Journal of Innovative Engineering and Research (JIER)
Vol.- 7, Issue -1, April 2024, pp. 12-17 (6 pages)

Advancing Early Detection and Prediction of Diabetes Mellitus: A Robust Soft Voting Ensemble Machine Learning Approach

SARKAR PRASANJEET JYOTIRMAY, DR. SANTOSH PAWAR

Department of Electronics & Communication Engineering,
Dr. A. P. J. Abdul Kalam University, Indore 452010, India
Correspond Author Email: sarkarprosanjeet08@gmail.com

Abstract— *Diabetes mellitus is an incurable illness related to an irregular and undeniable level of glucose in the blood. Nowadays, the healthcare industry has a large database. Machine learning and deep mining are fields that analyse huge datasets and find valuable information for the early prediction of disease. In this research paper the Pima India diabetes dataset are used, which collects diabetic and non-diabetic patient details like Glucose, Age, Blood pressure, etc. In our study, a soft voting ensemble machine learning model was used, where each base model would count toward the weighted average of probability to make the final prediction of whether a patient has diabetes or not. This model ensemble consists of eight machine learning algorithms, viz., NavieBayse, Logistic Regression, KNN, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and Light-GBM for the classification. The experimental result shows that the ensemble model has a maximum accuracy of 96.48%. Our proposed model helps the medical practitioner predict and diagnose the patient in time.*

Index terms: *Accuracy, Diabetes, Diagnosis Model, Ensemble, Machine learning, Prediction.*

I. INTRODUCTION

In the modern world, diabetes mellitus is a common metabolic disorder that cuts off human life at an early age. Diabetes mellitus is generally known as diabetes [1]. The blood glucose in the human body is the rich source of energy and which is generated from carbohydrate foods [2]. Insulin regulates the sugar level in the blood to prevent high and low sugar levels problem [3]. Diabetes mellitus, which causes the organ pancreases to be unable to produce enough insulin, due to a deficiency of insulin level in the human body, does not maintain blood glucose levels and causes many serious complications in the body such as stroke, kidney failure, eye blindness, nerve damage, etc [4]. Diabetes mellitus is divided into three parts. Type-1 diabetes allowed the body to destroy beta cells that live in pancreases and stop generation of insulin. Generally, it is related to genes and is mostly found in children [5]. Type 2 diabetes occurs more commonly found in 18 to 70 year-olds. In this type-2 diabetes, the human body organ pancreas produces an inadequate amount of insulin or resists the genration of insulin, which makes it unable to maintain blood glucose levels [6]. Type 3 diabetes is known as gestational diabetes; it developed during the pregnancy period of the woman and finally affects the generation of the baby after the birth and causing death [7].

As of now, current research in medical science does not provide full care for diabetes; only early complication growth symptoms identification and prediction of diabetes mellitus can control the spreading of the disease [8]. The development of intelligent systems in the medical field gives valuable information to the medical practitioner for the diagnosis of disease [9]. Machine learning and data mining technique have great strength in managing bulk amounts of dataset originated from several sources for auspicious examination and information extraction. This research work focused on building an ensemble model using a soft voting classifier for classifying diabetic and non-diabetic patient in given dataset. Nave Bayse, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and LGBM all machine learning model output have been combined and found the performance of the ensemble model achieved a better result as compared to the given base model classifier.

The best parts of this research paper are prepared as follows: Section II presents the literature review on machine learning model and ensemble methods used in this area. Section III gives the methodology for the ensemble approach, which uses soft voting for the classification of diabetes. Section IV gives the experimental findings using the suggested model. Section V presents the conclusion of the research work.

II. RELATED WORK

In last one decade many excellent research has been done for the classification of diabetes using machine learning models.

HafsaBinteKibria et al. (2022) developed an ensemble machine learning model for the classification of binary class diabetes dataset intodiabetic and non-diabetic patient. Algorithm ensemble of Artificial Neural networks (ANN), Random forests, Support Vector machines, Logistic regression, and XGBoost. In this paper, Pima India diabetes dataset was used for the experimental work. The missing value was imputed using the median value technique and to balance the dataset a technique of synthetic minority oversampling were used. This model used five-fold cross-validation to get a maximum accuracy of 90% [10].

RamyaAkula et al. (2019) discussed the prediction of Type-2 diabetes using an ensemble model. The model used Pima India diabetes dataset for the experimental work. The performance matrix was used to measure the robustness and accuracy of the model. Result obtained by the ensemble model with 89.1% accuracy [11].

GauravTripathi et al. (2020) give a comparative study of four machine learning model for the early prediction of diabetes mellitus. In this model, Pima India diabetes dataset was used for the experimental work and performance was computed by comparing all machine learning model, such as Linear Discriminant Analysis, K-Nearest neighbour, Support Vector machine, and Random Forest. Found that random forest outperformed the other four algorithms with 87.66% accuracy [12].

UmairMuneer Butt et al. (2021) presented a machine learning model for theprediction and classification of diabetes and also used anIoT-based hypothetical monitoring system for monitoring the person's blood glucose level. In this experiment, the Pima India diabetes dataset is used. Experimental results found that LSTM gave the highest prediction with 87.26% accuracy[13].

SaloniKumari et al. (2021) discussed the classification and prediction of diabetes patient using an ensemble model. They utilized the Pima India diabetes dataset and the breast cancer dataset for the experiment. In this model, the Pima India diabetes dataset is normalized using linear transformation of data and missing or null values are replaced by the mode of the individual column features. The accuracy of the ensemble classification using soft voting is 79.04% [14]..

YashiSrivastava et al. (2019) estimated gestational diabetes using a machine learning model. For the experiment, we used the Microsoft Azure AI service performed on the Pima India diabetes dataset. To compute the performance of the algorithm, divide the dataset randomly in the ratio of 70% and 30%, where 70% of the dataset sample was used for the training purpose and 30% dataset sample for the testing purpose. Two-class logistic regression is used to predict diabetes with an accuracy of 77.8% [15].

The research objectives are summarized as follows:

- The proposed model classifies the diabetes mellitus into diabetic and non-diabetic classes using Pima India diabetes dataset.
- To test the suggested model's robustness,F1-score Accuracy, Recall and Precision were used as the evaluation criteria.
- A comparison of the existing base machine learning model with the proposed ensemble model for finding the superior result

III. PROPOSED METHODOLOGY

In this research paper extensive focus on enhance the accuracy of the proposed model for the early and accurate detection of diabetes. Researchers have developed an ensemble of machine learning modelfor the classification of diabetes dataset using soft voting classifier. Thefigure1shows the flow chart of the suggested ensemble model using a soft voting classifier.

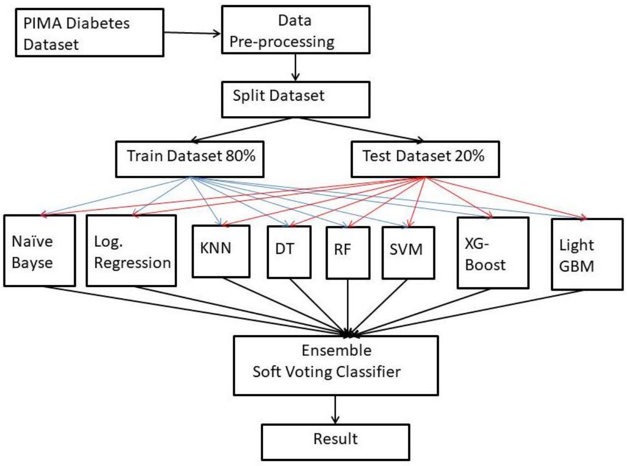


Figure 1. Flow chart for suggested ensemble model using soft vote classifier.

A. Dataset

In this experiment, the Pima India diabetes dataset was used and patient data has collected by the University of California Irvine for the research work. The diabetes dataset was downloaded from the given website www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. The dataset has eight valuable feature columns and one output column that specifies whether the person has non-diabetic (0) or diabetic (1). The given dataset has768 records, of which 268 are positive diabetes and the other 500 are negative diabetes.

B. Data Pre-Processing

Data pre-processing is utmost important step in machine learning model because model output was completely depend on the more valuable and efficient format of data. The first step in data pre-processing is addressing missing values in the feature column, and all the missing or null values are replaced by the mean of the particular feature column. In second step, data pre-processing is data normalization and in this min-max operation to perform linear transformations of the data to convert all eight independent feature column values into the same range between 0 and 1.

C. Data Split

In this step, the researcher randomly divided the dataset into two parts, 80% and 20%. The 80% part is used to train the model and the 20% part of the dataset is used to test the performance of the model.

D. Base Model

1) Naive Bayse

The working principle of Naïve Bayse algorithm depends on probabilistic method. It is a supervised machine learning algorithm used for solving classification problems. It uses the Bayes theorem by the consideration that the existence of one sample in a given class is independent of the existence of another sample in the same class. This independence consideration requires studying each parameter individually for every term.

2) Logistic Regression

It is more popular supervised machine learning classification algorithm. Logistic regression is almost identical to linear regression except that linear regression is used for the continuous data outcome and logistic regression is used for solving theclassification outcome. It works on the probabilities of an event occurring, and sigmoid function is used to map each data point in the dataset.

3) KNN

It is a supervised algorithm used in machine learning for both regression and classification problems. In this algorithm, the new feature class is determined based on the distance or similarity measurement of existing features in the same class. Generally, the Euclidean distance or Manhattan distance is used for measuring the similarity of a new feature with an existing class feature. This algorithm is very slow because it does not straightaway learn from the training dataset; rather, it stores the training dataset and applies an action to it when classifying.

4) Decision Tree

A decision tree model is used in both classification and regression problems, but it is more popular in solving classification problems. This work uses the graphical representation of a dataset to get the optimistic solution to a decision based on the given condition. A decision tree consists of two nodes, which are the decision nodes and leaf nodes. Decision nodes are used for to make the decisions, and if possible, n- number of branches and leaf nodes shows output of the decision tree.

5) Random Forest

It is one type of ensemble machine learning classifier algorithm. It consists of many trees and bootstrap aggregation techniques and applied training dataset to every tree in the algorithm. Each decision tree gives a high variance output; when it is combined and each of them is taken in parallel, the resultant variance is low, and the resultant output does not depend on single decision tree output but on multiple decision trees output.

6) Support Vector Machine

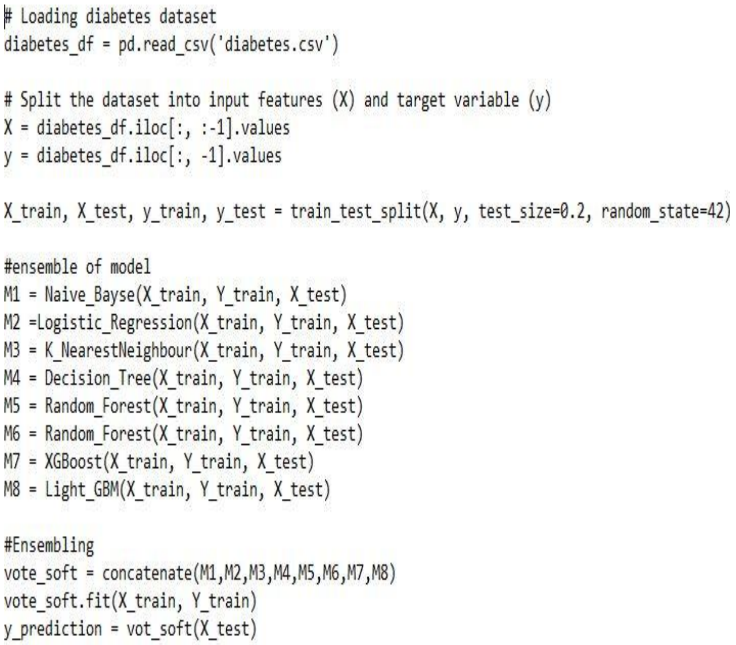
Support vector machines are used for both regression output and classification output problems.In this we can quickly place the new data sample in the appropriate category of the feature by creating the best decision boundary that divides n-dimensional space into classes. SVM chooses the optimum hyperplane in the multi-dimensional space, and the correcthyperplane has the maximum distance between the two ends of the data point. The unknown feature sample point is classified based on the hyperplane and fit into either one of the classes along the hyperplane.

7) XgBoost

It is a supervised machine learning algorithm used in both regression and classification problems. It uses a gradient boosting algorithm and an ensemble approach to train the model. The accuracy achieves of this model is higher than a single decision tree but sacrifices the basic interpretability of a decision tree.

8) Light GBM

It is a supervised ensemble machine learning classification algorithm used in both regression and classification problems. It is an extension of the gradient boosting algorithm, automatic feature selection and focusing on boosting examples with large gradients. It utilizes two techniques: gradient-based one-side sampling and exclusive feature building, which result in improved speed of training and accuracy of performance.



```
# Loading diabetes dataset
diabetes_df = pd.read_csv('diabetes.csv')

# Split the dataset into input features (X) and target variable (y)
X = diabetes_df.iloc[:, :-1].values
y = diabetes_df.iloc[:, -1].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#ensemble of model
M1 = Naive_Bayse(X_train, Y_train, X_test)
M2 =Logistic_Regression(X_train, Y_train, X_test)
M3 = K_NearestNeighbour(X_train, Y_train, X_test)
M4 = Decision_Tree(X_train, Y_train, X_test)
M5 = Random_Forest(X_train, Y_train, X_test)
M6 = Random_Forest(X_train, Y_train, X_test)
M7 = XGBoost(X_train, Y_train, X_test)
M8 = Light_GBM(X_train, Y_train, X_test)

#Ensembling
vote_soft = concatenate(M1,M2,M3,M4,M5,M6,M7,M8)
vote_soft.fit(X_train, Y_train)
y_prediction = vot_soft(X_test)
```

Figure 2. Programming structure of proposed model

E. Proposed Model

The proposed model for binary classification (1 or 0) consists of merging the same or dissimilar machine learning classification model form a strong meta-classifier model that predicts through majority voting. Hard vote classifiers and soft vote classifiers are the two types of voting classification used in the ensemble classification technique: Hard vote classifiers and soft vote classifiers. In hard voting, input data is classified based on the mode prediction made by each base model, with the class that receives the most votes as the final prediction. In soft voting, classify input dataset on the probability score of each base model for each class and calculate the weighted average of the probability to make the final prediction.

The presented ensemble soft voting classifier model achieves higher accuracy results than other base models classifier, as it merges the predictions of different models. This model ensemble Naive Bayse, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and LGBM. Ensemble soft voting classifier has been used to measure the probability of each base model and average of this probability makes the final decision. The basic programming structure of the proposed model is written in Anaconda platform using Scikit-Learn library, as shown in the figure 2.

IV. RESULTS

The proposed model uses the Pima India diabetes dataset for the experiment. The dataset has 768 patient records and 9 feature columns, with 268 positive samples of diabetes and 500 negative samples of diabetes. Figure 3 shows the ratio of non-diabetics (0) and diabetic (1) patients in the dataset.

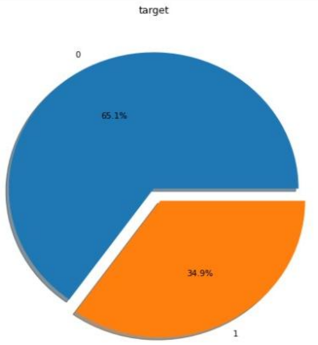
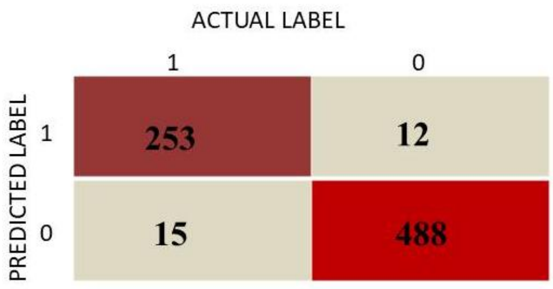


Figure 3. Ratio of Diabetic and non-diabetic patient in the dataset

The given dataset was split randomly into two parts. 80% of the dataset used in training the model and 20% of the dataset used in testing the model. The confusion matrix in figure 4 was used to determine the model's performance.



ACTUAL LABEL

PREDICTED LABEL	1	0
1	253	12
0	15	488

Figure 4. Confusion matrix of model

The most important evaluation parameters, Accuracy (1), Precision (2), Recall (3), and F1-Score (4) are used to measure the algorithm's robustness and effectiveness. They are determined using the formula below [16]

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \tag{1}$$

$$Precision = \frac{Tp}{Tp+Fp} \tag{2}$$

$$Recall = \frac{Tp}{Tp+Fn} \tag{3}$$

$$F1-Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Where,

Tp is the true positive means the value of the predicted outcome is 1 and actual outcome is 1.

Tn is the rue negative means the value of predicted outcome is 0 and the actual outcome is also 0.

Fn is the false negative means the value of predicted outcome is 0 but the actual outcome is 1.

Fp is the false positive means the value of predicted outcome is 1 and the actual outcome is 0.

The comparative study of the proposed ensemble machine learning algorithm with the base machine learning algorithm for the diabetes mellitus classification as shown in performance shown in Figure 5

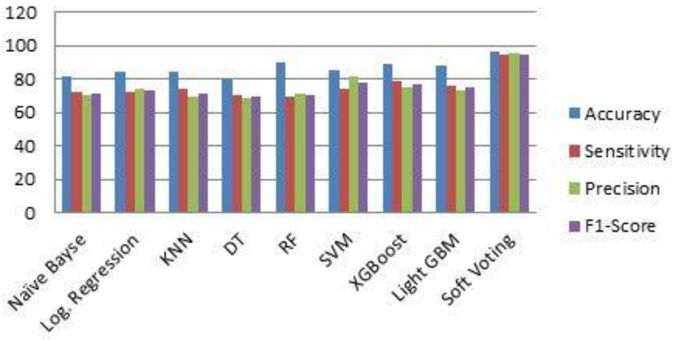
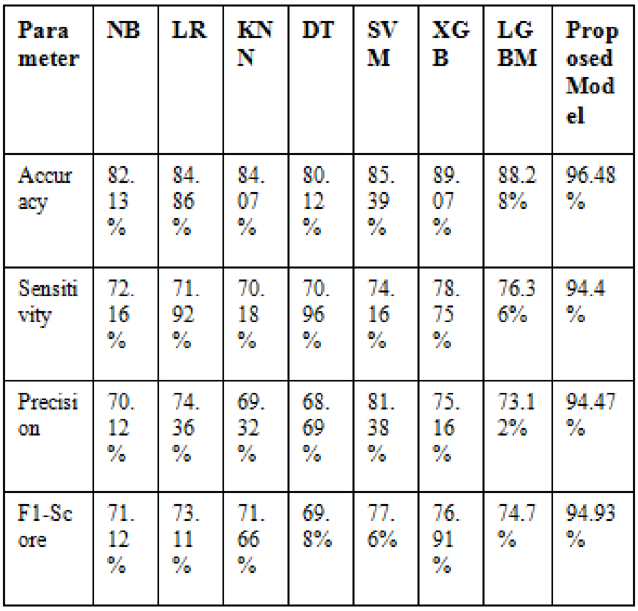


Figure 5. Performance comparison of all machine learning algorithm



Parameter	NB	LR	KNN	DT	SVM	XGB	LGBM	Proposed Model
Accuracy	82.13%	84.86%	84.07%	80.12%	85.39%	89.07%	88.28%	96.48%
Sensitivity	72.16%	71.92%	70.18%	70.96%	74.16%	78.75%	76.36%	94.4%
Precision	70.12%	74.36%	69.32%	68.69%	81.38%	75.16%	73.12%	94.47%
F1-Score	71.12%	73.11%	71.66%	69.8%	77.6%	76.91%	74.7%	94.93%

Table 1. Comparisons of the machine learning algorithm

Researchers has been observed from table 1 that performance of individual algorithm are less than the proposed ensemble model with accuracy 96.48% result.

V. CONCLUSION

Diabetes is a deadly disease that is normally found in young generation these daysand early prediction of this disease is always a challenging problem for practitioner. The aim of this research paper has to achieve high-accuracy performance model for predicting early and accurate diabetes disease. The researchers have proposed an ensemble model using a soft voting classifier. The PIDD dataset from UCI was used for the experiment work, and performance matrixes such as accuracy, precision, recall, and F1-scor have been evaluated. The ensemble model examined on the 768 sample and it was able to achieve an accuracy of 96.48%.R

REFERENCES

1. Hossam A. Shouip, "Diabetes mellitus,"
2. ResearchGate, Dec. 2014.
3. Aslam S, Martin A Holesh JE, Physiology, Carbohydrates. Treasure Island: StatPearls Publishing, 2023.
4. Wilcox G, "Insulin and insulin resistance," The Clinical biochemist. Reviews, vol. 26(2), pp. 19-39, May 2005.
5. American Diabetes Association, "Diagnosis and classification of diabetes mellitus," Diabetes care, pp. 62-67, Jan 2009.
6. B. O., Thomaidou, S., van Tienhoven, R., & Zaldumbide, A Roep, "Type 1 diabetes mellitus as a disease of the β-cell (do not blame the immune system?)," Nature reviews. Endocrinology, vol. 17, no. 3, pp. 150-161, 2021.
7. Y., Ding, Y., Tanaka, Y., & Zhang, W Wu, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention," International journal of medical sciences, vol. 11, no. 11, pp. 1185-1200, September 2014.
8. C. M., Arnegard, M. E., & Maric-Bilkan, C Silva, "Dysglycemia in Pregnancy and Maternal/Fetal Outcomes," Journal of women's health, vol. 30, no. 2, pp. 187-193, February 2021.
9. M. D., Malabu, U. H., Malau-Aduli, A. E. O., & Malau-Aduli, B. S. Adu, "Enablers and barriers to effective diabetes self-management: A multi-national investigation," PloS one, vol. 14, no. 6, June 2019.
10. T.Kalakota, & R. Davenport, "The potential for artificial intelligence in healthcare," Future healthcare journal, vol. 6, no. 2, pp. 94-98, June 2019
11. Nahiduzzaman M., Goni M.O.F., Ahsan M.,& Haider J. Kibria H.B., "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," Sensors, vol. 22, no. 19, September 2022.
12. Ivan & Akula, Ramya Garibay, "Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes.," in IEEE Southeastcon, Huntsville, Alabama, 2019.
13. Gaurav & Kumar, Rakesh Tripathi, "Early Prediction of Diabetes Mellitus Using Machine Learning," in 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India., 2020, pp. 1009-1014.
14. Umair & Letchmunan, Sukumar & Ali, Mubashir & Hassan, Fadratul Hafinaz & Baqir, Anees & Sherazi, Husnain Butt, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," Journal of Healthcare Engineering, vol. 7, pp. 1-17, October 2021.
15. Deepika Kumar, & Mamta Mittal Saloni Kumari, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40-46, June 2021.
16. Yashi & Khanna, Pooja & Kumar, Sachin Srivastava, "Estimation of Gestational Diabetes Mellitus using Azure AI Services," in Amity

International Conference on Artificial Intelligence (AICAI), Noida, Delhi, 2019, pp. 321-326.

17. Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R,& Isaza G Orozco-Arias S, "Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements," Processes, vol. 8, no. 6, May 2020.