# Prediction of Diabetes using Classification Algorithms

# Final Report

- **Priyank Patel**

## *Index*

## INTRODUCTION

Diabetes is an ailment which influences the intensity of the body in delivering the hormone insulin, which progressively makes the digestion of sugar anomalous and lift the level of glucose inside the blood. In Diabetes an individual for the most part experiences high blood glucose. Escalate thirst, intensify hunger also, frequent pee is some of the side effects made thanks high glucose. Numerous confusions happen if diabetes stays untreated. Some of the extreme inconveniences incorporate diabetic ketoacidosis and nonketotic hyperosmolar trance like state. Diabetes is analyzed as a significant genuine wellbeing matter during which the proportion of sugar substance cannot be controlled. Diabetes is not just experiencing different elements like stature, weight, inherited factor and insulin however the significant explanation considered is sugar focus among all variables. The primary ID is that the main solution for remain away from the inconveniences.

The principal aim of the fundamental objective gathering is to machine learning model that may foresee the probability of Diabetes inside the patients at its beginning phases with most extreme exactness reachable. As anticipating the likelihood of Diabetes fall inside the area of order issues with paired yields, we will use classification algorithms like Decision Tree, Support Vector Machine, and Naive Bayes during this exploration. After the modeling of the algorithms is finished, we may test and assess our outcomes based on its exhibition measures and with acceptable outcomes, would attempt to convey the machine learning model to be utilized in real-life cases.

## BENCHMARK(S)

- Sajida Parveen has designed a classification technique using J48 Decision Tree as the basis for prediction and explains the role of AdaBoost and Bagging ensemble methods used in his research considering the Diabetes risk factors. The outcome achieved in this research proves the AdaBoost ensemble techniques outperform in comparison to other methods.

- Pradhan recommended using Genetic Programming (GP) for the training and testing the prediction model, in which Diabetes data sourced for UCI repository is used. Results obtained from this experiment gives optimal accuracy compared to other techniques with faster classifier generation.

- Orabi has proposed a system of machine learning methods with the purpose of prediction of diabetes in patients at a particular age using Decision Tree techniques.

# METHODS

In this research, total three classification algorithms namely Naive Bayes, SVM and Decision Tree algorithms are utilized. Examinations are performed utilizing inner cross-approval 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures are utilized for the order of this work.

## Support Vector Machine (SVM):

SVM is one among the quality arrangement of supervised machine learning model utilized in classification. Given a two-class preparing test the purpose of a help vector machine is to locate the simplest most noteworthy edge isolating hyperplane between the 2 classes. For better generalization hyperplane ought not to lie nearer to the information focuses have an area with the opposite class. Hyperplane must be chosen which may be a great distance from the data focuses from every classification. The focuses that lie closest to the sting of the classifier are the assistance vectors.

## Decision Tree Classifier:

Decision Tree is a supervised machine learning algorithm used to take care of order issues. The fundamental goal of utilizing Decision Tree in this examination work is the expectation of target class utilizing choice standard taken from earlier information. It utilizes hubs and inter nodes for the forecast and arrangement. Root hubs arrange the occurrences with various highlights. Root hubs can have at least two branches while the leaf hubs speak to grouping. In each stage, Decision tree picks every hub by assessing the most elevated data gain among the properties. The assessed execution of Decision Tree method utilizing Confusion Matrix is as per the following:

This research is performed utilizing inner cross-approval 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures are classification of this work. Below table shows the accuracy measures which is according to our implementation.

## Gaussian Naive Bayes

Naive Bayes can be modified to real-valued attributes, most commonly by using a Gaussian distribution. This naive Bayes modification is called Gaussian Naive Bayes. Certain functions can be used to estimate the data distribution, but the Gaussian (or Normal Distribution) is the simplest to use, since you only need to measure the mean and standard deviation from your training data. One special form of Naive Bayes algorithm is a Gaussian Naive Bayes algorithm and it is generally used when there are continuous values in the functionality. It is often presumed that all the functions follow a gaussian distribution, i.e. normal distribution.
The Gaussian Naive Bayes method does not take parameters while building the model, so we will not use grid search method for this algorithm.

## Random Forest Classifier

Random forest, as its name implies, is made up of many individual decision trees which function as an ensemble. Every single tree in the random forest spits out a class prediction and the class with the most votes is the prediction of our model.
The basic idea behind random forest is a simple but strong one — the wisdom of crowds. Speaking in data science, the reason the random model of forests works so well is:

Many relatively uncorrelated models (trees) operating as a committee will exceed any of the individual models in the constituent.

The secret to this is the weak correlation between models. Much as investments with low correlations combine to create a portfolio greater than the sum of its components, uncorrelated models can generate predictions of the ensemble that are more accurate than any of the individual predictions. The explanation for this magnificent impact is that trees shield each other from their individual mistakes. Although some trees might be wrong, many other trees would be correct, and the trees will move in the right direction as a group.
So, the prerequisites for successful production of random forests are:

1. Our features need to contain some real signal so that models developed using those features perform better than random conjectures.
2. The predictions made by the individual trees (and hence the errors) need to have low correlations with each other.

## XGBoost Classifier

Extreme Gradient Boosting or XGBoost is a gradient boosting algorithm library that is designed for modern data science challenges and tools. It leverages the above techniques with boosting and comes packaged in a library that is easy to use. Some of XGBoost's major advantages are that it is highly scalable / parallelizable, easy to execute, and usually executes other algorithms.

Gradient boosting also involves an ensemble approach that adds predictors sequentially and corrects preceding models. Nevertheless, instead of assigning different weights to the classifiers for each iteration, this approach matches the new model to the previous prediction 's new residuals, and then minimises the loss when adding the latest prediction. And, at the end of the day, one update one's model with gradient descent and hence the term, gradient boost. It is sponsored both for regression and classification problems. Specifically, XGBoost implements this decision tree algorithm boosting in objective function with an additional custom regularisation term.

## Model Diagram

Proposed procedure is summed up in figure-1 underneath in the form of model diagram. The figure shows the pattern of the research conducted in constructing the model.

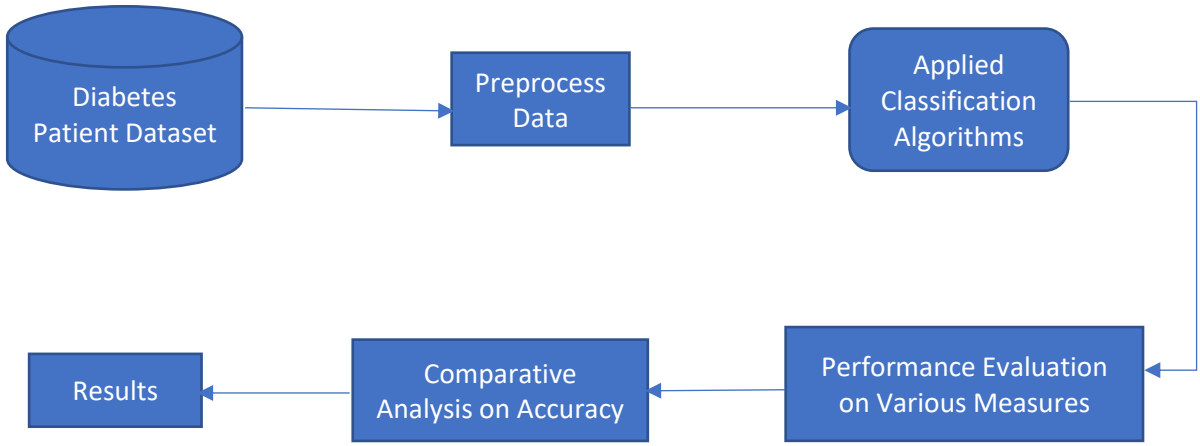


Fig-1: Proposed Model Diagram

## RESULTS

| | Model | Accuracy % | Recall | Precision | F-Measure | ROC |
|---|---|---|---|---|---|---|
| 0 | SVC | 0.82 | 0.62 | 0.76 | 0.68 | 0.77 |
| 1 | Decision Tree | 0.78 | 0.45 | 0.72 | 0.55 | 0.69 |
| 2 | GaussianNB | 0.76 | 0.62 | 0.67 | 0.64 | 0.74 |
| 3 | Random Forest | 0.81 | 0.55 | 0.74 | 0.63 | 0.73 |
| 4 | XGBoost | 0.78 | 0.55 | 0.67 | 0.60 | 0.72 |

Table: Performance Measure

According to implementation table-1 illustrates different performance values of all classification algorithms on all different measures and this implementation describes that Support Vector Machine (SVC) has highest accuracy and highest F-measure. For that, possibility of predicting highest accuracy among all classifier is SVC machine learning classifier.

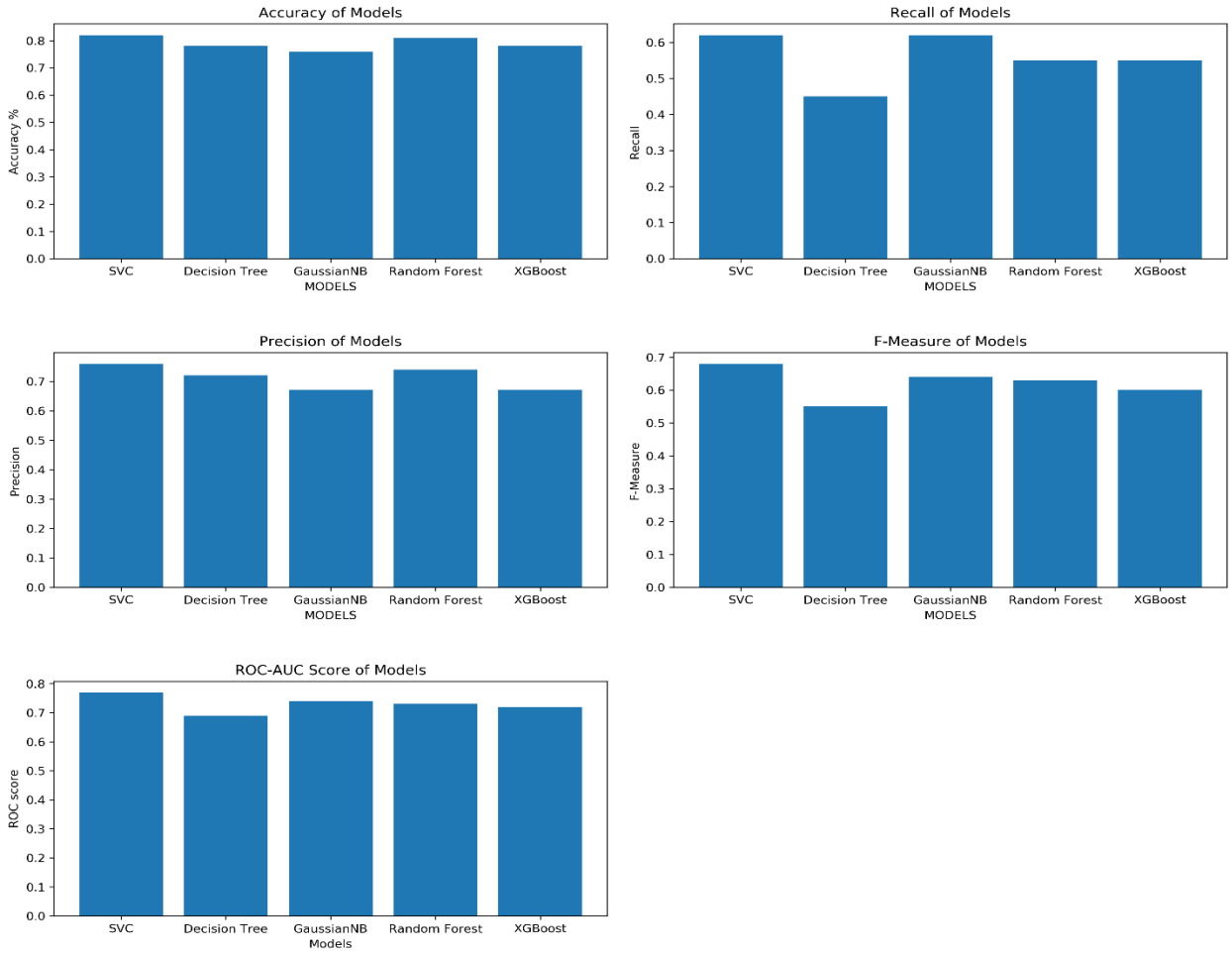


Fig-2: Performance Measures of each Classifiers

Mentioned figures-2 describes more accurately performance of each classifiers based on all measures according to implementation and figures-3 describes Receiver Operating Curve area of all classification algorithms.

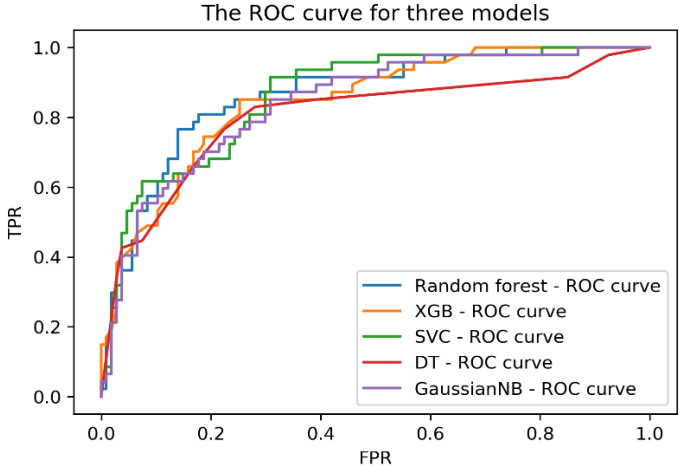


Fig-3: ROC area of all Classifiers

## DISCUSSION

As per the proposal, the acceptable model will be considered having accuracy of higher than 75%, which is the case in all of the machine learning models we have used for classification, in which Support Vector Classifier has performed exceptionally good giving the highest accuracy of 82%. This accuracy can be considered as an indicator that this model can be used in real time prediction of Diabetes using the same given parameters on which the model is trained to predict the outcome. This experiment opens a wide window of opportunity in the domain where more research can be done in the prediction of medical conditions with better accuracies than in the past. This can prove to be significant improvement over the manual prediction of medical conditions, where using machine learning algorithms gives faster and similar or in some case better results than the human decision.

## CONCLUSION

- The Motive of this research is to determine how accurately we can predict cases of Diabetes in a patient in the early stages. For measuring the performance of these models, we used various measures such as Accuracy, Precision, Recall, F-Measures and ROC which one can see in the table above the plots. For this study, we used Pima Indians Diabetes Database for training the models used for classification.
- The Three Classification algorithms used in this study are Naïve Bayes, SVM and Decision tree, from which Support Vector Classifiers (SVC) stands out with highest accuracy in predicting the cases with 82% and 0.76 Precision on test dataset in respective to other data models for this experiment.. This research also signifies that Machine learning algorithms can be used in prediction and diagnose of other medical conditions in its early stages and can prevent later complications.
- Although in our research paper we found that Naive Bayes performed the best out of three but as we implemented these Classifiers, we found that SVC is best giving results among the other Classification Algorithms. Although, the performance measure of other model is somewhat like the selected research paper and in our experiment. This difference can be because of selecting different parameter tuning in each model.
- We conclude during this experiment; we learn three different algorithms and evaluate them on various measures.

## CONTRIBUTIONS

- The portion of Data Cleaning has been combined effort by each member.
- The Exploratory Data Analysis section is also combined effort by each member.
- Working with 3 different classification algorithms, each member has built and evaluated one algorithm.

## REFERENCES


- https://www.sciencedirect.com/science/article/pii/S1877050918308548
- Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. DOI:10.1016/j.procs.2016.04.016
- Pradhan, P.M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V., Dhobale, V., 2012. A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research 2.
- Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in Industrial Conference on Data Mining, Springer. Springer. pp. 420–427
- https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabet s-detection
- https://thatsugarmovement.com/the-importance-of-early-diagnosis-of-pre-diabete s/
- https://www.pritikin.com/your-health/health-benefits/diabetes/1629-diabetes-the-benefits-of-early-action.html
- https://www.medicalnewstoday.com/articles/317074#lifestyle-changes-for-type-2-diabetes
- https://www.nih.gov/news-events/news-releases/youth-type-2-diabetes-develop-complications-more-often-type-1-peers
- Dataset:https://www.kaggle.com/uciml/pima-indians-diabetes-database?select=diabetes.csv
- Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 ,476–491. doi:10.1109/34.589207


## APPENDICES

1. **Final_Project.ipynb:** We have used juypter notebook for python programming, where data is imported and data transformation is performed and the machine learning models are defined in the notebook.