identified by this experiment which influences the readmission rate and features like 'num_medications', 'num_procedures', 'number_diagnoses' affects hospital stay duration of patients at the hospital.

There is a scope of improvement in this research where it can be optimized by considering the deep learning (DL) algorithms as a part of the experiment, it might enhance the stacked ensemble learning. The problem with deep learning is its lack of interpretability, but neural networks can be trained without much data cleaning and pre-processing which makes it useful as medical datasets are usually complex and have noisy data which makes it difficult to train machine learning algorithms. DL can also be utilized for identifying the important features even without data cleaning and processing, which might help researchers to focus on the important features and their pre-processing instead of finding the significant features in the later stage of research and losing time and effort on irrelevant attributes. In this research, the length of stay is having binary classification – long or short term stay, in future the models can be built to regress the numeric value of days for the duration of stay which could be more useful for the hospital management.

References

Al-Rubaei, F. and Alhanjouri, M. (2020). Generalization of deep neural network of hospital readmission prediction models for diabetes patients using apache spark clustering, *2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech)*, pp. 120–125.

Alahmar, A., Mohammed, E. and Benlamri, R. (2019). Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes, *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, pp. 38–43.

Alturki, L., Aloraini, K., Aldughayshim, A. and Albahli, S. (2019). Predictors of readmissions and length of stay for diabetes related patients, *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–8.

Alves, T., Laender, A., Veloso, A. and Ziviani, N. (2018). Dynamic prediction of icu mortality risk using domain adaptation, *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 1328–1336.

Assaf, R. and Jayousi, R. (2020). 30-day hospital readmission prediction using mimic data, *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–6.

Beniwal, S. and Arora, J. (2012). Classification and feature selection techniques in data mining, *International Journal of Engineering Research and Technology* **1**.

Bhuvan, M. S., Kumar, A., Zafar, A. and Kishore, V. (2016). Identifying diabetic patients with high risk of readmission.

Comino, E. J., Harris, M. F., Islam, M. F., Tran, D. T., Jalaludin, B., Jorm, L., Flack, J. and Haas, M. (2015). Impact of diabetes on hospital admission and length of stay among a general population aged 45 year or more: a record linkage study, *BMC health services research* **15**(1): 1–13.

Cui, S., Wang, D., Wang, Y., Yu, P.-W. and Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction, *Computer methods and programs in biomedicine* **166**: 123–135.

El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S. and El-Bakry, H. M. (2020). Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model, *IEEE Access* **8**: 133541–133564.

Im, S. J., Xu, Y., Watson, J., Bonner, A., Healy, H. and Hoy, W. (2020). Hospital readmission prediction using discriminative patterns, *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 50–57.

Kabir, S. and Farrokhvar, L. (2019). Non-linear feature selection for prediction of hospital length of stay, *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 945–950.

Ma, F., Yu, L., Ye, L., Yao, D. D. and Zhuang, W. (2020). Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods, *IEEE Journal of Biomedical and Health Informatics* **24**(9): 2651–2662.

Nguyen, Q. H., Do, T. T. T., Wang, Y., Heng, S. S., Chen, K., Max Ang, W. H., Philip, C. E., Singh, M., Pham, H. N., Nguyen, B. P. and Chua, M. C. H. (2019). Breast cancer prediction using feature selection and ensemble voting, *2019 International Conference on System Science and Engineering (ICSSE)*, pp. 250–254.

Pham, H. N., Chatterjee, A., Narasimhan, B., Lee, C. W., Jha, D. K., Fai Wong, E. Y., Ellyanti, S., Nguyen, Q. H., Nguyen, B. P. and H. Chua, M. C. (2019). Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis, *2019 International Conference on System Science and Engineering (ICSSE)*, pp. 273–278.

Pujianto, U., Luki, A., Ar Rosyid, H. and Salah, A. (2019). Comparison of naïve bayes algorithm and decision tree c4.5 for hospital readmission diabetes patients using hba1c measurement, *Knowledge Engineering and Data Science* **2**: 58.

Reddy, S. S., Sethi, N. and Rajender, R. (2020). Evaluation of deep belief network to predict hospital readmission of diabetic patients, *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 5–9.

Rosen, O. Z., Fridman, R., Rosen, B. T., Shane, R. and Pevnick, J. M. (2017). Medication adherence as a predictor of 30-day hospital readmissions, *Patient preference and adherence* **11**: 801.

Sharma, A., Muir, R., Johnston, R., Carter, E., Bowden, G. and Wilson-MacDonald, J. (2013). Diabetes is predictive of longer hospital stay and increased rate of clavien complications in spinal surgery in the uk, *The Annals of The Royal College of Surgeons of England* **95**(4): 275–279.

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* **2014**.

Sundaramurthy, S., C, S. and Kshirsagar, P. (2020). Prediction and classification of rheumatoid arthritis using ensemble machine learning approaches, *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 17–21.

Williams Batista, R. and Sanchez-Arias, R. (2020). A methodology for estimating hospital intensive care unit length of stay using novel machine learning tools, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 827–832.

Yu, K. and Xie, X. (2020). Predicting hospital readmission: A joint ensemble-learning model, *IEEE Journal of Biomedical and Health Informatics* **24**(2): 447–456.

Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M. and Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining, *Expert Systems with Applications* **42**(20): 7110–7120.

Zolbanin, H. M., Davazdahemami, B., Delen, D. and Zadeh, A. H. (2020). Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases, *Information & Management* p. 103282.
URL: *https://www.sciencedirect.com/science/article/pii/S0378720619301594*