Machine Learning Classifiers and Along with TPOT Classifier (Automl) to Predict the Readmission Patterns of Diabetic Patients

Phani Siginamsetty, V. Krishna Reddy

Abstract: ***Diabetes is seen as a common problem in the present running world. And till date 470million people globally in 2019, and it might be increased to 676million by the end of 2045.So day to day the diabetic has become a major problem, and due to the current technologies, we can easily predict the readmission of a patient based upon his digenesis. In this paper we are using classification algorithms to solve the problem by early predictions. And we can check it by using multiple hybrid classifiers, whatever the algorithm gives the best accuracy we are considering it as the generic model and it is going to predict the future diabetic patients. And we are considering the diabetic dataset mainly it consists of multiple features based upon the data we will consider as independent and dependent data, and solve the problem. Here, in this paper the algorithms which we are going to use are Logistic Regression(LR), Decision Trees, Random Forest (RF),XGboost,Gaussian-Naïve Bayes, TPOT(automl).Out of them Random Forest gives the best accuracy which is about 95.2%, the accuracy is attained by following pre-processing stage in a good manner, and handled all missing data.***

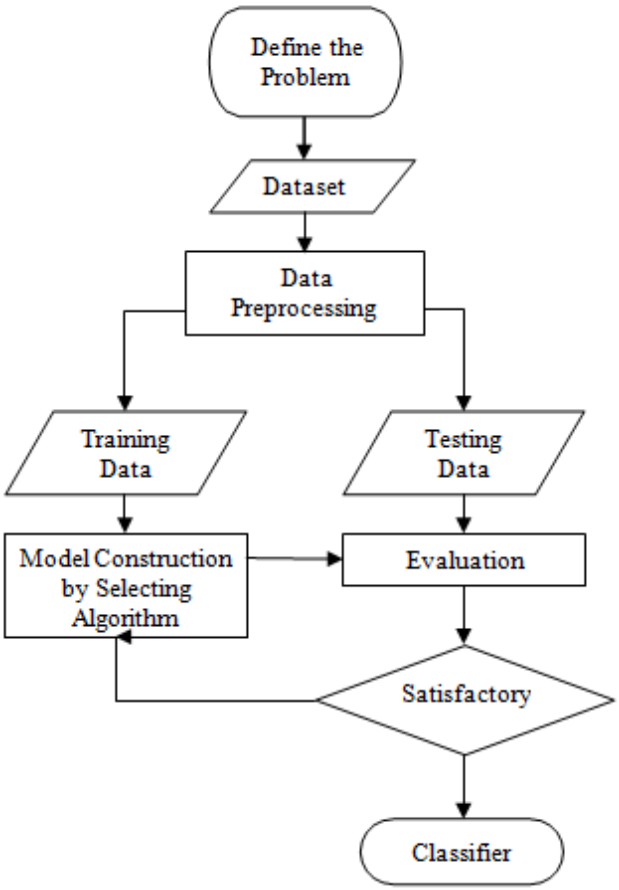
Index Terms: diabetes prediction, Logistic regression, decision trees, Random Forest, XGboost, Gaussian Naïve-Bayes, TPOT(AutoML)

I. INTRODUCTION

Diabetic is seen as a common problem; without proper treatment it can't be cured. We can easily identify diabetes as the glucose levels in blood increases than the normal level and more over if glucose levels in blood increase that might be damage of tissues like eyes, kidneys, heart, blood vessels and nerves. Mainly we can consider two kinds of diabetes, i.e. type 1and type 2. Type 1 diabetes is generally seen in younger mostly less than 25 years old, and the symptoms to identify the type1 diabetes is frequency of increase in thirst and urination and it can be solved by insulin therapy. [1] Type2 diabetes is most commonly seen in middle and old age people and symptoms are obesity, hypertension, and other diseases [2] [3][4]. Machine Learning can help people to predict the diabetes and analyze weather the person is readmitting or not. And we have to select a proper dataset and a generic model that actually classify the problem and it gives better accuracy. In general Machine Learning is playing a major role in solving problems in multiple domains. That including Healthcare, Stock Market, Finance, and others. Based on the problem statement we have to select the dataset and make the data to fit to the model then we can predict the data with a, better accuracy.

In this paper, we have taken the dataset from UCI Machine Learning Repository. And mainly this repository is consisting of multiple databased, hedge number of datasets and number of theories. Dataset we are using in this paper consisting of thousands of records, so we can observe each and every feature and finally analyze which are essential for us to analyze.[5]

And the process of the ML can be seen in the flow chart and we can easily understand the process of ML project. In each stage there are multiple tasks and according to the problem we are going to so solve it and finally we can achieve the good accuracy. And, we can check Recall, Precision, F-Score. Our problem is classification so we are going to consider recall, precision and F-score if our problem is regression then we can check the R^2 and adjusted R^2 values to check the accuracy of the model.



Revised Manuscript Received on April 21, 2020.
Phani Siginamsetty, MTech Computer Science, KL University Vaddeswaram, Guntur District siginamsettyphani@gmil.com
Dr.V. Krishna Reddy, Professor, KL University, Green Fields Vaddeswaram, Vijayawada

II. DIABETIC DATA SET

Features	Type	About Data
encounter_id	Unique Values	Information about the encounter id
Patient number	Unique Values	Information about patient number
Race	Categorical	Patient belongs to which continent
Gender	Categorical	Describes the gender
Age	Numerical	Describes the age of patient
Weight	Numerical	Describes the weight of patient
admission_type_id	Numerical	Source of admission
discharge_disposition_id	Numerical	Information about discharge
admission_source_id	Numerical	Information about admission id
time_in_hospital	Numerical	days between in and out
payer_code	Categorical	Describes the patient additional information.
medical_specialty	Categorical	Type of get through the admission
num_lab_procedures	Numerical	Count of lab tests
num_procedures	Numerical	Count of procedures
num_medications	Numerical	Count of distinct generic names
number_outpatient	Numerical	Count of outpatient visits
number_emergency	Numerical	Describing the emergency visitors of the patient.
number_inpatient	Numerical	Describing the number of inpatient visitors of the patient
diag_1	Numerical	The early diagnosis done for the patient
diag_2	Numerical	Secondary diagnosis done for the patient
diag_3	Numerical	And more secondary diagnosis done for the patient.
number_diagnoses	Numerical	Count of diagnoses
max_glu_serum	Numerical	Describes the range of glucose
A1Cresult	Numerical	Describes the range of level
Change of medication	Categorical	Describe information about change in diabetic medications
Diabetes medication	Categorical	Describe information about diabetic medication prescribed.
features for medications	Categorical	It's all about generic names.
Readmitted	Categorical	Days to inpatient readmission.

III. DATA EXPLORATION

A.Data

The dataset collected from UCI Machine Learning Repository it mainly consisting of multiple kinds of datasets and any kind of domain data is available in this repository. So, we can easily access the data for the future prediction [6].

B. Data Exploration

The distribution of the dependent feature was analyzed and come to know that 30% of the patients are readmitted within 30 days, so we are going to find out the insights present in the data and going to solve the problem.

C. Data Preparation

Extreme Values (Outliers) These values are predicted by statistics considering the interquartile range as the parameters beyond it whatever the data point in the dataset is considered as an anomaly.

E.Irrelevant Data

As the dataset consisting of multiple records each record is represented as a patient if patient died, then we should no consider the data points as the inputs to the model.

IV. DATA PREPARATION

Before creating the model, we have to analyze the dataset and want to know what are the important features and unimportant and it can be measured by correlation. Whatever the features having less correlation then we can eliminate those features and consider remaining features as input to the model. It is an important face and entire project 30% of time is to analyses the data and find the insights present in data.

Missing Data

Missing values plays an important role in data preprocessing stage, we have to handle those values if not model accuracy is very less so in our dataset the missing vales are handles by,

1) Race (75% missing). So, the race feature consisting of 75% of missing values and moreover it is not as much important to predict the readmission. So, we can drop the column.

2), payer code is not that much related to output so we can drop the column and it doesn't lead to overfit the model.

Features	Type	% Missing Value	Replace
Race	Categorical	2 %	Replace with mean

Weight	Numerical	94%	Replace with Mean
payer_code	Nominal	58%	Replace with Mode
medical_specialty	Categorical	55%	Replace with Mode
diag_1 diag_2 diag_3	Numerical	75%	Replace with Mode

Attribute Selection

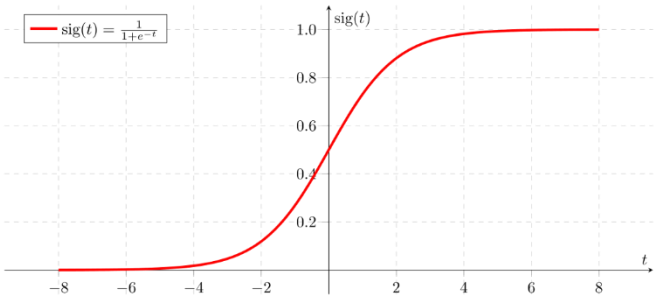
Encounter ID and **Patient Number**, these features are not related to output so even if we apply encoding technique it will take long time for data processing technique.

Algorithms

A. Logistic Regression

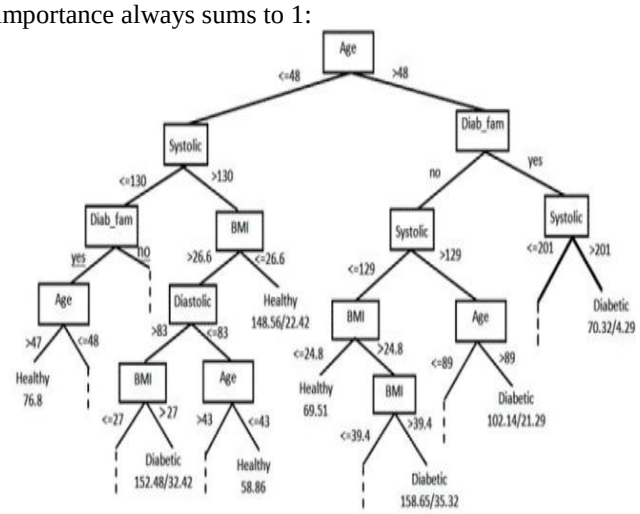
As we have seen our dataset, it consisting of multiple features so our problem is Multi-Class Classification. So, we have to predict weather the patient is going to readmit or not. And for this data problem we can use Logistic Regression. For multi-class classification, we can consider all the input features are independent and output feature is dependent.so our model is going to classify categorical problem. i.e. weather the patient is going to readmit or not and these kinds of problems can be easily solved by Logistic regression. The 40 features are formerly given to the Logistic Regression hypothesis, and output obtained can be in the form of 0 or 1.

Hypothesis => Z = WX + B
hΘ(x) = sigmoid (Z)



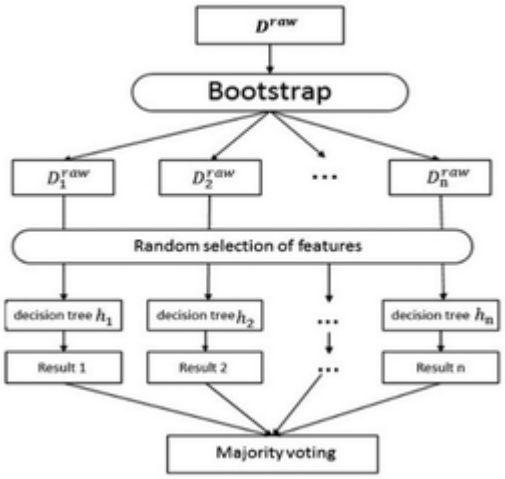
B. Decision Tree

Feature significance rates how necessary each function is for the choice a tree makes. It is a variety between 0 and 1 for each feature, the place 0 capacity "not used at all" and 1 skill "perfectly predicts the target". The characteristic importance always sums to 1:



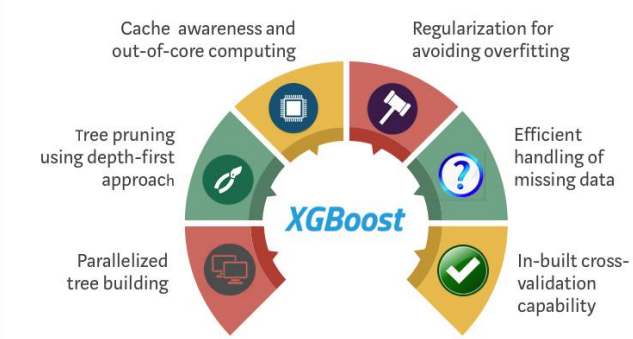
C. Random Forest

Random Forest is considered as the most used models for both classification and regression, Random Forest mean it consists of multiple Decision Trees. And mainly in RF we consider multiple DT and how many trees should we consider will be determined by the methodology hyperparameter tuning. So ,in this problem we are considering multiple DT and each tree is having some records of the dataset and it can be achieved by the row sampling and feature sampling with replacement .So each tree is going to train on some records and gives best accuracy ,initially we gives our some input to all the DT and when they get trained on data we are combining all the DT so our problem is classification and we consider the majority voting ,and predict the output. And if it a regression problem we consider the mean of the all the Decision Trees. We can call RF as Bootstrap process, where it is clearly mentioned in the fig. Our problem achieves high accuracy by using random forest. And gives an accuracy about 95.2%.



D. XGBOOST

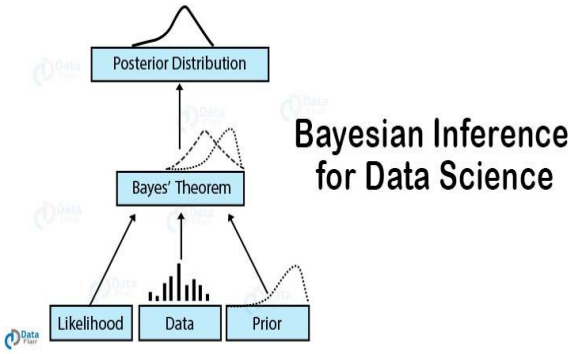
XGBoost is a kind of decision tree-based ML algorithm and moreover it is going to use the technique like gradient boosting. And in general, optimized gradient boosting techniques are using parallel processing mechanism, it will be handling missing values and regularization techniques.i.e. Ridge and Losso to remove the problems like overfitting.

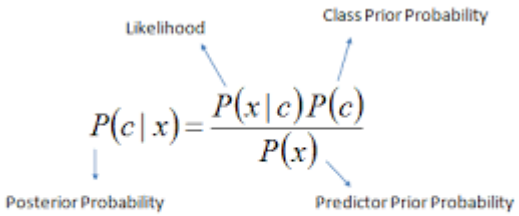


E. GAUSSIAN-NAIVE BAYES

Gaussian Naïve Bayes is based upon Bayesian theorem, and its user when the input dimensionality is high.

Mainly it is a classification problem. In machine learning we are going to select the best hypothesis (h) for a given data (D)
In Naïve Bayes classification we are considering probability and Bayes theorem provides a way to solve our problem. Naïve Bayes will consider the probabilities for each and every hypothesis is are simplified to make the calculations so simple.

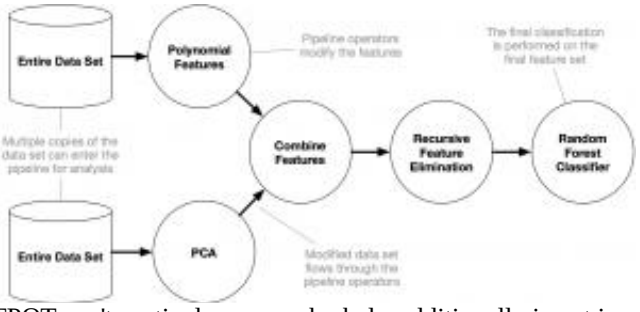




$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

F. TPOT (AUTOML)

TPOT is known as Tree-Based Pipeline Optimization Tool, is a genetic programming-based optimizer that generates machine
learning pipelines. It extends the scikit examine framework with its personal base regressor and classifier methods. It automates portions of the machine getting to know technique details in figure.
Like auto-sklearn, TPOT sources its information manipulators and estimators from sklearn and its search area can be limited through a configuration file. Time restrictions are applied to TPOT by means of altering the maximum execution time or the
population size. The optimization process also supports pausing and resuming. The most necessary characteristic of thisframework is the ability to export a mannequin to code to be further modified by means of hand.

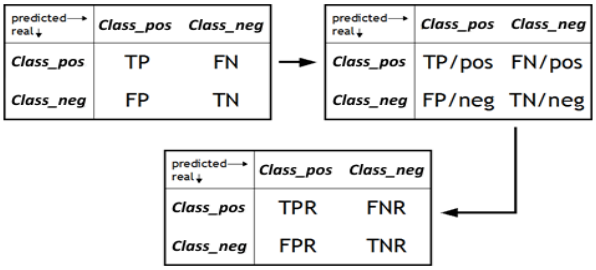


TPOT can't routinely manner herbal additionally is not in a position to techniques categorical strings which must be integer encoded earlier than passing in data. Also note that version 0.9 was used when testing this framework.

V. METHODOLOGY TO GET ACCURACY

Accuracy

Often when I talk to corporations that are searching to implement records science into their processes, they frequently ask the question, "How do I get the most correct model?". And I asked further, "What enterprise mission are you attempting to remedy the usage of the model?" and I will get the perplexing appear because the query that I posed does now not truly reply their question. I will then want to explain why I requested the question earlier than we start exploring if Accuracy is the be-all and end-all mannequin metric that we shall pick out our "best" model from. So, I notion I will explain in this blog put up that Accuracy want now not quintessential be the one-and-only model metrics statistics scientists chase and include easy rationalization of different metrics as well.



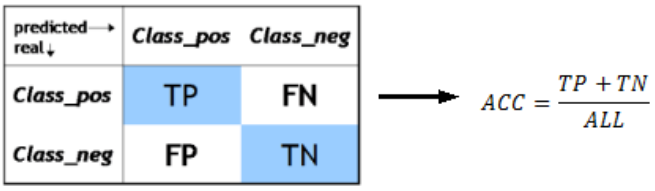


TABLE 4
PEER ASSESSMENT RATING INDEX COMPONENTS [2]

Name	Equation
TPR (True Positive Rate)	$TPR = \frac{TP}{FN + TP}$
FPR (False Positive Rate)	$FPR = \frac{FP}{TN + FP}$
TNR (True Negative Rate)	$TNR = \frac{TN}{TN + FP}$
FNR (False Negative Rate)	$FNR = \frac{FN}{FN + TP}$
Accuracy (ACC)	$ACC = \frac{TN + TP}{TN + TP + FN + FP}$
Precision (PRC)	$PRC = \frac{TP}{TP + FP}$
Geometrical Mean (GM)	$GM = \sqrt{TPR \times PRC}$

Precision and Recall

Precision and Recall are used in classification problem to determine the accuracy but in case of our problem is regression then we are going to use R^2 to determine the performance of model. And in this paper our problem is classification so we have to calculate the metrics i.e. Precision and Recall it mean Recall measures the original/actual data points where as Precision measures predicted data points. F1 score is also a metric and it is used to calculate model accuracy. Mainly there are multiple metrics used to test the model accuracy and based upon the problem statement we must use the metrics and check the accuracy. In machine learning we can use the default libraries and check the accuracy.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1 Score

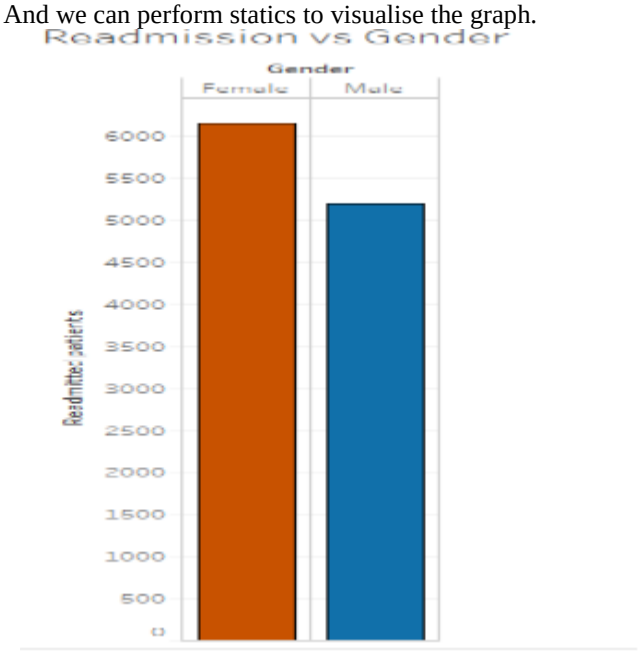
F1 score is used to check the model performance in terms of accuracy. Our problem statement is classification so we have taken F1 score metrics to measure the performance.

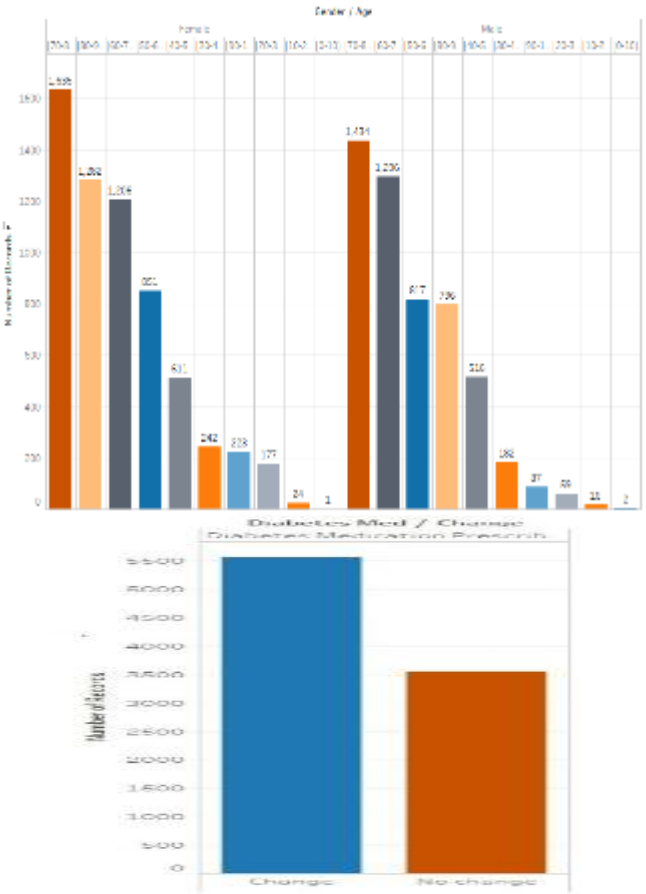
$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

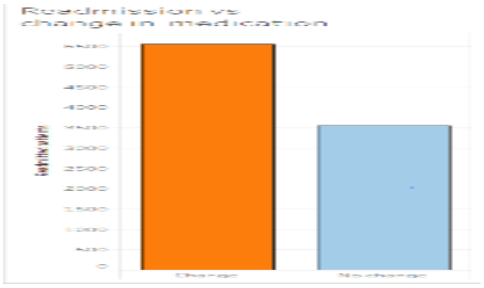
ERROR=1-ACCURACY

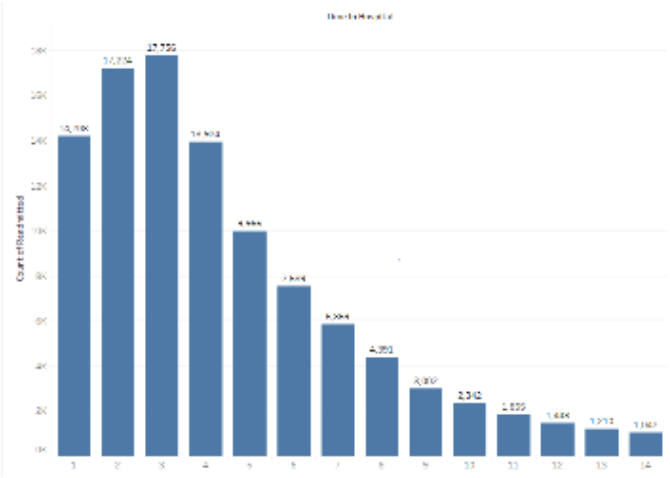
Readmission Vs Gender

Here we can plot a graph and it says the difference between the readmission and gender. In Multivariate analysis we can compare two or more features and in bivariate analysis we can compare two features. The comparisons of the graphs are mentioned below, so we can easily come to understand the dataset by using Seaborn and matplotlib. And we can compare features like gender and age and plot the graph. And we can perform statics to visualise the graph.









Models	Accuracy	F1-Score	Precision	Recall	Kappa value
Logistic Regression	76.42%	77%	76%	77%	52.7%
Decision Tree	80.54%	81 %	80%	83%	79.1%
Xgboost	88.1%	81%	82%	81%	88%
Gaussian NB	89.2%	88.1%	88%	89.2%	89%
TPOT	90.2%	90%	89%	89%	89%
Random Forest	95.2%	95%	94%	92%	91%

The visualisation of graphs is clearer and we can compare any two features by using multivariate analysis. And we can compare based on readmission on the time spent in hospital.

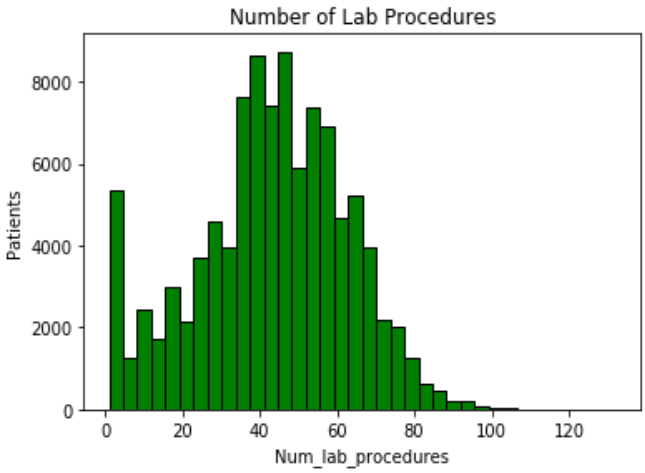
Distribution Plots

We can use distribution plots of more and better understanding and it will be useful of understanding of the analysis.

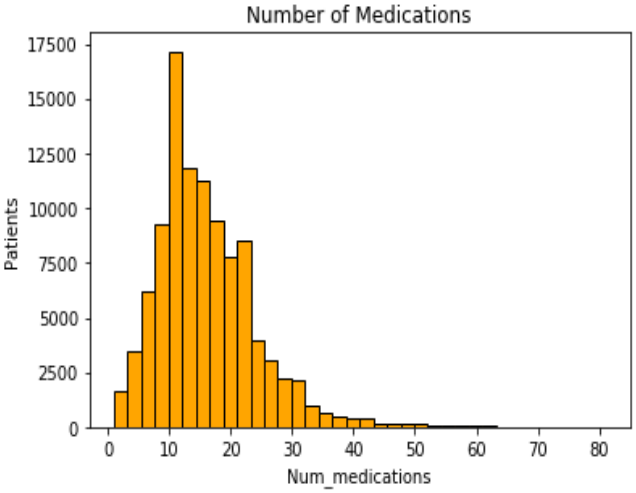
As we consisder the plots we can compare any two features i.e is known as the bivariate analysis. And the fig below listed are

comparision between the Num_lab_procedures vs patients.

And other distribution plot is Num_medications vs patients,and you can analyse the distrubutions using matplotlin.The distribution is as good to understand the actual problem .



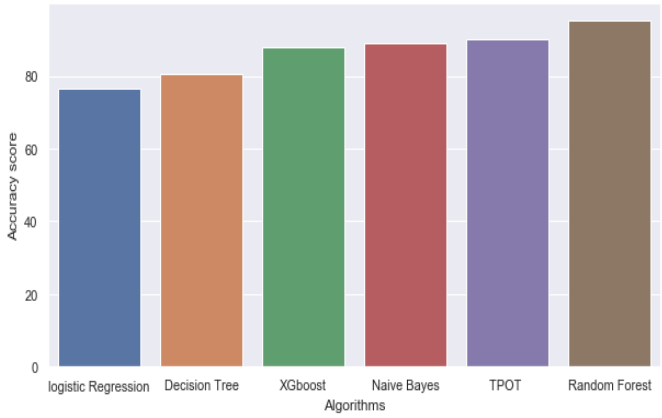
And other distribution is Num_medications vs patients. Basically, the distributions are considered for the better understanding of the data and more over based upon the distribution we can use Scaling Techniques and transform the data into a proper distribution.



VI. MODEL EVALUATION

And the accuracy of the models can be evaluated by using the below figure, as we can observer that the accuracy is increased in the Random Forest compared to all the algorithms.

And mainly we can easily get good accuracy in Random Forest because it is using multiple decision trees and moreover our problem is binomial distribution and it can be easily solved and it is going take the more repeated value as output.



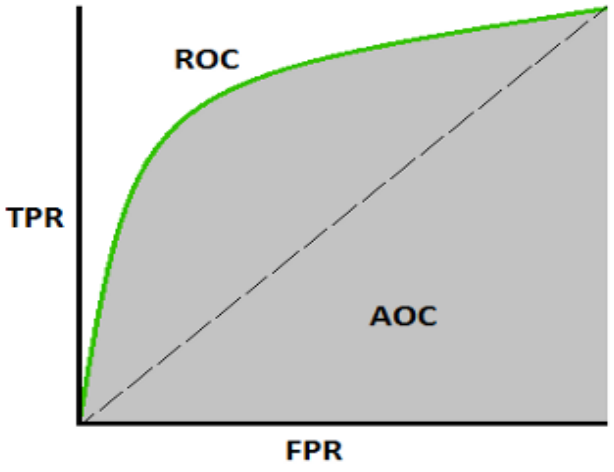
In general, our actual problem is classification so there are multiple algorithms to predict the result and beyond it there is an AutoML basically, it is a tool and it can check with all the possibilities of all the algorithms and based on problem it will consider algorithm itself and gives the result. And even today TPOT is still in development and we can just check the result after doing our problem by checking valid algorithms.

ROC CURVE

To know the model performance in binary classification we are using AUC curve. And it will tell us how good our model was? And higher the AUC then our model predicted the correct classification. By using Roc curve, we can easily determine any kind of model accuracy.

ROC curve mainly describes the performance of the classification if the value is high then our model performance was good. AUC it means Area Under ROC curve and it measures the performance of possible thresholds.

And we can see in fig the curve has been plotted true positive rate vs the false positive rate and considering multiple values of the thresholds.

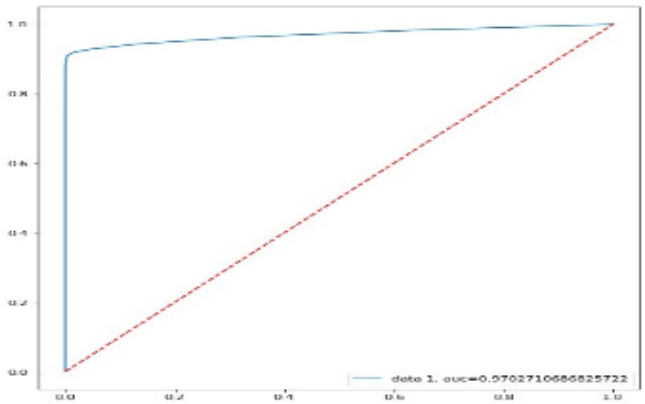


$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{FP}{TN + FP}$$

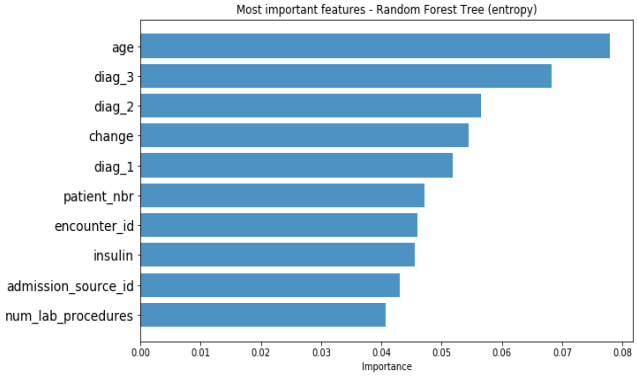


Confusion Matrix

Confusion Matrix's are used to check the predicted results on a classification problem. SO, we are comparing the actual values vs predicted values. i.e. here, actual means Recall values and Predicted means Precision values. So, if our problem statement is classification, we can use confusion matrix and let to know the type of errors occurred.

Actual /Predicted	Not Readmitted	Readmitted
Not Readmitted	17214	81
Readmitted	12	15560

Most important features to perform the task to predict the readmission, we consider Random Forest (Entropy)



VII. RESULT

As seen in the model evaluation stage the dataset was been tested with multiple classifiers and finally come to conclusion that the Random Forest Classifier has classified the patients into tow classes with respective having diabetic and non-diabetic. And not only consider the accuracy metrics and here in this metric classification it was clear that the F1 score, precision and recall was also with good accuracy. Moreover, the graph **auc roc** curve was also good and the model is perfectly classifying the new test data. And our model is going to consider the most important features by using the techniques like lasso and extra tree classifier methods are used for future selection. And entire like cycle of the project followed the data science life cycle and that included all the steps like data collection, data pre-processing, data selection, model creation and model evaluation was up to the mark. So finally achieved the accuracy of **95.2%** by following the entire life cycle.

VIII. CONCLUSION

Prediction of diabetes is done using Random Forest classifiers for diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 78% to 95% is achieved for data set by using 8-fold cross validation and by spitting data into 30% testing and 70% training.

REFERENCES

1. Ryd´en L., Standl E., et al. Guidelines on diabetes, pre-diabetes, andcardiovascular diseases: executive summary: The task force on diabetes and cardiovascular diseases of the European Society of Cardiology ESC) and of the European Association for the Study of Diabetes (EASD). European Heart Journal, 28(1):88–136, 2007.
2. International Diabetes Federation. IDF diabetes atlas, 8th edition. http://diabetesatlas.org/IDF Diabetes Atlas 8e interactive EN/, 2017. Brussels,Belgium.
3. Carol M. Ashton, David H. Kuykendall, et al. The Association betweenthe Quality of Inpatient Care and Early Readmission. Annals of InternalMedicine, 122(6):415–421, 03 1995.
4. Sara J Healy, Dawn Black, et al. Inpatient diabetes education isassociated with less frequent hospital readmission among patients withpoor glycemic control. Diabetes Care, page DC 130108, 2013.
5. Reena Duggal, Suren Shukla, et al. Predictive risk modelling for earlyhospital readmission of patients with diabetes in India. InternationalJournal of Diabetes in Developing Countries, 36(4):519–528, 2016.
6. B. Strack, J. DeShazo, C. Gennings, J. Olmo, S. Ventura, K. Cios and J.Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates:Analysis of 70,000 Clinical Database Patient Records," *BioMedResearch International*, vol. 2014, p. 11, 2014.

AUTHORS PROFILE

Mr. Phani Siginamsetty He has done his MTech (CSE) from KL University, Guntur, Andhra Pradesh. The area of specialization is computer science. And he is enthusiastic in machine learning, deep learning, Natural language Processing (NLP) and AutoML.

Dr. V. Krishna Reddy Is working as a Professor in Department of Computer Science and Engineering, KL University, Guntur, Andhra Pradesh. The area of specialization is cloud computing. And at present he is focused on Machine Learning and AutoML.