Diabetes Disease Prediction Using Machine Learning

Rohini Patil[1], Leena Majumder[2*], Manisha Jain[3], Vedanti Patil[4]
[1]*Assistant Professor, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India*
[2,3,4]*Student, Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India*
*Corresponding author: leenamajumder55@gmail.com

***Abstract*: Diabetes is considered as a chronic and deadliest disease, it is caused due to increase in amount of sugar or glucose which is condensed into the blood. It is even known as the "Slow Poison" that deteriorates the condition of patient in such a way that it causes failure of various organs particularly eyes, kidney, nerves, heart and veins. It also results into fatal diseases such as cardiac and respiratory diseases. Many difficulties and complications occur if this lethal disease doesn't gets diagnosed on time, it remains unidentified and untreated. Identifying process of Diabetes is monotonous and tedious as the glucose and sugar levels needs to be checked before and after meal, there are fluctuations before and after meal, this whole process of patient visiting a doctor is tiresome. The leap in Machine Learning approaches and algorithms helps us to solve this intense issue. The motive of this study and research is to make use of noteworthy features and try to predict the likelihood of the disease, Decision Tree, Random Forest and Support Vector Machine are the algorithm or approaches that have been applied to detect and predict diabetes at an early stage. Pima Indians Diabetes Dataset is sourced from Kaggle, a data science community. Precision, F1, Recall and Accuracy are the performance measures on the basis of which the three Algorithms have been evaluated. Random Forest outperforms and as a result it has the highest accuracy of 78.35% as compared to other algorithms. These results are verified using Receiver Operating Curve(ROC) in a proper manner.**

***Keywords*: Diabetes, Machine Learning, SVM, Decision Tree, Random Forest.**

1. Introduction

Increase in the blood glucose or sugar levels causes Diabetes, Mellitus, it is clearly classified into two categories such as Type 1 and Type 2.

Type 1 is caused due to sedentary lifestyle and lack of physical activity which eventually results into obesity. Type 1 causes the body to knock down the cells that are significantly important to produce energy. Type 1 can occur in teenage years. Type 2 occurs when there is no suffice amount of insulin produced in our body via pancreas. This type generally occurs in the middle or aged groups. Sedentary lifestyle, Eating Junk, Toxic and Chemical Contents in food are the major causes of this disease. Foot Ulcers, Retinopathy, renal issues, cardiovascular and respiratory complications are caused majorly via this disease. Sugar Concentration is the main reason behind this deadly disease. Height, Weight, Diabetes Pedigree Function (that is Hereditary Factor) and Insulin are the other factor that causes this deadly disease. Machine Learning Algorithms and approaches such as J48, SVM, Naive Bayes, ANN, Decision Tree are used for conducting experiments for diagnosing the disease.

Data Mining Techniques are also used to predict the disease using ETL process mechanism and various approaches such as K means, K mediod, Adaboost, Agglomerative Algorithms have been used in this area of research.

This research study brings into limelight the people that have certain probability of having diabetes. In this system we have used the pre-existing data set called Pima Indian from Kaggle to train and evaluate our model which is an open source dataset. Pima Indian Diabetes Dataset consisting of 768 records has been selected from Kaggle to train, test and evaluate our model. The paper has been systematically divided into sections. Relevant and related work which was performed in the past is discussed in Section II. Proposed Architecture of the model is demonstrated in Section III. Section IV presents the design and methodology of different algorithm and approaches used. Utterly we have concluded our work in section V.

2. Relevant Work

Deeraj Shetty et al., [1] prediction of Diabetes Mellitus is made via Data Mining. The goal of Data Mining Methodology is to extract, transform and load data from a dataset and change it into reasonable structure for further use. The patterns are utilized for better clinical decisions predicting it at an early stage.

Dr. Pramananda et al., [3] uses Data Mining Techniques and Methods for Diabetes Prognosis. In this predicting model, Frequent Pattern Growth a 2step approach is used which allows frequent item set discovered without candidate item set generation. FP Tree is built which is comparatively smaller than the database and thus reduces the expenses in subsequent mining process.

Vrushali Balpande et. al., [2] suggested model is used to predict the severity of diabetes. Parameters like Age, BMI, HbAlc, FBG, PMBG are used in this approach where each and every attribute is tested. Eclat Algorithm calculates the condition of Diabetes.

3. Proposed Architecture

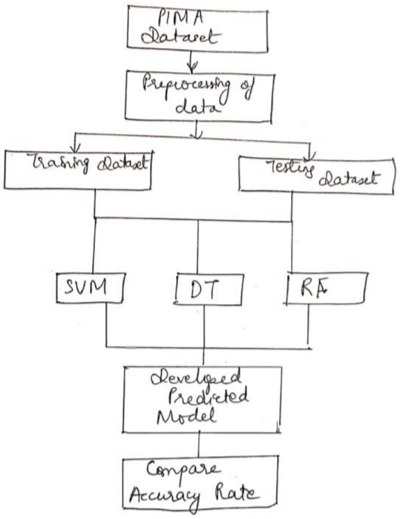


Fig. 1. Architecture of the diabetes prediction model

We have considered 9 attributes from dataset which determines whether the patient suffers from diabetes or not.

Steps in Diabetes Prediction Model:

1. Collection of Data
2. Data Preprocessing Technique
3. Machine Learning Model
4. Training of Data
5. Testing of Data
6. Analysis of Data
7. Output/Prediction Result

Table 1
Different attributes of Indian pima dataset

Class	Attribute
Numbers of time Pregnant	1
Glucose Test	2
Blood Pressure	3
Triceps skinfold thickness	4
2-Hour Serum Insulin	5
Body Mass Index	6
Diabetes Pedigree function	7
Age	8

Number of Times pregnant results into Gestational Diabetes. Glucose Test measures the amount of simple sugar in blood, it is done for type 1, type 2 and gestational diabetes.

With the help of contraction of heart muscle, the pressure of blood within arteries is produced a person suffering from diabetes suffers from high blood pressure as well.

Triceps Skin fold thickness gives us the information regarding fats of the body. The fats are stored as in glutes, which results into development of cellulite in women.

2 Hours Serum Insulin checks the insulin levels before fasting and after intake of glucose via food into body.

BMI is relation of body weight and height.18.5 kg/m^2 to 25 kg/m2 indicates a normal weight, anything less than this is considered as underweight, between 25 kg/m^2 and 29.9 kg/m^2 is considered as overweight.

Diabetes Pedigree Function is Heredity of your family.

A. *Support Vector Machine*

SVM is used for classification and regression purpose.

It is a supervised algorithm. SVM divides the whole dataset into samples such as positive and negative via a linear hyper plane we segregate the data points on the graph.

Hyper plane should be chosen such that it should be far from the data points from each category. We will maximize the distance between the hyper plane which is defined by $w^T x + b = -1$ and the hyper plane defined by $w^T x + b = 1$

This distance is equal to 2 This means we want to solve max 2 Equivalently we want min ‖w‖| The SVM should ‖w‖. ‖w‖.2. also correctly classify all x(i), which means

yi(wT xi + b) >= 1, ∀i ∈ {1, ¢¢, N}.

B. *Decision Tree*

With the help of prior data we are predicting the target class using decision rule, this is the main motive using Decision Tree in this research work. For the purpose of classification and prediction nodes and internodes are used. Features or attributes that can be easily differentiated between classes are considered as the root node. Leaf nodes represent classification, while root nodes can have two or more branches where the features are further segregated.

C. *Random Forest*

Random forest is made up from Decision Trees where by means of Voting Classifier we choose the best one, it reduces the over fitting by averaging the result. Bootstrap Dataset i.e. Random samples are selected for the purpose of classification duplicate samples are considered here. Create decision tree via bootstrap dataset and segregate the classes. By means of Voting Classifier we choose the best solution. Bagging is the resultant of Bootstrap and Aggregation. Data which has not been considered for the purpose of classification is called “Out of Bag Dataset”. It is an ensemble method better than single Decision Tree because it reduces the over fitting by averaging result.

With the help of 9 diagnostic measures present in Pima Indian Diabetes Dataset we can predict whether the patient is suffering from diabetes or not. The dataset has 768 records which has been selected from Kaggle. Pre-processing of data is done. Dataset is an excel sheet where 8 to 9 parameters are considered of each and every patient. If there is a certain block or cell which is empty then this particular cell is replaced by the average value of that particular diagnostic parameter, via such techniques the accuracy rates are enhanced for our model.

Machine recognizes the pattern, structure in the data. For better accuracy and efficiency of approaches and algorithms we go for cross validation of data via training data. To make predictions of unknown, unseen data testing is used. How well the model performs based on new data is testing. In our model

80% training and 20 % testing of data is done. Better the training and testing model, results are more accurate and precise.

Support Vector Machine, Decision Tree and Random Forest are the approaches that have been considered for our model. Based on performance measures such as f1, recall, precision and support the accuracy of the above mentioned algorithms can be known, the one which outperforms other two is selected for prediction.

4. Design and Implementation

768 records consisting of 9 attributes determine whether the person suffers from this deadly disease or not. Based on diagnostic measurements, the aim of Pima Diabetes Dataset is to forecast whether the patient is suffering from this disease or not. This dataset is originally from the National Institute of Diabetes and Digestive Kidney Diseases.

For the purpose of Training set we have considered 70% of data while for Testing 30% is considered. Training dataset is already tested, the machine recognises the similar structures and cross validations is also done for the purpose of accuracy and efficiency of these approaches.

Testing is done for unseen, data. In Testing we observe how well the model performs based on new data

available. Performance metrics such as f1-score, precision, recall, support are used.

ROC graphs deal with Sensitivity and Specificity. Receiver Operating Graphs are useful for consolidating the information from a ton of confusion matrices into a single easy to interpret graph. The y-axis shows True Positive Rate, which is the same thing as Sensitivity.

True Positive Rate = True Positive / True Positive + False Negative

The x-axis shows the False Positive Rate, which is the same thing as “1-Specificity”

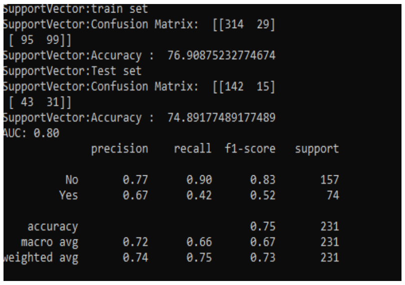
False Positive Rate = (1- Specificity) = False Positive / False Positive + True Negative

ROC curve is created by plotting true positive rate against false positive rate at threshold setting.

A. Support Vector Machine

The evaluated performance of SVM algorithm for prediction of Diabetes using Confusion Matrix is as follows:

The snapshot gives us a clear picture of performance metrics and the accuracy of Support Vector Machine approach.



```
SupportVector:train set
SupportVector:Confusion Matrix:  [[314  29]
 [ 95  99]]
SupportVector:Accuracy :  76.90875232774674
SupportVector:Test set
SupportVector:Confusion Matrix:  [[142  15]
 [ 43  31]]
SupportVector:Accuracy :  74.89177489177489
AUC: 0.80
              precision    recall  f1-score   support

          No       0.77      0.90      0.83       157
         Yes       0.67      0.42      0.52        74

    accuracy                           0.75       231
   macro avg       0.72      0.66      0.67       231
weighted avg       0.74      0.75      0.73       231
```

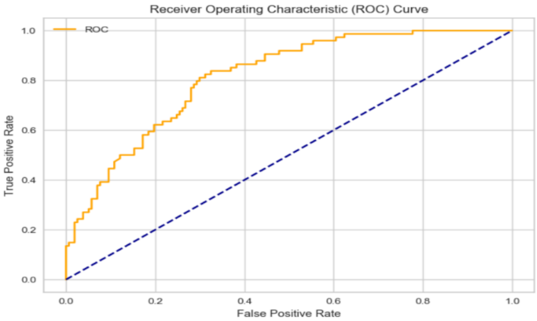


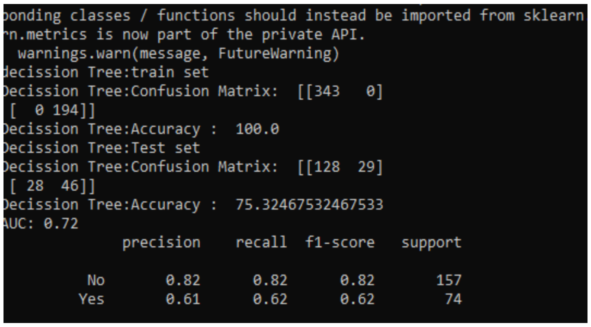
Fig. 2. ROC Curve of SVM

B. Decision Tree

The evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

	Diabetes	Non-diabetes
Diabetes	True positive (256)	False positive (116)
Non-diabetes	False negative (150)	True negative (246)

The snapshot gives us a clear picture of performance metrics and the accuracy of Decision Tree algorithm



```
ponding classes / functions should instead be imported from sklearn
rn.metrics is now part of the private API.
  warnings.warn(message, FutureWarning)
decission Tree:train set
Decission Tree:Confusion Matrix:  [[343   0]
 [  0 194]]
Decission Tree:Accuracy :  100.0
Decission Tree:Test set
Decission Tree:Confusion Matrix:  [[128  29]
 [ 28  46]]
Decission Tree:Accuracy :  75.32467532467533
AUC: 0.72
              precision    recall  f1-score   support

          No       0.82      0.82      0.82       157
         Yes       0.61      0.62      0.62        74
```

	Diabetes	Non-diabetes
Diabetes	True positive (236)	False positive (176)
Non-diabetes	False negative (160)	True negative (196)

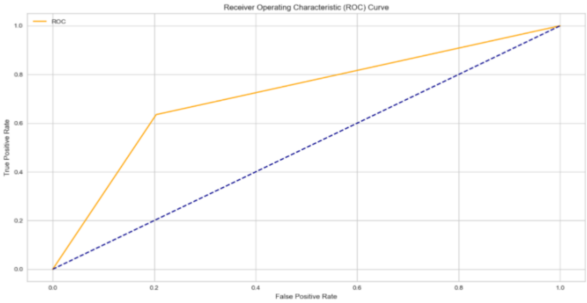


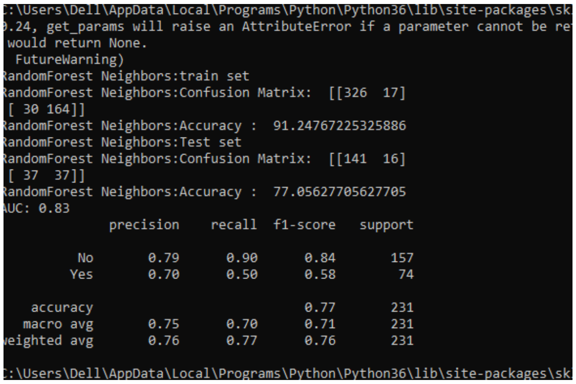
Fig. 3. ROC Curve of Decision Tree

C. Random Forest

The evaluated performance of Random Forest technique using Confusion Matrix is as follows:

	Diabetes	Non-diabetes
Diabetes	True positive (305)	False positive (105)
Non-diabetes	False negative (148)	True negative (210)

The snapshot gives us a clear picture of performance metrics and the accuracy of Random Forest algorithm



```
:\Users\Dell\AppData\Local\Programs\Python\Python36\lib\site-packages\sk
.24, get_params will raise an AttributeError if a parameter cannot be re
 would return None.
  FutureWarning)
andomForest Neighbors:train set
andomForest Neighbors:Confusion Matrix:  [[326  17]
 [ 30 164]]
andomForest Neighbors:Accuracy :  91.24767225325886
andomForest Neighbors:Test set
andomForest Neighbors:Confusion Matrix:  [[141  16]
 [ 37  37]]
andomForest Neighbors:Accuracy :  77.05627705627705
UC: 0.83
             precision    recall  f1-score   support

          No       0.79      0.90      0.84       157
         Yes       0.70      0.50      0.58        74

    accuracy                           0.77       231
   macro avg       0.75      0.70      0.71       231
eighted avg       0.76      0.77      0.76       231

:\Users\Dell\AppData\Local\Programs\Python\Python36\lib\site-packages\sk
```

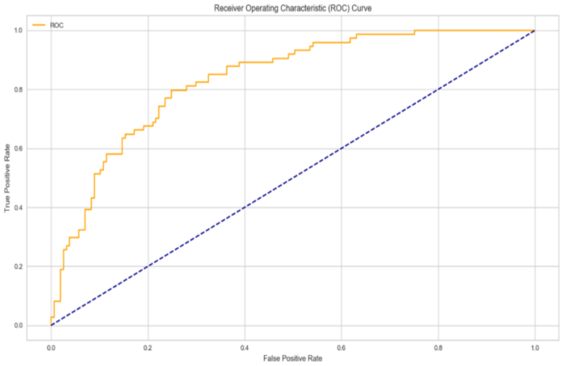


Fig. 4. ROC Curve of Random Forest

Following picture depicts the average accuracy rate varies with training dataset size. We have experimented with different size of training data for SVM, Decision Tree, Random Forest.

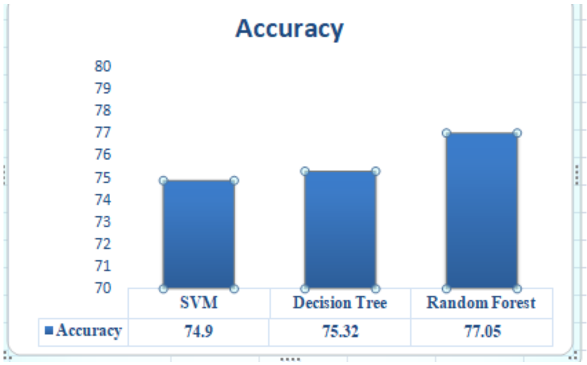


Fig. 5. Accuracy

So from the results we come to know that Random Forest outperforms other two with an accuracy rate of 77.05 whereas Decision Tree has an accuracy of 75.32 whereas Support Vector Machine has an accuracy of 74.9.ROC and AUC are based on logistic regression.

The table above gives us a brief description of weighted average value of performance metrics used in Decision Tree, SVM, Random Forest.

So the above table and graph gives us a brief scenario of various performance metrics of different algorithms used.

Algorithm	Precision	Recall	F-Score	AUC
SVM	0.74	0.75	0.73	0.8
Decision Tree	0.75	0.75	0.75	0.72
Random Forest	0.76	0.77	0.76	0.83

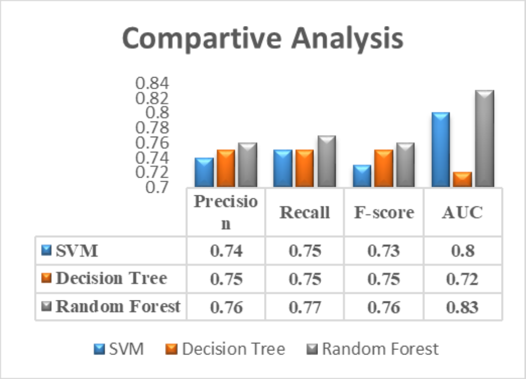


Fig. 6. Comparative analysis

5. Conclusion

In this paper, from different machine learning algorithms Random Forest provide us highest accuracy with ROC Method on Indian Pima Dataset. During this work, three machine learning classification algorithms are studied and evaluated on various measures such as Precision, Recall, f1 and support which is purely based upon confusion matrix. Random Forest outperforms Decision Tree and Support Vector Machine with an accuracy of 77.065%. By referring Data Mining papers, we understood the concept and we tried to apply these approaches via Machine Learning. With the help of Automation Diabetes Analysis including some machine learning algorithms the work can be extended and can be enhanced further. This research work has significant potential

Acknowledgment

We have taken efforts in building this Paper. However, it would have not been possible without the kind support. A hearty thanks to our Project Guide Mrs. Rohini Patil for constantly guiding us and helping us to make this project efficiently.

References

[1] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining," 2017.

[2] Vrushali Balpande, Rakhi Wajgi, "Prediction and Severity Estimation of Diabetes Using Data Mining Technique," 2017.

[3] Sankaranarayanan S, Pramananda Perumal T, "Diabetic prognosis through Data Mining Methods and Techniques," 2014.

[4] Pima Indians Diabetes Dataset: https://www.kaggle.com/mehdidag/pimaindians/home

[5] Y. Christobel, C. Sivaprakasam, "A new class-wise k nearest neighbor (CKNN) method for the classification of diabetes dataset," 2013.

[6] R. Priyadarshini, N. Dash, and R. Mishra, "A Novel approach to predict diabetes mellitus using modified Extreme learning machine," 2014.