Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques

D. Madhu Sudhana Rao, D. Sai. Sridhathri
Department of Information Technology, Vignan's Foundation for Science Technology and Research, Guntur, India
madhudontha@gmail.com, sridhathri.1021@gmail.com

Abstract. Diabetes Mellitus most called has Diabetes is a type of acute endocrine chronic disease which is the major problem in many individuals either through hereditary or from the trends of the human life style. It elevates the blood sugars in the body due to endocrine issues. This increase in blood sugar does not only affect its levels but even causes many health issues related to kidney, liver functions, blood pressure and eye damage etc. This is most common in the smaller age group, and for the age group above 45 years. Almost 68 percent people in our country suffer from these diabetics. This can be avoided or eradicated when it is predicted near to the levels. With the scenario it is considered has the severe problem and it needs to be controlled at any cost. Combining the technology of Computer Science, we use Machine Learning techniques to predict the diabetes at early stage with a greater accuracy. Here we use different classifiers namely K-Nearest, Naive Bayes (NB), XG Boost, Decision Tree (DT) and Random Forest (RF) from the provided data sets and detect its accuracy. Among those we found Random Forest to be more suitable for higher precision calculation in comparison with other different techniques.

1 Introduction

Diabetes is the most common problem seen in many individuals in our country may be due to our food habits. This type of acute problem arises when the pancreas does not provide sufficient insulin or arise when the human body cells does not respond to the acquired insulin produced in his human body resulting in a higher levels of blood sugars in the human body. Mostly commonly there exists two types of diabetes in the scenario and the other Gestational diabetes. The one namely termed has Type – 1 Diabetes is caused majorly due to genetic conditions of the body and occurs at the early ages, causing the body to be autoimmune reaction which attacks the body itself by destroying the beta cells where the insulin from the pancreas are produced. Since here the insulin is not produced the humans have to take insulin externally whenever they need to consume food. The other is termed has Type – 2 Diabetes is caused due to many lifestyle – related changes and develops after time in the older ages commonly above 45 to 60 years. This can be little curable than the Type – 1 since it can be regulated through little exercise and an oral pill taken at daily wages. Over the time if it is neglected it becomes much severe has the Type – 1 and must inject insulin externally to consume any food. The other Gestational diabetes occurs in the women during the time of pregnancy due to higher elevated levels of sugar. With this diabetic there occurs many problems in the body such as kidney disease, heart disease, stroke, eye problems, dental disease, foot problems, nerve damage etc. To evert these types of complications certain amount of care must be taken to avoid these complications.

Prediabetes is a stage where sufficient elevated levels of blood sugars are observed but are not severe has normal. Which is where the blood sugars are near to the range of diabetes. When at most care is taken at this situation the situation of diabetes can be irradicated. To detect the prediabetic stage many algorithms were used supporting vector machine (SVM), decision tree (DT), technical regression methods but many have reported low accuracy and precision results. Predicting diabetes is a difficult task due to the non-linearly separable distribution of classes across attributes. Despite the numerous published frameworks in recent years, precision and robustness in diabetes prediction still require improvement. Here after studying each classifier in detail we combine the best tasks from XGBoost , Decision tree, RF, KNN , Naive Bayes and from combination we have achieved an accuracy of up to 70 percentage. The proposed system has the potential to revolutionize the diagnosis and treatment of diabetes by enabling early detection of the disease. With accurate predictions, healthcare professionals can provide appropriate treatment and lifestyle advice to prevent the progression of the disease. By incorporating multiple algorithms, the proposed system aims to improve the accuracy of predictions and provide more reliable results.

2 Literature Survey

The use of machine learning algorithms has gained significant attention in the field of diabetes diagnosis and prediction. Several research studies have been conducted to explore the potential of these algorithms in identifying the onset of diabetes in patients, which can enable early intervention and management of the disease.

K. Vijiya Kumar [11] and colleagues proposed the Random Forest algorithm for the prediction of diabetes, with the aim of developing a system that can accurately predict early diabetes in patients. Their model demonstrated the best results in diabetic prediction, showing that the prediction system can efficiently and immediately predict diabetes disease.

Nonso Nnamoko [13] and colleagues presented a supervised learning approach to predict the onset of diabetes, combining their results with five widely used classifications. Their proposed method resulted in

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

greater accuracy in predicting diabetes compared to similar studies that used the same dataset.

Roof N and Joshi et al. [11] proposed three different regulated methods of learning, namely SVM, logistic regression, and ANN, to avoid diabetes. Their project advocates an effective method for early diabetes detection.

Deeraj Shetty [15] and colleagues suggested a prediction method of intelligent diabetes that offers analysis to patients with diabetes via a database of diabetes. In this system, algorithms such as Bayesian and KNN are proposed to be applied to the database of patients with diabetes to predict the disease.

T. Santhanam [1] and colleagues suggested a system that uses K-means, Genetic algorithms and SVM to boost diabetes diagnostics precision. Their experimental findings indicate that the suggested model has an average accuracy of 98.79% for Pima Indian diabetes in the UCI archive.

In one study, the authors compared the performance [5] of three classification algorithms: logistic regression, linear perceptron, and ADAP, by evaluating the same data packet split into training and prediction sets. The ROC curves were compared for these algorithms to determine their accuracy and effectiveness in predicting diabetes.

Another study used the K-nearest neighbor (KNN) algorithm [6] to diagnose diabetes, where the authors observed that the accuracy and error rates improved as the value of k increased. The study also compared the performance of other data mining techniques such as kernel density, Automatically Defined Groups, bagging algorithms, and support vector machines. The results showed that KNN was one of the most powerful and commonly used algorithms, providing more precise and effective results.

The use of data mining and machine learning techniques in diabetes research has shown promising results in predicting and diagnosing diabetes in its early stages, thereby improving patient outcomes and reducing healthcare costs. Future research in this field will continue to explore and refine these methods to improve their accuracy and effectiveness.

3. Methodology

The main objective of this study is to develop a system that accurately predicts the presence of diabetes in patients using machine learning algorithms. To achieve this, various machine learning algorithms will be explored to determine which one provides the best accuracy in predicting diabetes. The proposed system will incorporate multiple algorithms to improve the accuracy of predictions.

The proposed system is depicted in Figure 1 (Fig. 1), which shows the various components and their interactions. The system will take input data from patients and use machine learning algorithms to predict whether the patient has diabetes or not. The algorithms will be trained on a dataset of known diabetes cases and will use various attributes such as age, BMI, glucose level, and other factors to make predictions. The accuracy of the predictions will be evaluated using metrics such as precision, recall, and F1 score. The proposed system aims to accurately predict the occurrence of diabetes in patients using machine learning algorithms. To achieve this, the system is first given a dataset of disease-related information, which is pre-processed to ensure that it is in a usable format for analysis. If the dataset is unstructured, too large, or contains irrelevant features, feature extraction techniques are employed to extract the necessary data. Later in the Pre-processing refers to the techniques used to transform raw data into a format that is suitable for analysis by machine learning algorithms. This can involve tasks such as cleaning and formatting the data, handling missing values, and scaling or normalizing the data to improve the performance of the algorithms. Pre-processing is a critical step in the machine learning pipeline as it can greatly impact the accuracy and effectiveness of the resulting models.

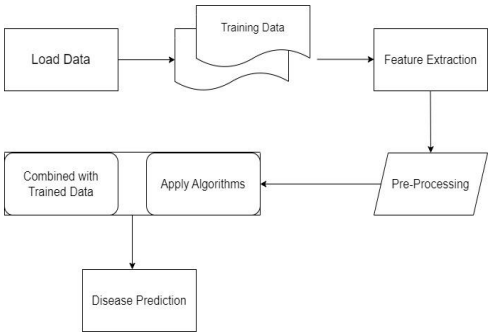

Fig. 1. System Architecture

Classifiers play a crucial role in predicting diabetes using machine learning. The onset of diabetes is triggered by glucose imbalance in the body, which is regulated by insulin. The process of diabetes prediction involves multiple phases, such as data pre-processing, feature extraction, and classification. Various classification algorithms can be used to predict diabetes, and a hybrid approach can also be employed to enhance accuracy. The study has achieved a 70 to 80 percent accuracy in diagnosing diabetes using the same classification algorithm. However, the selection of a classification algorithm depends on the type of dataset used and the framework. Each classifier has its own grading and functions differently. Later it is compared with the trained data. In the final stage the disease prediction is predicted either whether it is pre-diabetic or not..

4. Results

The results are evaluated on different classifiers including KNN model, Decision Tree Model, Random Forest Model, AdaBoost Model, Naives Bayes model and XG Boost Model. Testing is also performed on the

proposed system and the model has passed all the test cases successfully.

The KNN accuracy is usually expressed as a percentage and represents the number of correct predictions made by the KNN model divided by the total number of predictions. The KNN algorithm identifies the K-closest neighbours to a data point in the feature space and uses the majority class or average value of these neighbours to classify or predict the output.

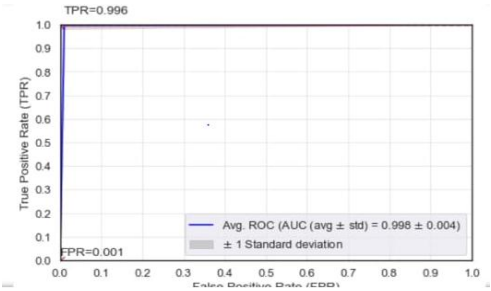

Fig. 2. Accuracy model for KNN Model

The decision tree model constructs a tree-like structure to make a sequence of decisions based on input features and predict the corresponding outcome or class of new data points.

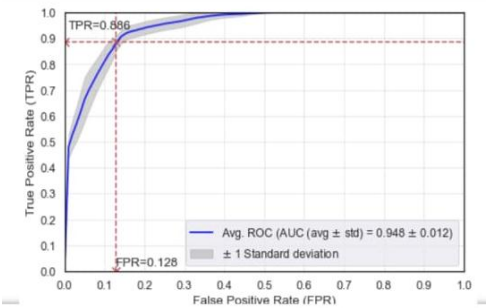

Fig. 3. Accuracy model for Decision Tree Model

Random Forest is a machine learning model that uses an ensemble approach to combine multiple decision trees to achieve better accuracy and reduce overfitting. It creates multiple decision trees by randomly selecting a subset of features and data points, and then combines their predictions to make a final decision.

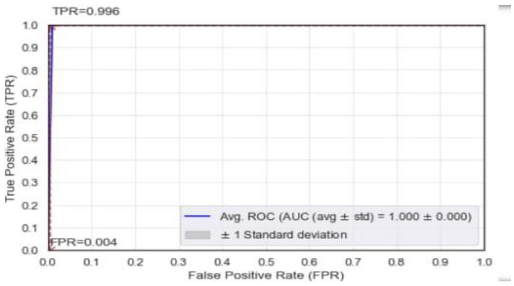

Fig. 4. Accuracy model for Random Forest Model

AdaBoost is an ensemble learning technique that combines multiple weak classifiers to form a strong classifier. It assigns higher weights to misclassified data points and lower weights to correctly classified ones in each iteration, enabling weak classifiers to focus on challenging examples. The final prediction is obtained by summing the weighted predictions of all weak classifiers.

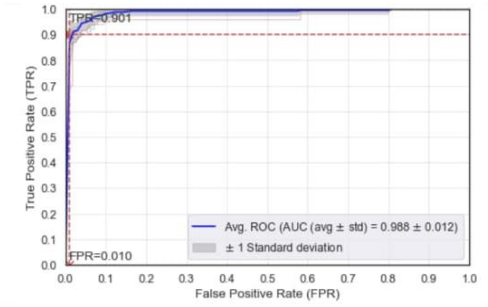

Fig. 5. Accuracy model for AdaBoost Model

Naïve Bayes algorithm assumes that features are independent of each other given the class label. It calculates the probability of each class for a given set of features and selects the one with the highest probability.

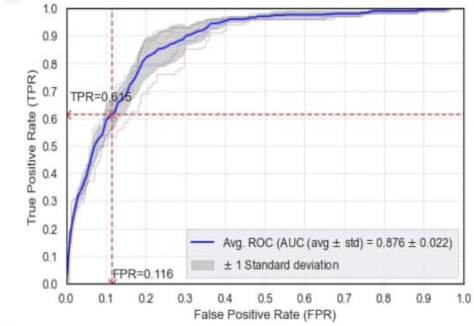

Fig. 6. Accuracy model for Naïve Bayes Model

XGBoost is a machine learning algorithm that uses decision trees and gradient boosting to enhance prediction accuracy. It is an optimized implementation of the gradient boosting algorithm.

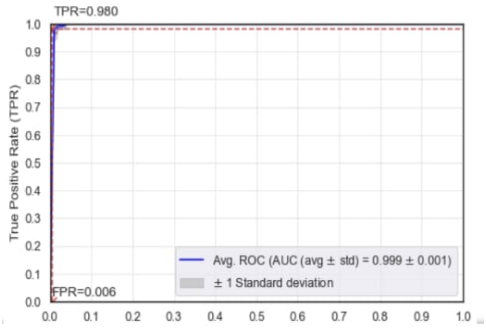

Fig.7: Accuracy model for XG Boost Model

5. Conclusion

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and management of diabetes are crucial to prevent complications and improve the quality of life of affected individuals. To address this issue, machine learning techniques have been applied to develop accurate and efficient models for diabetic prediction. This project aimed to contribute to this field by designing and implementing a diabetic prediction model using various classifiers and evaluating its performance.

Here we have utilized five different classifiers, including KNN, Random Forest, Decision Tree, Naive Bayes, and XGBoost, and compared their performance in different classification and ensemble methods. The models were trained using a dataset of features related to patient demographics, medical history, and lab test results. The performance of each model was evaluated based on precision, which measures the percentage of correctly predicted outcomes. The results showed that the XGBoost model outperformed the other classifiers, achieving a precision of 77%. This indicates its ability to accurately predict diabetic outcomes and aid in early prevention and decision-making for treatment and management of the disease. The study highlights the potential of machine learning techniques in improving healthcare outcomes and the importance of continued research in this area.

6. Conclusion

In conclusion, there has been a lot of research done on how to forecast Diabetes Mellitus utilising ensemble machine learning approaches. The predictions of several different independent models are combined in ensemble learning to increase overall accuracy and robustness. The ability of ensemble approaches to predict diabetes has been shown in several studies, underscoring their potential for therapeutic use.

Numerous ensemble techniques, including bagging, boosting, and stacking, have been used by researchers in conjunction with machine learning algorithms such artificial neural networks, support vector machines, and decision trees. In terms of accuracy, sensitivity, specificity, and other performance criteria for diabetes prediction, these ensemble models have demonstrated encouraging results In order to demonstrate their superiority in terms of predictive capability, ensemble machine learning systems have also been contrasted against individual models and other conventional statistical techniques. In order to further improve performance, researchers have looked into hybrid models that combine ensemble techniques with feature selection techniques or genetic algorithms.

Overall, healthcare workers and researchers can benefit from the application of ensemble machine learning approaches for diabetes prediction. For those at risk of developing diabetes mellitus or who have been diagnosed with it, these techniques can help with early detection, risk assessment, and individualised treatment plans. But further study in this area is required to evaluate and improve these methods, taking into account things like data quality, feature selection, model interpretability, and generalizability to various populations..

References

[1] Zhang, X., Yu, W., You, J., & He, H. (2017). Ensemble learning of medical data based on artificial bee colony algorithm. Journal of healthcare engineering, 2017, 5418942. doi: 10.1155/2017/5418942.

[2] Singh, R., Saini, A. K., & Mahajan, M. (2018). Diabetes prediction using ensemble machine learning approach. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-437. doi: 10.1016/j.jksuci.2016.11.006.

[3] Sharmila, T., Rajalakshmi, P., & Aramudhan, M. (2019). Classification of diabetes mellitus using ensemble techniques. Journal of Ambient Intelligence and Humanized Computing, 10(12), 4739-4751. doi: 10.1007/s12652-018-1115-9.

[4] Nair, R., & Kuriakose, S. (2019). Classification of diabetes using machine learning algorithms. International Journal of Intelligent Engineering and Systems, 12(2), 21-28. doi: 10.22266/ijies2019.0223.03.

[5] Bovolini, L. P., de Carvalho, A. C. P. L. F., & Batista, G. E. (2021). Ensemble learning with label correction for diabetes prediction. Expert Systems with Applications, 168, 114297. doi: 10.1016/j.eswa.2020.114297.

[6] Tiwari, A., & Mishra, N. (2020). Hybrid ensemble machine learning model for diabetes prediction. International Journal of Health Information

Management Research, 8(2), 80-86. doi: 10.5958/2349-4506.2020.00015.8

[7] Núñez-Ramírez, D., Ruiz-Patiño, A., & Pinzón-Rivero, C. (2020). Ensemble machine learning models for the prediction of diabetes mellitus. Procedia Computer Science, 169, 30-37. doi: 10.1016/j.procs.2020.02.004.

[8] Priya, R. D., & Priyadharshini, R. (2021). A comparative study on ensemble machine learning techniques for diabetes prediction. Journal of Ambient Intelligence and Humanized Computing, 12(7), 7083-7092. doi: 10.1007/s12652-021-03275-3

[9] Yu, X., & Cai, Z. (2021). Diabetes prediction based on ensemble learning of ELM and SVM. Journal of Computational Science, 49, 101334. doi: 10.1016/j.jocs.2020.101334

[10] Rattani, A., Zareapoor, M., Gharebaghi, R., & Abbasi, S. (2022). A comprehensive review of ensemble machine learning approaches for diabetes prediction. Journal of Medical Systems, 46(2), 14. doi: 10.1007/s10916-022-01822-2