Use of Machine learning models to find the relationship between hospital management system for hyperglycemia and readmission rate

Eesha tur razia babar, Student ID: 45369117

June 29, 2022

Abstract

The management of hyperglycemia of patients during their visit to hospitals is essential for better treatment of disease, to decrease mortality and the hospital readmission rate. Unfortunately, very few studies have been conducted for assessment of diabetes care in the hospitals until now. In this project, a data set collected from 130-US hospitals, recorded between the years 1999-2008 was used to determine the relationship between the treatment provided to patients with blood glucose problem and their readimission rate to hospital with the same problem. Multiple advanced machine learning models were employed for this purpose and a final accuracy of 75 percent was obtained using multilayer perceptron (MLP) through Machine learning pipeline, Gridsearch and SMOTE. This relationship will help to implement proper protocols for diabetes treatment for better outcomes in terms of both mortality and morbidity, in the hospitals.

1 Introduction

According to CDC USA, 11.3% of the US population is suffering from diabetes while 38.0% of the adult US population is suffering from pre-diabetes [1]. As of 2021, diabetes is considered as 8th leading cause of death in USA [2]. Management of diabetes related symptoms has very important relation with probable life expectancy of patients. According to Deborah et.al, hyperglycemia can be controlled through proper management of inpatient treatment [3]. Diabetes management system in the hospitals has positive outcomes in terms of prevention of disease and control [4]. If the relation between inpatient care and possible outcomes can be fully understood, targeted protocols can be set for diabetic patients for better outcome. As more and more hospitals are employing the customized protocols for diabetic patients, there is still a lot of room for improvement since there are various challenges in terms of studying the clinical data and to find out the relationship between diabetes management process in the hospital with the outcome. The actual clinical data generally have missing values/outliers and is difficult to deal with as compared to data obtained during an experiment in a controlled environment because of possible noise, redundancy and sparsity of labels. To deal with this issue Beata Strack et.al [5] decided to use the method based on HbA1c measurement for the patients of hyperglycemia and applied Machine learning based methods to find the relationship between these measurements and the hospital readmission rate of patient for same problem. The motivation behind this project is research conducted by Beata et.al to find out the relationship between hospital management system for hyperglycemia and the resulting impact on readmission rate of patients in the hospital for same disease. A multivariate data set which was obtained from 130 US hospitals during the year 1998-2008 was used in this project [6]. There are three different class labels in this data. These three classes are No readmission, readmission within 30 days of discharge, and readmission after 30 days of discharge from hospital. This data set includes multiple procedures that were conducted on the patients, various vital sign which were measured and several medications which were given to the patients. Then the class label of patient is also recorded in the data set in order to find out the relationship of diabetes management system with the readmission rate of patient in the hospitals. In this project multiple machine learning models were used, from simpler decision trees to advanced neural networks to increase the accuracy of class prediction of patient. Multiple advanced techniques including machine learning pipelines and SMOTE were used to increase accuracy and deal with class imbalance. Regularization and hyper-parameter tuning was employed for better results. Finally multi-fold cross-validation techniques were employed to validate the results.

2 Problem Statement

The purpose of this project is to predict the hospital readmission rate of patients having hyperglycemia after being treated for the same disease in the hospital. This hospital readmission rate of patient will help to understand the relation between diabetes management system in the hospitals with the possible outcomes in terms of mortality, morbidity and cure of disease. This understanding will help to use targeted and customized protocols for diabetes management in the hospitals.

3 Data Exploration

In order to get better understanding of data, it was initially explored through manual reading of documents and then using statistical analysis tools. The data was explored to find out the missing and/or invalid values and to get some numerical statistics. After finding theoretical information and numerical statistics, multiple graphs were plotted in order to visualize feature distribution and important features for class prediction.

This data set consists of 101766 instances, which means the information regarding 101766 patients during their stay in the hospital. The original data set had 55 features which includes the number of days a patient lived in the hospital anywhere between 1 to 14, demographic information of patients for example age, gender and race. It also includes a list of 24 different medications, stating whether or not patient was taking that medication. It also includes information about the lab test performed for a patient. It have the information regarding the insulin level of patient and the information regarding whether or not a patient

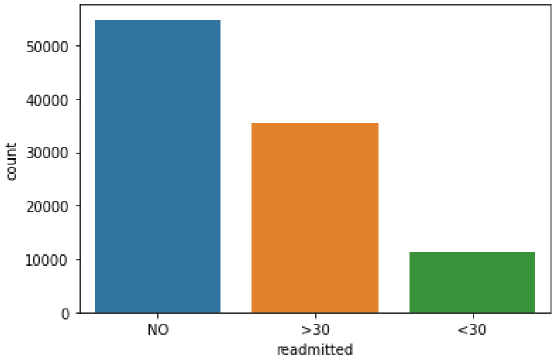

Figure 1: Data distribution across different class labels.

had diabetes and whether or not the same patient was readmitted to the hospital with same problem within 30 days of discharge. In depth detail about data set can be obtained from the paper of Beata Strack et.al [5] After getting theoretical information of data, it was analyzed statistically to find out number of features, unique class labels, type of values etc.

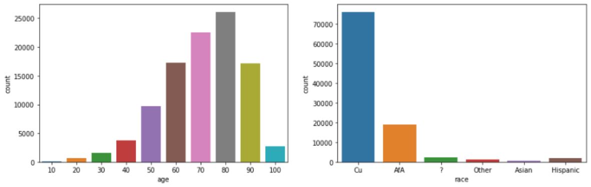

Figure 2: Demographics based distribution of patients.

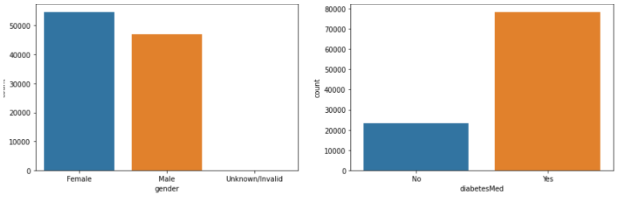

Figure 3: Demographics and medication based distribution of patients.

These statistics helped us to find out the number of unique classes (no readmission, readmission within 30 days and readmission after 30 days), types of features and missing values donated by ?. Figure 1 shows the histogram of class label distribution across patient counts using raw data directly from CSV files. Figure1 helps to realize the class imbalance, showing that a very small number of patients were readmitted within 30 days of discharge and this is one of the challenging aspect of this data set. In the figure2, mapped data set was used instead of raw data set for better visibility for example Cu in the race column means Caucasian and AFA means African Americans.

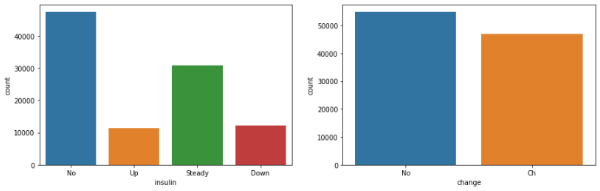

Figure 4: Insulin level and change in blood glucose level based distribution of patients.

Table 1: Statistical analysis of data

Item	Value
Number of instances	101766
Number of attributes	50
Number of features	49
Number of class labels	3
Attribute types	integer(12), object(38)
Missing values	Yes

4 Data Processing

For data processing, initially it was explored for missing values. It was realized that weight column had a lot of missing values. It was not possible to deal with this problem by replacing missing weights with mean or median weight because the number of entries with valid weight values were very small as compared to those having missing weight values. That is why it was decided to drop weight column. Also, medical specialities column had a lot of missing values too hence it was decided to drop that feature too. All the other features were mapped to their numerical equivalents using the information given in the documents for mapping. The class labels were also given numerical representation. 0 being no admission, 1 being admission within 30 days and 2 being admission after 30 days. All the other tests for example maxgluserum, A1Cresult, metformin were distributed into classes. Some features related to registration forexample encounter_id, patient_nbr, payer_code had no relation with

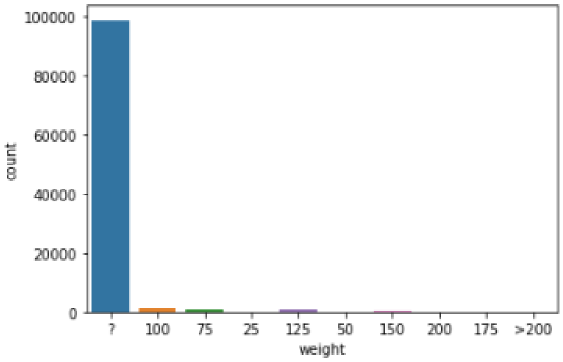

Figure 5: Weight column with missing values represented as ? .

class prediction hence these were dropped. It was also realized that there was no direct relation of race with the class label hence race column was also dropped.

5 Model Exploration for multi-class classification

After data exploration, visualization and processing, data frame was converted into numpy array in order to implement machine learning models on it.

5.1 Phase 1

After data exploration and visualization, it was realized that the linear classification model will be very simple for given problem statement since the problem statement falls into the category of multi-class classification. That is why it was decided not to use linear classifier but to start with decision trees. For the first run, mean squared error was used as performance metric and the models from open source library UCIML Library were used. Initially, the data was divided into train and validation sets of 75% and 25%. Firstly, decision tree classifier was implemented using variation in maximum depth. The minimum error which was obtained using decision tree classifier by varying maximum depth for multi-class classification was 0.45 using maximum depth of 5, on validation data set. The minimum error by varying maximum depth and minimum parents was obtained as 0.4385 using maximum depth as 5 and minimum parents as 2. Then a random forest classifier with the 10bags was implemented using maximum depth of 45 and minimum parents as 25 and minimum error was obtained as 0.4179. On the same data set KNN classifier was used, with varying K (neighbour of nearest neighbours) and A (regularization). Two dimensional graph was plotted for training and validation error. The behaviour of graph depicts the general behaviour of KNN where error on training set increases by increasing the values of K and error on validation set decreases by Increasing the value of K. The minimum error using KNN was obtained using K = 100 and A = 1 as 0.475.

Table 2: Minimum mean squared error of models

Model	MSE
K Nearest-Neighbor	0.475
Decision Trees	0.4385
Neural Networks	0.434
Random Forest	0.4179

After the KNN and Decision Trees, the Neural Network model was employed. Multiple hyper-parameters were also tested for this model including several hidden layers, several different numbers of nodes per layer, and activation functions including but not limited to htangent and logistics. The minimum error obtained using Neural Network was 0.434. After first run the performance of all the models DT, RF, KNN and NN was compared. Using the error as a performance metric it was realized that Random Forest has the best performance

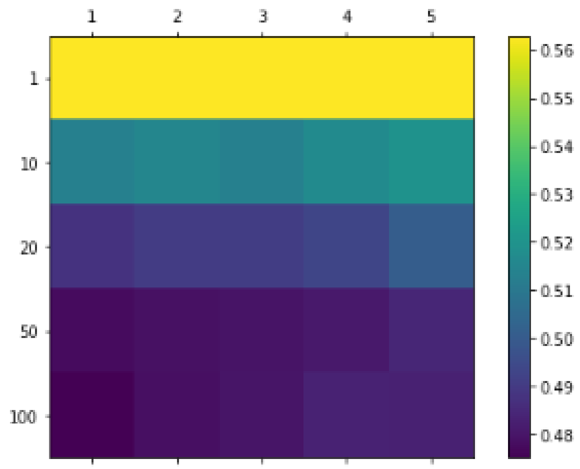

Figure 6: KNN validation error graph by varying K and A.

among all employed algorithms having a small error rate of 0.4179. Also, this model is highly efficient since it is very fast compared to other models and used less system power.

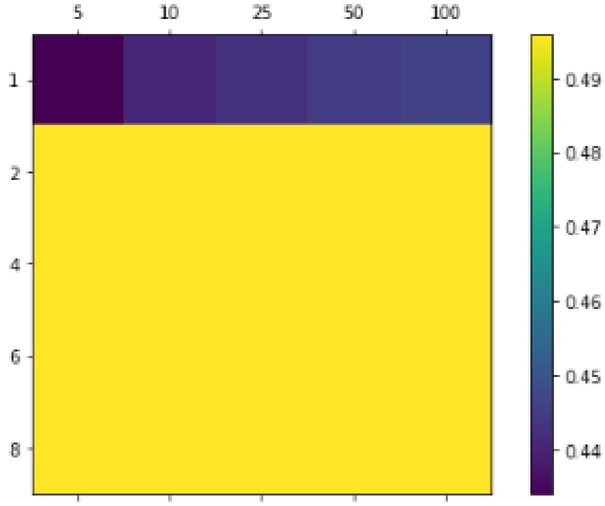

Figure 7: Validation error graph for NN varying hidden layers and nodes per layers.

5.2 Phase 2

After initial phase it was realized that Accuracy is more interpretable metric for performance comparison then mean squared error as MSE can be greater than 1 too while AUC can be only between 0 to 1 hence for second phase the AUC performance metric was used for model. Also, during the second phase more advanced libraries were used including sklearn, Keras and Tensorflow. For this section, again the same train-test split 75%/25% was used. Initially the Random Forest classifier was used from Sklearn. Multiple hyperparameters were tested in order to get maximum accuracy. The maximum accuracy of 0.56 was obtained using

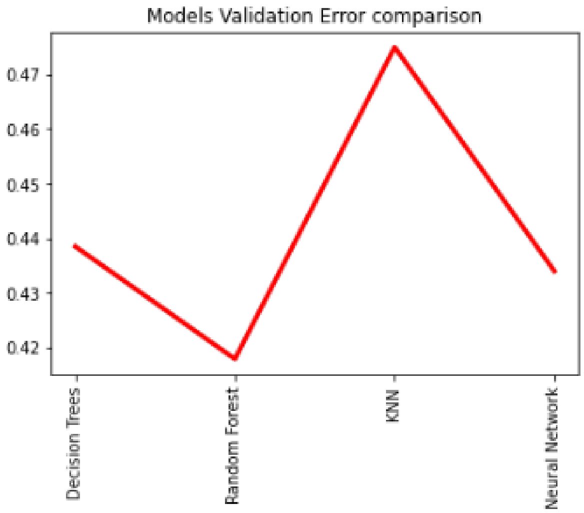

Figure 8: Graph comparing the performance of multiple algorithms using MSE for multi-class classification.

nestimators of 100, maximum depth of 180 and maximum features of 11. After Random Forest, from SKlearn library multilayer perceptron was used. Using hyperparameter tuning, testing accuracy of 0.58 was obtained using hidden layer size of 200 and activation function of htangent. Then again from Sklearn, KNN classifier was used. Number of neighbours were varied and maximum accuracy of 0.5245 was obtained. The forth model was again from Sklearn which is called logistic regression. Using balanced class weight and L1 penalty, the maximum accuracy obtained was 0.539. Fifth model which was employed for this problem was neural network from Keras and highest accuracy of 0.5035 was obtained using neural network.

Table 3: Highest accuracies of models for multi-class classification

Model	AUC
K Nearest-Neighbor	0.5245
Random Forest	0.56
MLP	0.58
Neural Network	0.5035
Logistic Regression	0.539

6 Model Exploration for binary classification

After first phase of model exploration and testing, class distribution was down-sampled. Both the no admission and readmission after 30 days was treated as class 0 and the readmission within 30 days was treated as class 1 to check the performance of our models. Hence, in this

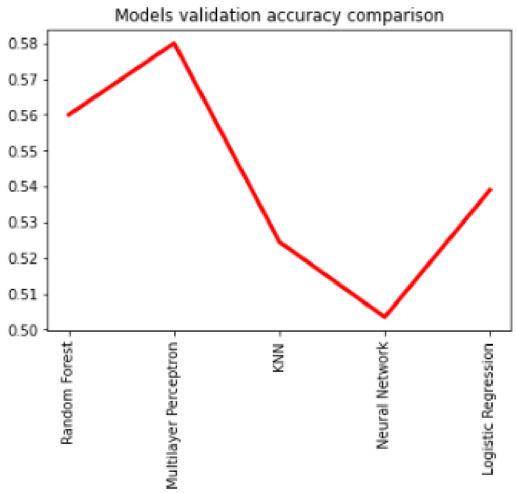

Figure 9: Graph comparing accuracy of different models for multi-class classification.

way multi-class classification was converted into binary classification for ease of analysis. All the previous mapping methods were employed in this case too. Once more Random Forest classifier was used which gave the maximum accuracy of 0.64 using minimum parents parameter as 150 and maximum depth as 150 also. Multilayer perceptron from sklearn gave the testing accuracy of 0.66 using hidden layer size of 200 and activation function as tanh. Neural network from Keras gave accuracy of 0.65 using batch size of 30 and epoch of 200. KNN gave the highest AUC as 0.649 and logistic regression gave accuracy of 0.59 using balanced classes and penality of L1.

Table 4: Highest accuracies of models for binary classification

Model	AUC
Random Forest	0.646
Multilayer perceptron	0.66
KNN	0.649
Neural Network	0.65
Logistic Regression	0.59

Although down-sampling of class distribution from multiple classes to binary classes increased the accuracy of prediction but it is still very low. One of the main reason for smaller accuracy is that, the data used in this project is real clinical data hence it is very redundant and non-uniform even after pre-processing and filtering and labels are very sparse instead of uniformly distributed. In addition to this, it has multiple missing values for important features for example weight. Therefore, after using simpler models and hyperparameter tuning, it was decided to use more advanced techniques e.g. Gridsearch, pipelines and cross-validation.

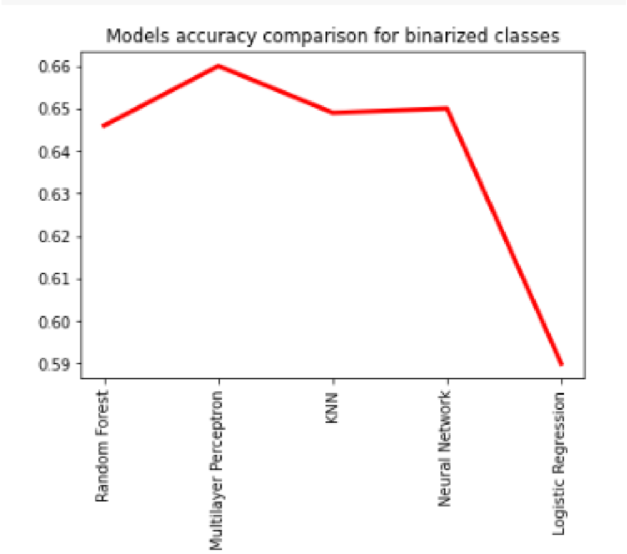

Figure 10: Models accuracy comparison for binary classification.

7 SMOTE and Grid Search

As explained in the beginning, there is very clear class imbalance in the data. Most of the instances in this data set belong to either no admission or readmission after 30 days. There are various approaches that can be used to deal with class imbalance. One approach is to use higher penalty for minority class in loss function. But in this project, it was decided to use a more sophisticated technique to deal with class imbalance which is called SMOTE. SMOTE is defined as Synthetic Minority Oversampling Technique which is often used to deal with class imbalance in machine learning problems [7].

7.1 Logistic Regression Grid Search

In this model,Data was pre-processed using sklearn to avoid convergence issue. Logistic Regression with balanced class and L1 penalty was implemented initially. And then logistic regression with L2 penalty and balanced classes was employed. 4 fold cross validation technique was used to find out the accuracy of this model. Both models with L1 and L2 gave validation accuracy of around 0.63. The Gridsearch method was implemented using logistic regression and SMOTE. The best estimator gave the accuracy of 0.60.

7.2 Multilayer perceptron Grid Search

Using same splits, neural network pipeline method was employed. SMOTE was used as sampling method, MLP classifier from sklearn was used as the model. In the grid search, 3 fold cross validation was used, with the activation function of 'relu'. The best estimator used the alpha=0.015 and hidden layer size (20, 20, 20) and gave best accuracy of 0.75.

7.3 Random Forest Grid Search

Third time, in the pipeline, random forest classifier was used with sampling type as SMOTE. Using maximum depth as 12, nestimators as 500 and 4 fold cross validation through grid search Random Forest classifier gave the best accuracy of 0.68.

Table 5: Highest accuracies of models using SMOTE for binary classification

Model	AUC
Logistic Regression	0.60
Multilayer perceptron	0.75
Random Forest	0.68

Using the pipeline and gridsearch method, the best accuracy of 0.75 was obtained using multilayer perceptron from Sklearn. This accuracy is the best accuracy obtained so far in this project. This is why this model is final recommended model for this problem. This model has accuracy of 0.75, precision of 0.149, recall of 0.334.

8 Performance Validation

Multiple methods were used to validate the performance of our model and to avoid under-fitting and over-fitting. In the beginning, all the data was shuffled in order to generalize the class distribution and to get same result each time. A three fold cross validation was used on training set using SKlearn cross validation module. And finally the model was tested on a held out data set. This type of method allowed to avoid any possible under-fitting, over-fitting or reduced accuracy problem between runs.

9 Conclusion

During this project, the relationship between hospital management system and rate of hospital readmission rate was studied using machine learning model. The accuracy of 75 percent was obtained using multi-layer preceptron through grid search, pipelines and SMOTE. Through the understanding of this relationship, customized protocols for patients of hyperglycemia can be implemented in the hospitals to achieve better outcome both in the terms of mortality and morbidity.

10 Future Directions

Multiple different models were used for the purpose of this project and performance of those models were compared to keep the model with best accuracy. Although, the accuracy obtained for binary classification, 75 percent, is a good number but the accuracy for multi-class classification is still very low. That is due to the fact that real clinical data comes with a lot of noise, missing values and uneven distribution of labels. In the future, author would like to work to increase the accuracy for multi-class classification too. Also, in the future author would explore SNORKEL technique [8] for model training. In addition to this, it was realized that machine learning models can help to find out the hospital readmission rate for patients using information related to hyperglycemia care system in hospital but we still don't know the relationship of each individual feature with the outcome. Hence, in the future shap values technique [9] will be used to get better insight regarding relationship of each individual feature with outcome.

References

[1] 2022. Center for disease control and prevention. [online] Available at: https://www.cdc.gov/diabetes/data/statistics-report/index.html [Accessed 28 June 2022].

[2] 2022. Center for disease control and prevention. [online] Available at: https://www.cdc.gov/nchs/fastats/diabetes.html [Accessed 28 June 2022].

[3] J, D., 2007. Prevalence of Hyper- and Hypoglycemia Among Inpatients With Diabetes: A national survey of 44 U.S. hospitals. [online] Available at: https://diabetesjournals.org/care/article/30/2/367/28379/Prevalence-of-Hyper-and-Hypoglycemia-Among [Accessed 28 June 2022].

[4] S, M., 2013. [online] The Chronic Care Model and Diabetes Management in US Primary Care Settings: A Systematic Review. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604796/ [Accessed 28 June 2022].

[5] Strack, B., 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. [online] https://www.hindawi.com/journals/bmri/2014/781670/. Available at: https://www.hindawi.com/journals/bmri/2014/781670/ [Accessed 28 June 2022].

[6] Archive.ics.uci.edu. 2014. UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 [Accessed 28 June 2022].

[7] Docs.microsoft.com. 2021. SMOTE - Azure Machine Learning. [online] Available at: https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/smote [Accessed 28 June 2022].

[8] Snorkel.org. 2019. Introducing the New Snorkel. [online] Available at: https://www.snorkel.org/blog/hello-world-v-0-9 [Accessed 28 June 2022].

[9] Kaggle.com. 2022. SHAP Values. [online] Available at: https://www.kaggle.com/code/dansbecker/shap-values/tutorial [Accessed 28 June 2022].