Prediction of Readmission of Diabetic Patients

Aman Aggarwal
2018327
aman18327@iiitd.ac.in

Gavish Gupta
2018390
gavish18390@iiitd.ac.in

Samarth Chauhan
2018410
samarth18410@iiitd.ac.in

Abstract

Diabetes is a chronic metabolic disorder in which prevalence has been increasing steadily all over the world. As a result of this trend, it is fast becoming an epidemic in some countries of the world. The number of people affected is expected to double in the next decade. Due to an increase in the ageing population, they are thereby adding to the already existing burden of hospital management, especially in poorly developed countries. This project is our contribution to the doctors and hospital management in these difficult times (COVID).
The purpose of this study is to classify the patients by proposing a machine learning model to classify that a patient will be readmitted within 30 days or not. This model will include risk factors like age, race, gender, etc. (Dataset Description). In a broader sense, the goal of this study is to help the hospital management by providing them with a reasonable estimate of whether a patient will be readmitted or not. So our research will help to increase the quality of care and will increase efficiency and reduce the burden of the hospital management staff.
The github repository for this project can be found here.

1. Introduction

Given the data of patients suffering from diabetes mellitus, we want to propose a model to predict that the patients admitted once will be readmitted again within 30 days period or not.

2. Literature Review

1. [4] In this study, pre-processing was done by dividing the by dataset by age groups, all features with none values were dropped, and all similar columns were clustered into one. This study didn't mention any specific data selection technique, but separate feature selection was done for each dataset. Train- test ratio taken was 3:1. And three different varieties of modelling algorithms were used. All the models were benchmarked using 10 fold cross-validation. LACE model, which is used in real life, was used along with the metrics like f1 score, accuracy to provide classification reports.
2. [2]In this study, extensive feature engineering was done as a part of pre-processing. In the study, Approximate Bayesian Bootstrap was used to fill the none values. And columns were clustered to create new columns. P-values with logistic regression were used for selection, but it gave bad results so random forest was used for feature selection. The dataset used in this study was prone to class imbalances, so to handle that SMOTE was used. For modelling, MLP was used with L2 error, Adams optimiser, and PRelu activation. For benchmarking, results from previous studies were taken, which included results of neural networks, random forests, etc.

3. Dataset

There are 50 features present in the dataset. Out of 50, 24 marks the dosage of medications like Metformin, Glipizide, etc., and the remaining 26 included features like age, gender, weight, time in the hospital, etc. The 24 features which were marking the dosage of medications can possess four different types of values:- Up(Increase dosage), Down(Decrease dosage), Steady(Same dosage) and No(Drug Not Prescribed). The feature "age" was present in 10 groups showing the age groups like [0,10), [10,20) etc., and 66.2% of the total dataset population were older than 60. The feature "race" consisted of values like African American, Caucasian(74.7% of the entire dataset population), Other and also there were some missing values. The feature "readmission rate" consisted of three types of values:- "NO", "30" days and "30" days. There were also the results of the A1c test were shown, which is a preliminary test for diabetes mellitus. And there were many more features which can be looked upon in the given.
UCI Machine Learning Repository

3.1. Preprocessing

Firstly, it was identified which features have missing values. There were seven such features. Out of these,

Features Affected	Pre-Processing Technique
Race, gender	String values converted to discrete integers
Age	Rounded off to nearest value in the set:{5,15,...,85,95}
Admission_type_id, discharge_id, discharge_disposition_id, admission_source_id	Similar values merged together and converted to discrete values.
Diagnosis features	The severity of illness categorised and converted into integers
24 Medicine doses	Converted into two categories - dose level changed or not
Readmission(output class)	No readmission and readmission after >30 days clubbed together as it has the same impact on the problem at hand

Figure 1. Feature-wise Preprocessing.

"weight", "payer code", "medical speciality" had the majority of values missing, so these features were dropped. For other features, the corresponding rows with missing values were dropped. After this step, 94% of the data was retained. The majority of features had string values, so they were converted to discrete integer values.

3.2. Feature Creation

A new feature was synthesized named 'service taken'. This feature was the sum of the columns 'number outpatient', 'number inpatient' and 'number emergency'. This feature summarised the comprehensive services utilized by the patient before being admitted to the hospital for the first time. The features 'Service taken', 'number outpatient', 'number inpatient', 'number emergency' had a lot of skewness, To overcome this, the log transform was applied to these features. This helped in synthesizing the next feature. Some of the features were multiplied with each other to find a better relationship between them. Some of these are (num of medications, time in hospital), (age, num of diagnoses). [5] [1]

3.3. Outlier Removal

The dataset included multiple occurrences of a single patient [3] [6]. For each patient, the subsequent data has been removed. The reason behind this being characteristic of the patient might change in terms of lifestyle, age, health, etc., which can affect the data. The data on which log transformations was applied have also been standardized.

3.4. Feature Encoding

The dataset included features like race, gender, admission type id, discharge disposition id, admission source id, max gluserum, A1C result, level1 diag1 which can be divided into classes, so we used One-Hot Encoding to encode them.

3.5. Feature Selection

Several techniques including ANOVA were tested to perform feature selection. The initial test using logistic regression and regularization technique coefficient's p-values appeared to drop highly significant features from a medical perspective. Hence, the random forest feature selection was performed. The variable importance was then computed during 48 iterations. A total of 25 predictors were thus selected after 48 iterations.

4. Methodology

We divided the dataset into three parts based on the 'age' feature after doing all feature engineering. Age groups in the dataset were 0-29, 30-69 and 70-100 years [4]. We treated each age group as an individual dataset. In the original dataset, output was divided into three classes no readmission, readmission within 30 days and after 30 days. No readmission and readmission after 30 days were merged into one class as readmission after 30 days contained records which are years apart hence they can create noise.

Once done with the preprocessing, we moved on to

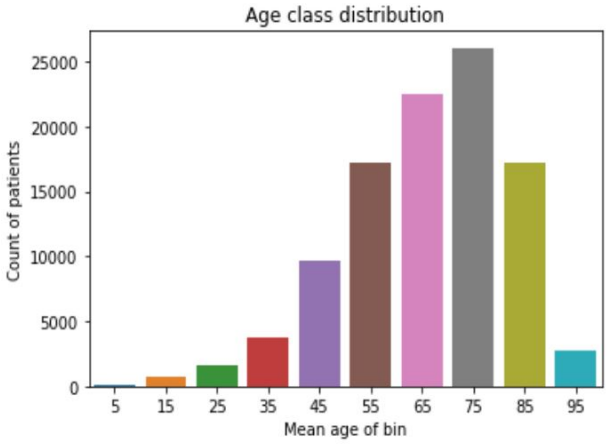

Figure 2. Age Distribution

apply the classification algorithms. The classifiers were trained and evaluated independently for each of the three age groups. Each of the classifiers was validated through K-fold cross-validation. We chose K=5. This was the most optimal choice considering the size of the dataset. A smaller K would have decreased the size of the training set resulting in lower accuracy as the classifiers won't be able to learn the parameters efficiently. On the other hand, a larger K would have added the computational overhead.

The first classifier chosen was the logistic regression *(C= 10, multi_class='ovr', penalty='l2', solver='saga').*

The results were not satisfying, so we applied SMOTE. After using SMOTE, we used and evaluated various classifiers. Logistic regression is the simplest of them. We found the best parameters of all the models using a grid search with the respective parameters. After logistic regression, we moved on to the Bayesian theory-based classifier (Naive Bayes) with an assumption that all the features are independent of each other and have zero correlation amongst them. Then tree-based models were used like decision trees(criterion='entropy', max_features='auto'), random forest *(criterion='entropy', max_features='log2', n_estimators=50)* and boosted decision trees along with regularisation *(L1 and L2 regularisation)* parameters (alpha and lambda) to prevent overfitting. Then we used KNN *(algorithm='auto', leaf_size=10, weights='distance')* and SVM *(C=1000, gamma='scale', kernel='rbf', max_iter=-1)* based models for our further analysis. [3] [6] [1]

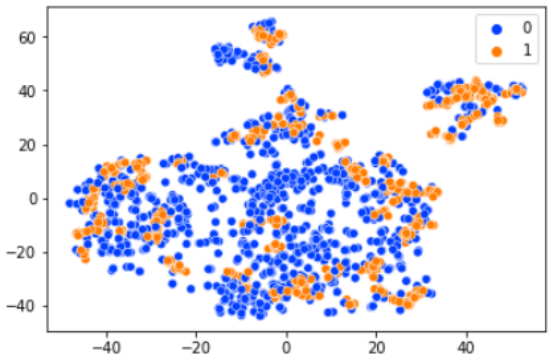

Figure 3. t-SNE plot of input features

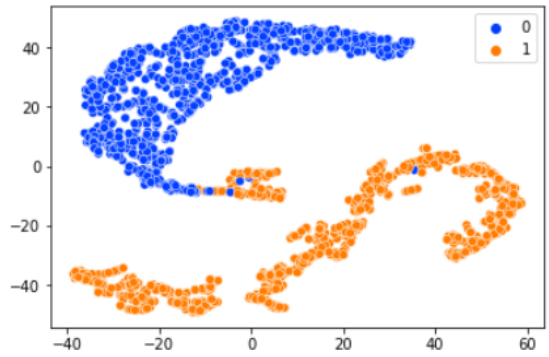

Figure 4. t-SNE plot of features of final hidden layer of NN

The last classifier used was Neural Network (NN). For the age group 0-30, the data was comparatively small and easier to classify. The architecture used for this dataset was [# input, 256, 64, 16, #output]. While for the other two age groups, i.e. 30-70 and 70-100, the data was comparatively large and difficult to classify. The network architecture used for these is [#input, 1024, 512, 128, 32, #output]. This was computationally hard but necessary for good results. [5]

4.1. Evaluation Metric

Since we divided the dataset into three groups based on the age of the patient, we decided different evaluation metrics for each of them. For data corresponding to age group 0-30 and 30-70, MCC score was a sufficient measure to determine the best classifier. Along with these other metrics like precision, recall, f1-score were also taken into consideration. Those classifiers which were giving excellent accuracy but performed poorly on other metrics were not chosen.

For data corresponding to age group 70-100, we focused more on having low false-negativity. The reason for the following being that this group denotes the senior citizens. If the classifier predicts the patient doesn't need to be readmitted within 30 days, but actually the patient had to be readmitted, this can lead to loss of life, which is a severe consequence. The model with decent accuracy and the comparatively low false-negativity (high MCC score) was chosen in this case.

5. Results and Analysis

The main objective of this paper was to classify the patients according to their readmission rate correctly. We have described the metrics used for the following classification task in section X.Y. For the given task, we started from linear models like logistic regression and then moved on to the advanced models.

We started with the logistic regression, but the results were not satisfying. There could be two reasons behind this- Non-linearly separable data and class imbalance. The low precision values for one of the class pointed out that class imbalance could be the potential reason for low precision and recall values. We solved the class imbalance problem by applying SMOTE.

Logistic regression now gave better results except for the MCC Score, which was relatively low. So this depicted that we need to use some advanced complex models which will better maintain the relationships hidden in the data. So we ruled out Logistic Regression. Then, We used Naive Bayes, which provided even worse results.

In the below figures, first/green boxplot of every point in x represent MCC score, second/yellow boxplot represent accuracy and third/purple boxplot represent ROC AUC.

	Precision	Recall	F1-Score	MCC
Logistic Regression	0.85	0.84	0.84	0.68
Naive Bayes	0.79	0.67	0.63	0.44
Decision Tree	0.91	0.91	0.91	0.82
Random Forest	0.96	0.96	0.96	0.91
XGBoost	0.96	0.96	0.96	0.92
NN	0.97	0.97	0.96	0.91
SVC	0.93	0.93	0.93	0.85
KNN	0.87	0.83	0.83	0.70

Figure 5. For Age between 0-30

	Precision	Recall	F1-Score	MCC
Logistic Regression	0.72	0.72	0.72	0.43
Naive Bayes	0.61	0.52	0.40	0.09
Decision Tree	0.91	0.91	0.91	0.81
Random Forest	0.96	0.96	0.96	0.92
XGBoost	0.96	0.96	0.96	0.92
NN	0.94	0.93	0.93	0.85
SVC	0.72	0.72	0.72	0.43
KNN	0.88	0.85	0.85	0.73

Figure 6. For Age between 30-70

	Precision	Recall	F1-Score	MCC
Logistic Regression	0.69	0.69	0.69	0.389
Naive Bayes	0.68	0.56	0.47	0.20
Decision Tree	0.88	0.88	0.88	0.76
Random Forest	0.94	0.94	0.94	0.88
XGBoost	0.94	0.94	0.94	0.88
NN	0.93	0.92	0.92	0.85
SVC	0.70	0.70	0.70	0.40
KNN	0.87	0.84	0.83	0.70

Figure 7. For Age between 70-100

Then we tried to use some tree-based models, so we started with decision trees, and the results were quite

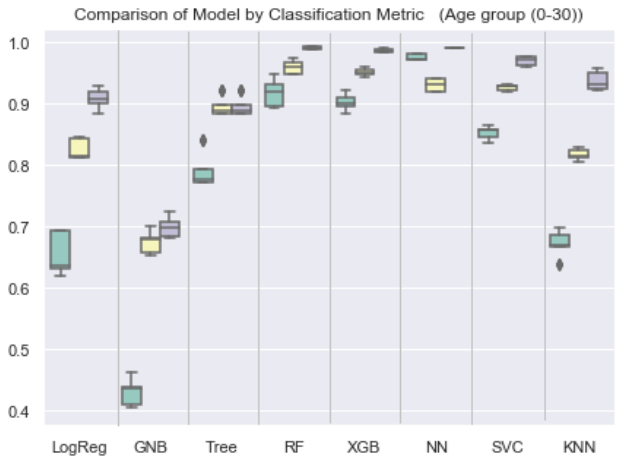

Figure 8. Model comparison for Age 0-30

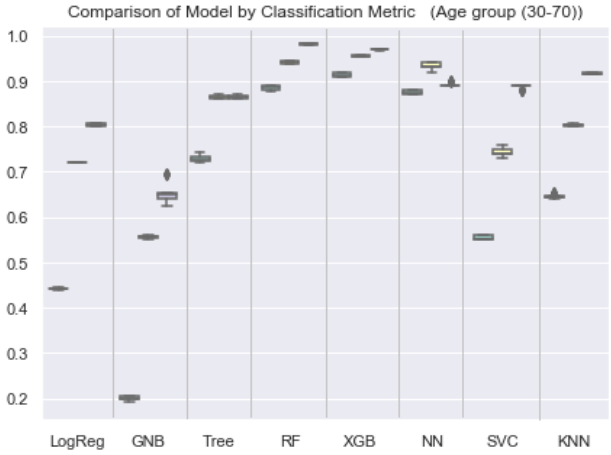

Figure 9. Model comparison for Age 30-70

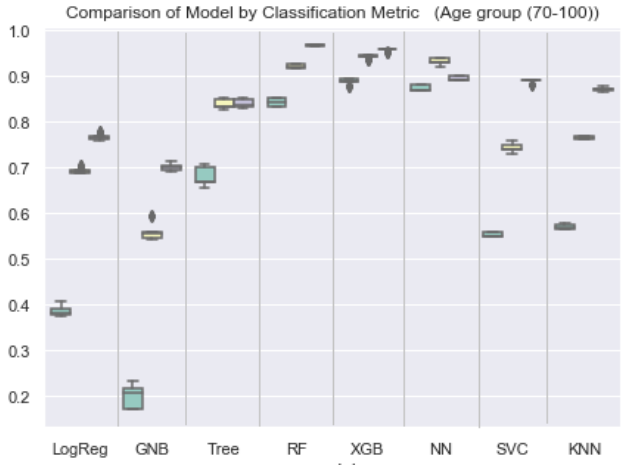

Figure 10. Model comparison for Age 70-100

astonishing. We obtained the best possible results till now as depicted in the fig X. We received 90% accuracy with reasonably good MCC Score. So then we tried using Random forests which use Ensemble Learning, and the results were even better than decision trees (95% accuracy), and the MCC Score was also very high. Then we tried to use boosted decision trees using XGBoost, but there was no significant advancement. The formula for MCC is given in Fig 11.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Figure 11. Formula of MCC score

Then we greedily tried out KNN, but we failed badly as the results obtained were not up to the mark of Tree-based classifiers. Then we finally used SVM, but the results were still not satisfactory which can mean that data has more than one complex(or simple boundaries) which the linear models like logistic regression and SVM are incapable of finding.

So Tree-Based models i.e. **Random Forest, XGBoost** performed best in our case, and these even outperform the neural networks. The plausible reason for this can be that the Tree-based models use a deterministic based approach. In contrast, a neural network's probabilistic based approach tends to over complicate a simple problem.

6. Conclusion

Our research suggests that applying a machine learning model with sound feature engineering and feature selection can outperform the current models in use like LACE by a huge margin, improving patient outcomes and lowering inpatient cost to hospitals. The highest performing models were those developed around age groups rather than a general "all" age groups. This study is for diabetic patients only, but we believe that it can be used for other medical conditions as well. With an updated data set added with some other valuable features can increase the already increased accuracy.

6.1. Individual Contribution

1. **Samarth Chauhan :** Literature Review, Statistical Analysis, Modelling, Hypertuning, Feature Creation, Benchmarking

2. **Aman Aggarwal :** Dataset study, SMOTE, Feature Selection, Modelling, Hypertuning

3. **Gavish Gupta :** Dataset study, Statistical Analysis, Feature Selection, Feature Creation, Benchmarking

References

[1] A. Hammoudeh G. Al-Naymat I. Ghannam and N. Obied. Predicting hospital readmission among diabetics using deep learning, 2018.

[2] Ti'jay Goudjerkan and Manoj Jayabalan. Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron, 2019.

[3] C.Y. Lin H. S. Singh R. Kar and U. Raza. What are predictors of medication change and hospital readmission in diabetic patients, 2018.

[4] Damian Mingle. Predicting diabetic readmission rates: Moving beyond hba1c, 2007.

[5] D. Rithy. Simulation of imputation effects under different assumptions, 2016.

[6] C. Chopra S. Sinha S. Jaroli A. Shukla and S. Maheshwari. Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients, 2017.