ISSN No. 0976-5697

VOLUME 12 SPECIAL ISSUE 2, JUNE 2021

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# DIABETES DISEASE PREDICTION USING MACHINE LEARNING ENSEMBLE METHOD

Yerasi Jayasimha Reddy
School of Computer Science and Engineering
REVA UNIVERSITY
Bengaluru, India
jayasimha1357@gmail.com

Yerasi Kedarnath Reddy
School of Computer Science and Engineering
REVA UNIVERSITY
Bengaluru, India
kedarnathreddy07@gmail.com

Rahul David
School of Computer Science and Engineering
REVA UNIVERSITY
Bengaluru, India
rahuldavid026@gmail.com

Rahul Vasishta
School of Computer Science and Engineering
REVA UNIVERSITY
Bengaluru, India
hsrahulv3@gmail.com

Anilkumar Ambore
School of Computer Science and Engineering
REVA UNIVERSITY
Bengaluru, India
anil.ambore@reva.edu.in

***Abstract:*** Machine learning incorporates AI, and is used to solve many problems in data science. The machine reads patterns from existing databases, and then inserts them into an unknown database to predict the outcome. Classification can be a powerful machine learning method commonly used for prediction. Some classification algorithms provide satisfactory accuracy, while others provide restricted accuracy. This paper examines a method called ensemble classification, which is often used to improve the accuracy of weak algorithms by combining multiple categories. Tests for this tool are performed using a diabetic database. A comparative analytical approach was performed to find out how the ensemble process is often used to improve diabetes prognosis. The goal of this paper is not only to increase the accuracy of weak classification algorithms, but also to implement an algorithm on a medical database, to demonstrate its ability to detect the disease at an early age. The results of the study indicate that integrated strategies, such as the random forest, are effective in increasing the predictive accuracy of weak classifiers, and have shown satisfactory effectiveness in identifying the risks of diabetes. A seven-point increase in the accuracy of the weak classifiers was achieved with the help of an ensemble classification.

***Keywords:*** Machine Learning; Classification; Random Forest; Ensemble Classification; Weak Classifiers

## I. INTRODUCTION

Diabetes is a chronic disease that has the potential to cause health care problems worldwide. According to the International Diabetes Federation 382 million people are living with diabetes worldwide. By 2035, this could double as $ 592 million. DM can be a disease caused by high blood sugar levels. A variety of traditional methods, both physical and chemical tests, are available to diagnose diabetes However, early diagnosis of diabetes is a difficult task for medical professionals due to severe dependence on a variety of factors as diabetes affects human organs such as kidneys, eye, heart, nerves, foot. By shedding new light on common questions. An important function is to assist and make predictions on medical information. The ultimate goal of this project is to develop a system that can enable the early diagnosis of diabetes in a patient with better accuracy by combining the results of various ML methods. Therefore, this project uses one among the study methods of ensemble classification called random forest to predict the risk of diabetes in hazardous substances. It also attempts to increase the accuracy of predicting the risk of diabetes using a strategy called ensemble.

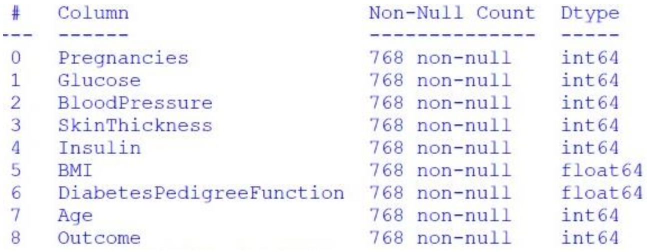
## II. LITERATURE SURVEY

In paper [1] the author describes how prediction is done on Big Data Healthcare. Here they emphasized on 4 important machine learning algorithms. To analyze they used WEKA tool. According to paper [2] A Simplified Indian Diabetes Risk Score is for screening the Undiagnosed Diabetic Subjects. They used simplified versions and applied in 3 phases. For plotting results they used regression analysis. [3] Author used Hybrid Machine Learning Technique they suggested it that its accuracy is 81.33% better than SVM, ANN, KNN. [4] Depicts the comparative analysis of machine and deep learning algorithms. [5] Here also they analyzed machine learning algorithms and plotting WEKA tool is used.

## III. METHODOLOGY

### A. *Data Source*

The data is gathered from UCI repository which is known as Pima Indian Diabetes Dataset. The dataset has many attributes of 768 patients.



```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Pregnancies                768 non-null    int64
 1   Glucose                    768 non-null    int64
 2   BloodPressure              768 non-null    int64
 3   SkinThickness              768 non-null    int64
 4   Insulin                    768 non-null    int64
 5   BMI                        768 non-null    float64
 6   DiabetesPedigreeFunction   768 non-null    float64
 7   Age                        768 non-null    int64
 8   Outcome                    768 non-null    int64
```

### B. *Class Imbalance Problem*

SMOTE: Synthetic Minority Oversampling Technique

SMOTE is an oversampling method where the synthetic selected samples are produced for the minority class. This algorithm is used to overcome the over fitting issue posed by random oversampling.

### C. *Proposed System*

To develop a web application where:
For diabetes prediction:

1. The user enters the data manually with attributes such as their Age, insulin level, BMI, and etc.

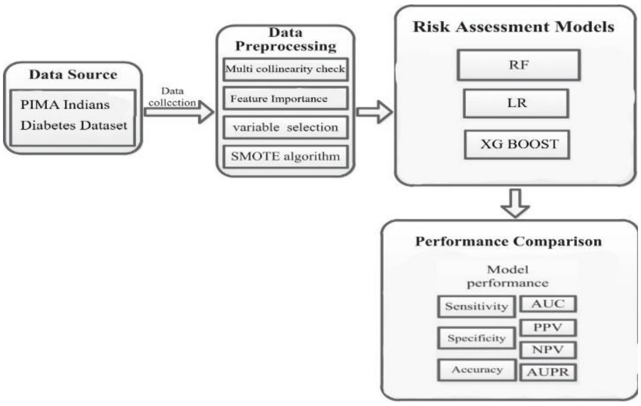
2. The data collected from user through the application is further processed using the machine learning model.

3. The prediction for the individual instance of data is done.

4. The predicted result is displayed on the web application.

For risk calculation of diabetes:

1. The user enters the 4 parameters – person physical activity, person's family history, age and abdominal obesity.
2. 
3. The data collected from the user through the application is used by the python program to calculate the risk of diabetes.



4. The calculated score is displayed on the web application

Fig: Proposed Architecture

The following figure depicts the approach that has been applied to perform the comparative analysis in order to recommend the best algorithm for building classification model in order to predict the diabetes disease.

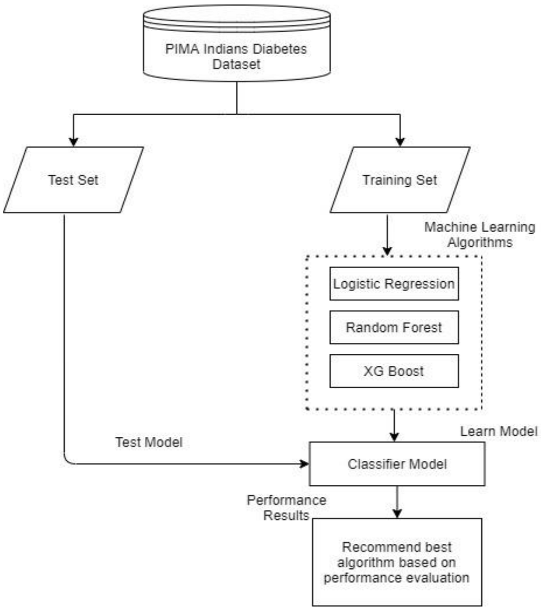


Fig: Proposed Methodology Flowchart

Stepwise Procedure of Proposed Methodology

• Step 1 – Perform data preprocessing on the input

dataset of diabetes disease.

• Step 2 - Split the dataset by 80% to divide dataset as Training set and Test set

• Step 3 - Select the machine learning algorithm i.e. Random Forest, Logistic Regression and XGBoost.

• Step 4 - Build a classification model from the selected machine learning algorithm based on training set.

• Step 5 - Test the classifier model for the selected machine learning algorithm based on test set

• Step 6 - Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

• Step 7 - After analyzing based on various measure conclude the best performing algorithm.

### D. *Description of Algorithms*

Ensemble methods are the classifier combination methods that combine multiple machine learning techniques into one powerful predictive model.

The two categories of ensemble methods used are:

1. Bagging

Bagging method creates the classifiers parallelly. The base learners in bagging are independent of each other and results in the decrease of variance. It bootstraps the subsamples and then aggregates the results over the weak learners, hence the name Bootstrap Aggregation.

a. Random Forest

Random Forest is a bagging ensemble technique that utilizes decision tree as the base learner. The sample of records is created using row sampling with replacement and feature sampling. Each of the decision trees is provided with the set of records for training. For a given test data, majority vote is performed on the predicted results and the cardiovascular disease prediction is done.

2. Logistic Regression

Logistic regression (also known as logit regression) is a statistical representation that in its basic form uses a logistic function to model a binary dependent variable.

3. Boosting

In boosting technique, the base learners are created sequentially and combined to form a strong learner to increase the accuracy. The objective here is to train the weak learners successively, each trying to rectify the formerly misclassified records.

a. Extreme Gradient Boosting

XGBoost provides the gradient boosting framework that braces many languages like C++, R, Python, and Java. The proposed work is developed using python programming. XGBoost is a good candidate for cloud integration. The pros of using xgoost are cache optimization, speed, portability, regularization, and auto pruning

### E. *Model Performance Measures*

1. Confusion matrix:

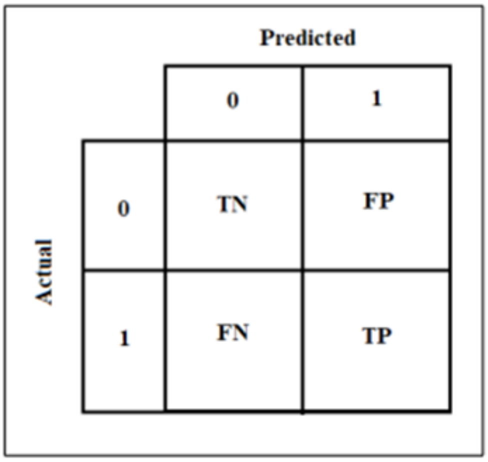
For binary classification, it is a 2 x 2 matrix. It describes four important properties of any model. These properties are used as input to various performance measures.

a. True Positive (TP): The number of instances correctly predicted as a patient.

b. False Positive (FP): The number of instances incorrectly predicted as a patient.

c. True Negative (TN): The number of instances correctly predicted as healthy.

d. False Negative (FN): The number of instances incorrectly predicted as healthy.



| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | TN | FP |
| | 1 | FN | TP |

2. Performance Measures:

a. Accuracy: The proportion of correct predictions form an overall number of predictions. Mathematically, this can be stated as:

Accuracy = (T P + T N)/ (T P + F N + T N + F P)

b. Sensitivity or Recall: The proportion of correct positive classifications from instances that are actually positive. Mathematically, this can be stated as:

Sensitivity = T P/ (T P + F N)

c. Specificity: The proportion of correct negative classifications from instances that are actually negative. Mathematically, this can be stated as:

Specificity = T N/ (T N + F P)

## IV. RESULT & DISCUSSION

This project intends to design and implement a web application for the prediction of diabetes disease, Risk factor for obtaining diabetes for undiagnosed diabetic patients and to analyze the effective performance of the applied ensemble techniques. Random Forest attained an accuracy of 90 percent. By solving the class imbalance problem, the prediction accuracy of the model was increased. The accuracy has been computed using the Confusion Matrix Metric and ROC Curve Metric.

Table 1 depicts the ensemble classification methods that are applied for diabetes disease prediction and their corresponding accuracies. It is inferred that the efficiency of random forest is high when compared to decision trees.

TABLE 1: Comparison of classification techniques

| **ALGORITHM NAME** | **ACCURACY** |
|---|---|
| Random Forest | 0.9065161868651518 |
| Logistic Regression | 0.8168451358615254 |
| XGBoost | 0.8495852868852459 |

## V. CONCLUSION

In the proposed work, a web application is developed to predict the threat of diabetes disease and calculate the risk factor for diabetes for undiagnosed diabetic patients. It is perceived that the bagging algorithm, random forest renders highest accuracy when contrasted to other algorithms.

## VI. FUTURE WORK

Various web and mobile applications can be developed to diagnose other chronic diseases like cancer, cardiovascular, arthritis, etc. The work can be blended with IoT for real-time prediction. The diabetes disease prediction can be done using artificial neural networks or through machine learning cloud platforms for improved accuracy.

## VII. REFERENCES

[1] Ayman Mir, Sudhir N. Dhage. (2018).Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare. Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm in WEKA to predict Diabetes. Random Forest turns out to be an accuracy of 78% over Naive Bayes, SVM and Simple CART.

[2] V Mohan, R Deepa, M Deepa, S Somannavar, M Datta (2015).A Simplified Indian Diabetes Risk Score for Screening for Undiagnosed Diabetic Subjects. The Indian Diabetes Risk Score is developed based on results of many logistic regression analysis. Internal validation is performed on the identical data. IDRS has mainly four risk factors - abdominal obesity, family history of diabetes, age and physical activity.

[3] Rajawat, P. S., Gupta, D. K., Rathore, S. S., & Singh, A. (2018). Predictive Analysis of Medical Data using a Hybrid Machine Learning Technique. Hybrid Machine learning approach to predict if a person is in risk of diabetes. Hybrid Technique turns out with an accuracy of 87.33% better than SVM,ANN,KNN.

[4] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. Machine learning algorithms (SVM,RF) and Deep Learning is based on algorithms which are used for predicting of diabetes. The results have showed that R F is more effective for the classification of diabetes which produced overall accuracy for diabetic prediction to be 80.67%.

[5] Ramzan, M. (2016). Comparing and evaluating the performance of WEKA classifiers on critical diseases. Naive Bayes, Random Forest and J48 Decision Tree are the ones used to compare classifiers to predict critical diseases using the WEKA tool. Random forest however turns out with a higher accuracy which is more than both J48 and Naïve Bayes.

[6] Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.