**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

# Implementing a Model to Detect Diabetes Prediction using Machine Learning Classifiers

**[1]P J Sireesha, [2]Mr. K Prakash [3]Dr. D Sumathi**

Department of Computer Science and Engineering, Kuppam Engineering College, Kuppam, India

Abstract: Diabetics is a metabolic disease caused because of high glucose level in a body. It may cause many complications like kidney damage, heart related problems, eye problem, blood pressure and it can affect other organs of human body. Now a days there is an increase in number of people suffering from diabetics. Diabetics can be treated if it is predicted earliest stage. In this approach various machine learning techniques can be used from predicting diabetics with higher accuracy. To achieve the goal, we will use different ML and ensemble techniques which are K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Ada Boost, Naive Bayie and XG Boost. Our Result shows that XG Boost achieved higher accuracy compared to other machine learning techniques.

Index Terms - Diabetes, Machine, Learning, Prediction, Dataset, Ensemble

## I. INTRODUCTION

Diabetics is a chronic long-lasting disease that affects our human body. If you have a diabetes your body either doesn't make enough insulin or can't use the insulin it makes as well as it should and it can be caused because of obesity, high glucose level where patients suffer from blood glucose problems due to abnormal production and release of insulin. As per (WHO) World Health Organisation about 422 million people suffering from diabetes. Without any resources and appropriate models for early prediction, people with diabetics may only be diagnosed after the onset of complications. Early prediction of diabetes can be controlled and save the human life. To achieve this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose, we use the Pima Indian Diabetes Dataset and we apply various ML techniques. Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus, for this purpose we apply popular classification and ensemble methods on dataset for prediction.

The organizational framework of this study divides the research work in the different sections. The literature review is presented in section 2. Further, in section 3, Modules discussed. Moreover, in next section IV, briefly explain about Data and data set and data processing and in section V, the description of Modules building in different steps mentioned. In section VI, experimental work after implementing machine learning classifiers. And finally, the experimental results discussed in section VII. Conclusion and future work are presented by last sections VIII.

## II. LITERATURE SURVEY

K.VijiyaKumar et al. [11] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Nonso Nnamoko et al. [13] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method,

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

diabetes onset prediction can be done with higher accuracy.

Tejas N. Joshi et al. [12] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease.

Deeraj Shetty et al. [14] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

## III. MODULES

**Module Description:**

- Data Pre-Processing
- Feature Selection
- Classification Modeling
- Performance Measures

***Dataset Description:***

The objective of the dataset is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

*Pregnancies:* Number of times pregnant

*Glucose:* Plasma glucose concentration 2 hours in an oral glucose tolerance test

*Blood Pressure:* Diastolic blood pressure (mm Hg)

*Skin Thickness:* Triceps skin fold thickness (mm)

*Insulin:* 2-Hour serum insulin (mu U/ml)

*BMI:* Body mass index (weight in kg/(height in m)^2)

Diabetes Pedigree Function: Diabetes pedigree function

*Age:* Age (years)

*Outcome*: Class variable (0 or 1) 268 of 768 are 1, the others are 0

***Data Pre-Processing***

Diabetic disease data is pre-processed after collection of various records. The dataset contains a total of 769 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 763 patient records are used in pre-processing.

*Preprocessing Selection*

| Process Selector | Preprocessing |
|---|---|
| P | Outlier Rejection |
| Q | Filling Missing Value |
| R | Standardization |

Each P, Q, and R process has four techniques for feature selection which are N/A, PCA, ICA, and correlation-based feature selection

***Feature Selection***

From among the 8 attributes of the data set, one attributes pertaining to age is used to identify the personal information of the patient. The remaining 7 attributes are considered important as they contain vital clinical records.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

Clinical records are vital to diagnosis and learning the severity of diabetes disease.

***Classification Modeling***

The clustering of data sets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error.

- Decision Trees Classifier
- Support Vector Classifier
- Random Forest Classifier
- Logistic Regression
- K Nearest neighbors
- Naive Bayes

***Performance Measures:***

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model.

## IV. PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase. A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

**Table 1: Dataset Description**

| S.N | Attributes | Comments |
|---|---|---|
| 1 | Pregnancies | Number of times pregnant |
| 2 | Glucose | Plasma glucose concentration 2 hours in an oral glucose tolerance test |
| 3 | Blood Pressure | Diastolic blood pressure (mm Hg) |
| 4 | Skin thickness | Triceps skin fold thickness (mm) |
| 5 | Insulin | 2-Hour serum insulin (mu U/ml) |
| 6 | BMI | Body mass index (weight in kg/(height in m)^2) |
| 7 | Diabetes Pedigree Function | Diabetes pedigree function |
| 8 | Age | Age (years) |

### A. DISTRIBUTION OF DIABETIC PATIENT

We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic.

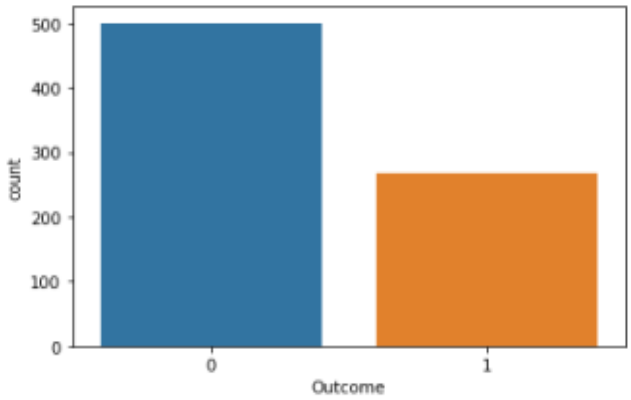


**Figure 1: Ratio of Diabetic and Non Diabetic Patient**

### B. DATA PRE-PROCESSING

Data pre-processing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data pre-processing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset, we need to perform pre-processing in two steps.

**1) Missing Values removal-**

Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore, this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

**2) Splitting of data-**

After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, aim of normalization is to bring all the attributes under same scale.

**C. APPLY MACHINE LEARNING**

When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

The Techniques are follows

**1) K-Nearest Neighbor -**

KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique.KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other.KNN helps to group new work based on similarity measure.KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set it's nearest neighbours. Here K= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance.

The Euclidean distance between two points P and Q i.e. P (p1, p2, …Pn) and Q (q1, q2 ,.. qn) is defined by the following equation:-

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

**Algorithm-**

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Then, Decide a random value of K. is the no. of nearest neighbours
- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.

If the values are same, then the patient is diabetic, otherwise not.

**2) Decision Tree-**

Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc. Steps for Decision Tree.

**Algorithm-**

- Construct tree with nodes as input feature.
- Select feature to predict the output from input feature whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a subtree using the feature which is not used in above node.

**3) Naive Boost-**

Naive Boost is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Naive Boost used when we want to classify or distinguish some data items into categories. It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of Naive Boost is to best fit which is responsible for describing the relationship between target and predictor variable. Naive Boost is a based on Linear regression

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

model. Naive Boost model uses sigmoid function to predict probability of positive and negative class.

**4) Ensembling-**

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as – Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

**5) Random Forest –**

It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

**Algorithm-**

- The first step is to select the "R" features from the total features "m" where R<<M.
- Among the "R" features, the node using the best split point.
- Split the node into sub nodes using the best split. • Repeat a to c steps until "l" number of nodes has been reached.
- Built forest by repeating steps a to d for "a" number of times to create "n" number of trees.

**6) XG Boosting –**

XG Boosting is most powerfull ensemble technique used for prediction and it is a classification technique. It combine week learner together to make strong learner models for prediction. It uses Decision Tree model. It classifies complex data sets and it is very effective and popular method. In gradient boosting model performance improve over iterations.

**Algorithm-**

- Consider a sample of target values as P.
- Estimate the error in target values.
- Update and adjust the weights to reduce error M.
- P[x] =p[x] +alpha M[x]
- Model Learners are analyzed and calculated by loss function F
- Repeat steps till desired & target result P.

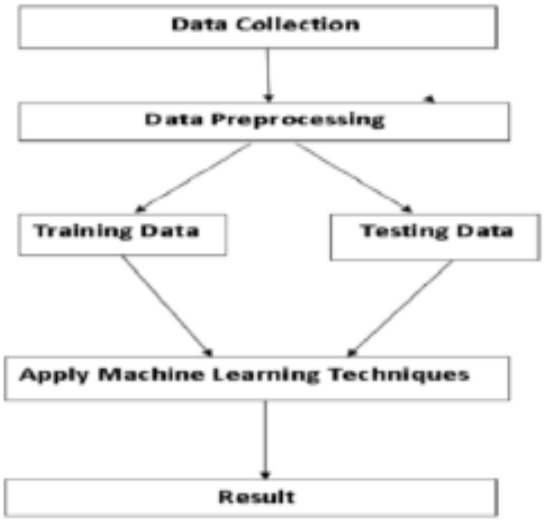


**Figure 2: Overview of the Process**

## V. MODULE BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction. Procedure of Proposed Methodology

**Step1:** Import required libraries, Import diabetes dataset.
**Step2:** Pre-process data to remove missing data.

**Step3:** Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

**Step5:** Build the classifier model for the mentioned machine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier

**Step8:** After analyzing based on various measures conclude the best performing algorithm

## VI. ANALYSIS

We get the ROC curve after implementing the classifier. Please see Fig.3 to Fig. 8 for the reference. The ROC curves for KNN, Decision Tree, Random Forest, ada boost, Naive Bayes, XG Boost, classifiers.

### KNN Classifier

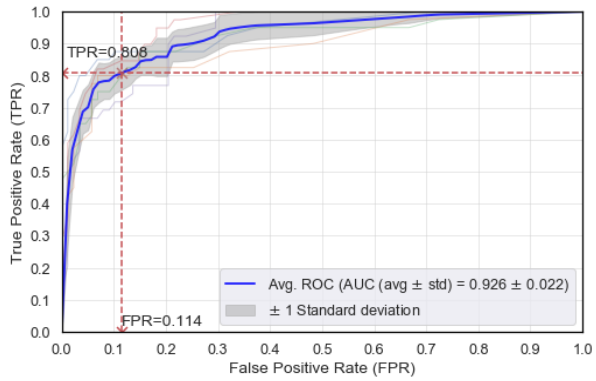


**Figure 3: ROC curve: KNN**

### Decision Trees Classifier

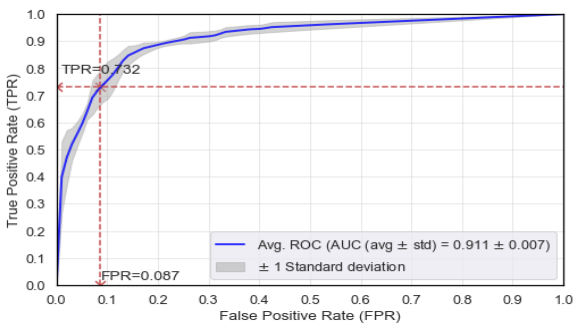


**Figure 4: ROC curve: Decision Tree**

### Random Forest Classifier

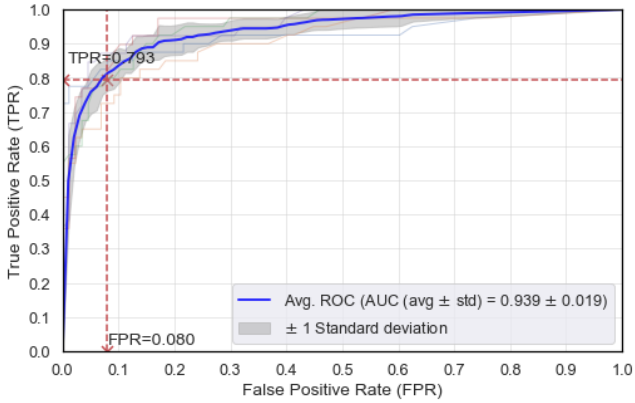


**Figure 5: ROC curve: Random Forest**

### AdaBoost Classifier

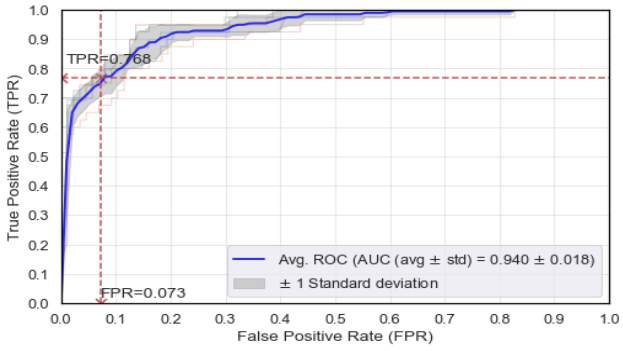


**Figure 6: ROC curve: ADABoost**

### Naive Bayes Model

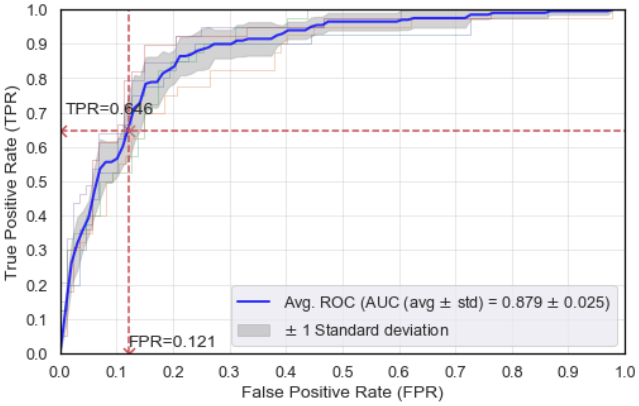


**Figure 7: ROC curve: Naive Bayes**

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 1, 2022, p. 558-566
https://publishoa.com
ISSN: 1309-3452

**XGBoost Model**

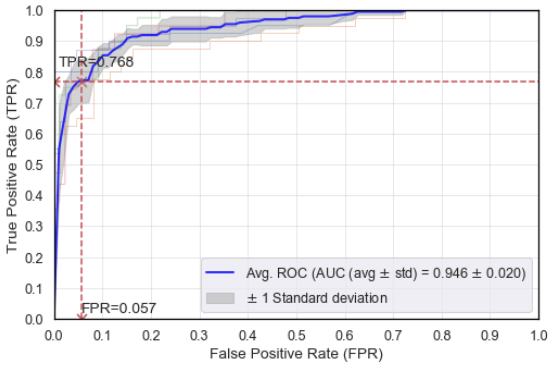


**Figure 8: ROC curve: XGBoost**

## VII. EXPERIMENTAL RESULTS

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy. Figure shows the result of these Machine Learning methods.

As discussed in the earlier sections, we have used five different classifiers to predict and improve the accuracy of the disease diabetes, Comparisons of these classifiers have been shown below in the accuracy table.

**TABLE 2: Showing AUC and Accuracy for different classifiers**

| MODEL | AUC | ACCURACY |
|---|---|---|
| Decision Trees Classifier | 0.911 | 0.857 |
| KNN Classifier | 0.926 | 0.862 |
| Random Forest Classifier | 0.939 | 0.881 |
| Ada Boost Classifier | 0.940 | 0.877 |
| Naive Bayes Classifier | 0.879 | 0.807 |
| XG Boost Classifier | 0.946 | 0.888 |

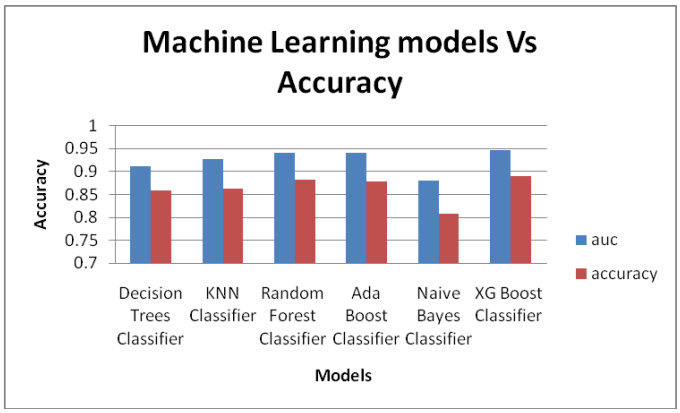


**Figure 9: Machine Learning models Vs Accuracy and AUC**

We have found that our system provides us with 85.2 % of accuracy for Decision Tree Classifier, 88.8% accuracy for XGBoost Classifier, 86.2% accuracy for KNN Classifier, 88.1% accuracy for Random Forest Classifier, 87.7% accuracy for Ada Boost Classifier, 80.7% accuracy for Naive Bayes Classifier. We have therefore found that the best among all the above classifiers is XG Boost Classifier.

## VIII. CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

***Future work***

In future, if we get a large set of diabetic dataset we can perform comparative analysis for analyzing the performance of each algorithm as well as the Hybrid algorithm so that the best one can be applied for predictive analysis. A particular method to identify diabetes is not very sophisticated way for initial diabetes detection and it is not fully accurate for predicting diseases. That's why we need a smart hybrid predictive analytics diabetes diagnostic system that can effectively work with accuracy and efficiency. We can use data mining , neural network for exploring and utilizing to support medical decision, which improves in diagnosing the risk for pregnant

diabetes. Due to the dataset we have till date are not upto the mark , we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and explore it, which may improve the accuracy of predicting diabetes. We can also study the causes of diabetes and how to avoid having diabetes

## REFERENCES

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
2. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
3. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
4. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
5. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.
6. A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
7. T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," IEEE Transactions on Electronic Computers, vol. 14, no. 3, pp. 326-334, Jun. 1965.
8. G. I. Webb, J. R. Boughton, and Zhihai Wang, "Not So Naive Bayes: Aggregating one-dependence estimators," Machine learning, vol. 58, no. 1, pp. 5-24, Jan. 2005.
9. S. B. Belhouari and A. Bermak, "Gaussian process for nonstationary time series prediction," Computational Statistics & Data Analysis, vol. 47, no. 4, pp. 705-712, Feb. 2004.
10. C. Cortes and V. Vapnik , "Support-vector networks," Machine Learning, vol. 20, pp. 237-297, Sep. 1995.
11. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.
12. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13
13. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
14. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.
15. D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578-1585, Jan. 2018.
16. S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115-121, Mar. 2016.
17. M. Pradhan and G. R. Bamnote, "Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming," in Proc. third International Conference on Frontiers of Intelligent Computing: Theory and Applications, Nov. 2015, pp. 763-770.
18. N. Nai-arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132-142, Dec. 2015.
19. M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," Computer Methods and Programs in Biomedicine, vol. 152, pp. 23-34, Dec. 2017.
20. R. Bansal, N. Gaur, and S. N. Singh, "Outlier Detection: Applications and techniques in data mining," in Proc. sixth International Conference-Cloud System and Big Data Engineering, Jan. 2016, pp. 373-377.
21. D. Cousineau and S. Chartier, "Outliers detection and treatment: A review," International Journal of

Psychological Research, vol. 3, no. 1, pp. 58-67, Mar. 2010.

22. C. R. Rao, "The use and interpretation of principal component analysis in applied research," Sankhya: The Indian Journal of Statistics, Series A, pp. ¯ 329-358, Dec. 1964.
23. A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and applications," Neural Networks, vol. 13, no. 4-5, pp. 411-430, Jun. 2000.
24. F. Han and H. Liu, "Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution," Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability, vol. 23, no. 1, pp. 23-57, Feb. 2017.
25. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics surveys, vol. 4, pp. 40-79, Jul. 2010.
26. D. Krstajic, L. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," Journal of Cheminformatics, vol. 6, no. 1, pp. 10, Mar. 2014.
27. X. Zeng and T. R. Martinez, "Distribution-balanced stratified crossvalidation for accuracy estimation," Journal of Experimental & Theoretical Artificial Intelligence, vol. 12, no. 1, pp. 1-12, Nov. 2000.
28. P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers," Multiple Classifier Systems, vol. 34, no. 8, pp. 1-17, Mar. 2007.
29. T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785-794.
30. S. Hsieh, S. Hsieh, P. Cheng, C. Chen, K. Hsu, I. Wang, and F. Lai, "Design ensemble machine learning model for breast cancer diagnosis," Journal of Medical Systems, vol. 36, no. 2012, pp. 2841-2847, Jul. 2011.
31. B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," Journal of Biomedical Informatics, vol. 86, pp. 25-32, Oct. 2018.
32. A. S. Miller, B. H. Blott, and others, "Review of neural network applications in medical imaging and signal processing," Medical and Biological Engineering and Computing, vol. 30, no. 5, pp. 449-464, Jan. 1992.
33. D. E. Rumelhart, G. E. Hinton, and J. Ronald, "Review of neural network applications in medical imaging and signal processing," Nature, vol. 323, no. 6088, pp. 533-536, Oct. 1986.
34. A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The Diagnostic Odds Ratio: A single indicator of test performance," Journal of Clinical Epidemiology, vol. 56, no. 11, pp. 1129-1135, Nov. 2003.
35. P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv:1710.05941, Oct. 2017.
36. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929-1958, Jan. 2014