An approach to prevent Diabetes using Machine Learning in early stage

A PROJECT REPORT

Submitted by

MADHULIKA G (2116210701139)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“An approach to prevent Diabetes using Machine Learning in early stage**” is the bonafide work of “**MADHULIKA G (2116210701139)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on___________________

Internal Examiner **External Examiner**

Abstract

Insufficiency in the body's response to insulin causes diabetes, a chronic disease that impacts everyone. Both insufficient insulin production by the pancreas and inefficient insulin utilization may lead to this condition. Discovering and treating dangerous illnesses early is the only way to reduce their impact, given there is currently no cure for diabetes. In order to detect diabetes individuals at an early stage, many research have used various machine learning algorithms. But improving the precision of predictions is never an easy job. Various classification systems have varied performance metrics when it comes to diabetes prediction. Adaboost shows a high level of accuracy of 97.15% when tested with several models, including KNN, XGboost, logistic regression, and random forest.By preprocessing the data and selecting features based on the correlation coefficient, a generalized model and a resilient framework are suggested.A dataset's correlation coefficient reveals the nature of the connection between any two variables.Using k-fold cross validation, one may learn about the system's stability and performance by looking at its standard deviation and mean accuracy.The patient is able to avoid becoming a chronic condition with the use of this model, which aids in the early stages of diabetes prevention.

Keywords:

Diabetes ,Machine learning, Adaboost

1. Introduction

Diabetes is one of the diseases that is widely known and presents significant issues.The pancreas secretes the hormone, which allows glucose from meals to enter the bloodstream. Absence of the hormone results in pancreatic dysfunction and diabetes, which can have dangerous side effects.

In order for people to live healthy lives, early detection of diabetes is crucial for preventative actions.The healthcare business places a high value on diabetes prognosis.A hemoglobin A1c blood test is often used to detect diabetes.A person's way of life might also play a role in triggering the condition. Three distinct forms of diabetes are recognized. Machine learning allows us to determine whether a person is diabetic or not. A categorization issue is what it is. This is a binary categorization since the result is either "yes" or "no" about the patient's diabetes status. Datasets undergo preprocessing to eliminate inconsistencies and improve system efficiency. The characteristics will be divided into two groups: X train and y train, and X test and y test, after the processing that came before. Next, the features are separated to train the system, and finally, the testing data is used for prediction. It determines the system's performance score.Overfitting occurs when it produces accurate results throughout training. It is referred to be underfitting if it yields satisfactory results after testing.

To determine whether or not the model is generalizable, cross validation using the K-fold is carried out.It generates the mean accuracy and standard deviation by folding the data into k subsets, which accounts for data that has not yet been viewed.A generalised model has a low standard deviation and a high mean accuracy.

Numerous studies on diabetes, prediabetes, and the possibility of patients being readmitted to hospitals as a result of their diabetes have been conducted in an effort to forecast the onset of a number of diseases. The used categorization and prediction methods have been investigated, and by comparing the various models, a strong framework is suggested.

2. Literature survey

Researchers have published many research works that identify whether the person is diabetic or not.In [1],the researchers have applied ontology based machine learning classifications to identify the diabetic patients.The study was done by the Pima Indians Diabetes Database.They used six classifiers to predict the diabetes.The output of the study showed that ontology and SVM has 77.5% and 77.3% accuracy respectively.Ontology classifier and SVM classifier has high accuracy compared to other classifiers.

The researchers in [2] have proposed the ensembling of different ML models to predict diabetes.They have used the Pima Indian Diabetes Dataset.According to the study's findings, combining Adaboost with XGboost is the best approach.

Fused Machine Learning has been proposed in [3] study to predict diabetes.They made use of the ANN and SVM models.The output of the classifier mentioned above serves as the logic's input.Cloud storage is used. Their accuracy rate was 94.87.

The research team in [4] has suggested methods for calculating the risk involved in type 2 diabetes.The data used in this research work is from the ELSA database.Then they have applied different ensemble learning approaches to get the output.

A method for predicting the blood glucose level after a meal in women with diabetes who are pregnant was suggested by the authors in [5]. A gradient boosting strategy based on trees was suggested by them.Clinical trials provided the source of the data used in this study.

In [6] the study used two different datasets for prediction,they are synthetic and Pima Diabetes Data.Results showed that the ANN model has attained 81.69% which is the highest accuracy among other classifiers.

Through the use of multivariate classification algorithms, the researchers in [7] have developed a predictive model for metabolic syndrome in diabetes growth. The algorithm includes naive bayes and decision trees. The outcomes proved that the Naïve Bayes under-sampling strategy with K-medoids is superior to random under-, over-, and no sampling.
In [8] the researchers have used a multimodal dataset which has four types of sensor data.The different sensors include Electrocardiogram sensor,Accelerometer, breathing sensor data and

sugar level.XGboost gives 98.2% accuracy.The Xgboost gives this accuracy with multisensor data.

The recommended method [9] makes use of an ensemble model that is voted on in relation to KNN-associated features.This research looks at the problem of left data in diabetes care and shows how useful KNN features are.

In this study[10], they have used XBG and six more classifiers.The SVE method combines the output of multiple classifiers by giving each classifier an equal weight. The features which contain details on individuals with and without diabetes, are used in the investigation. It has been demonstrated that the value that was missed imputation outperforms conventional classifiers for data pre-processing and ensemble classification.

AdaBoost, RF, and bagging strategies were used in the study [11]. When it came to predicting diabetic illness, RF received the greatest rating among the three algorithms.

In this study [12] the selected ensemble learning method shows remarkable ability to induce complex patterns that are essential for the prediction of chronic diseases.By comparing the suggested hybrid model—which combines XGBoost and Bayesian Optimization—with a number of well-known ML models, this study showed its exceptional performance. SVM, DT, RF, KNN, GNB, LR, and MLP were all included in a comprehensive comparison analysis.

3. Proposed Methodology

The present study uses machine learning to predict diabetes at an early stage. The algorithms are employed to avoid the condition of diabetes.

The different models are evaluated and their performance metrics are calculated.The methodology includes several steps.

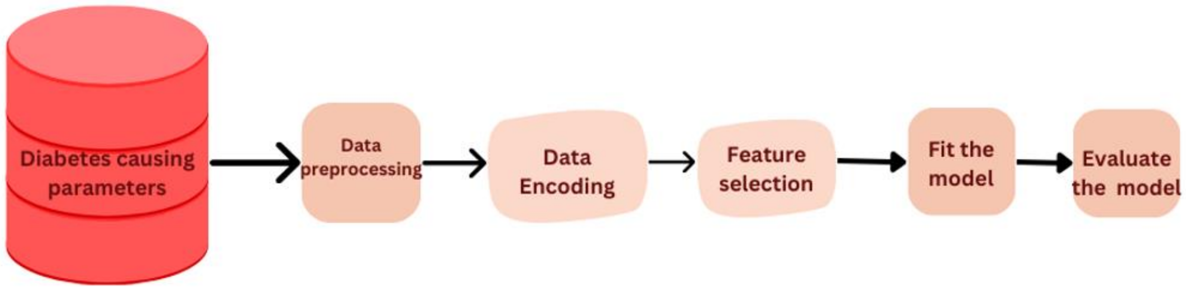

Fig.Proposed method

3.1 Gathering Data:

The characteristics from the Kaggle data pool are used in the study. This dataset has nine columns.
The age data type is changed from float to int in the dataset.The blood glucose level data type is changed to float.

The dataset contains information on gender, including male, female, and other.There is more data for no info and never than for former, present, ever, not_current. Accordingly, the smoking history has a negligible impact on the accuracy of the model.This might lead to discrepancies in the results.Consequently, the smoking_history column is removed.To transform the gender characteristic from categorical to numerical data, one hot encoding is used.The independent features are parameters which identify the presence of disease.The dependent feature is diabetes.The proposed method results are helpful to predict if the patient is diabetic or not.

3.2 Data Preprocessing:

Data discrepancies may be eliminated and accuracy improved by processing the data beforehand. We examine the feature for null values and eliminate them. Numeric numbers are used to encode the gender data.In order to divide the data and create the heatmap, this encoding is helpful.

3.3 Feature Selection:

Improving the precision of a machine learning model relies heavily on dimensionality reduction. We have reduced the dimensions. Combining characteristics allows one to choose the most significant one and then extract it. To find out how the attributes are related, we compute their correlation. A correlation coefficient is used to choose features. The connection between two variables is described by the correlation coefficient. There is a very low level of correlation between smoking history info, gender, current smoking history, never smoking, and other variables. Diabetes is inversely connected to the following variables: gender (other), smoking history (no info), and female gender. Diabetes is strongly associated with elevated levels of blood glucose and hemoglobin A1c.Accordingly, the smoking_history is removed. There is a 0.42 association between blood glucose levels and HbA1C values and a 0.40 correlation between the two.The coefficients of association between each attribute and the independent variable diabetes are shown below.From that we can infer the relationship.

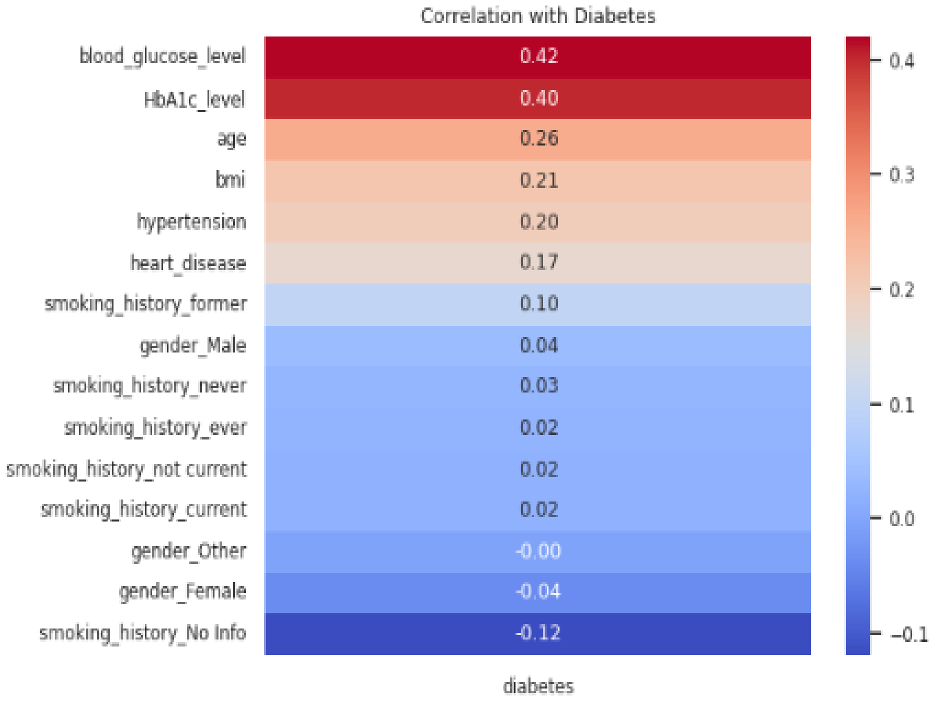

Fig.Correlation coefficient with diabetes

3.4 Split the training and testing data:

There are two sets of features: train data and test data.The test data set has an 80% size.Training is done with the last 20% of the data.

Logistic classifier:

An example of a supervised algorithm is LR.The input is a set of independent characteristics, and the output is a probability between zero and one.For binary classification, it is used.

Random Forest:

The Random Forest method is applicable to issues involving both classification and regression. All of the decision trees get its own unique subset of the data. Predictions of the production are made by averaging or voting.

Adaboost:

Converting weak learners into strong learners is what Adaboost does.Minimizing bias is the goal of this machine learning approach. Errors made throughout the training session are known as bias. The weak learner's error rate is inversely related to the value of the alpha parameter, which is different from other boosting strategies.

XGboost:

Extreme Gradient Boosting is what XGboost stands for. Furthermore, it transforms sluggish learners into ace learners. It can manage datasets of any size.

Multi Layer Perceptron:

Like other ANNs, MLPs are used to solve categorization difficulties.When training complex models with non-linear data correlations, this learning approach is used.

It is used to train the model once the data has been separated.Next, the model is fitted and then tested using testing data to see how well it predicts.Lastly, in order to choose the optimal model, performance measures for each model are computed.

Performance evaluators:

Accuracy, precision, recall score, and f1 score are the model's performance measures. The performance measures plays a crucial role for comparing different models and a model’s performance.

The accuracy refers to the outcome being correct. It doesn't point up the errors made by the model.

Precision is related to the selection of the output class correctly.

Recall scores means how it selects the positive that is the true sample.

The f1 score is calculated using both the above explained two values.

4. Results

Each model's performance metrics are computed once the classifiers have been assessed.

Model	LR	RF	MLP	Ada boost	XG boost
`Accuracy	96.05%	96.97%	95.6%	97.15%	97.01%
Precision	95.78%	96.90%	95.72%	97.10%	96.92%
Recall score	96.01%	96.97%	95.6%	97.15%	97.01%
F1 score	95.73%	96.75%	94.97%	96.95%	96.81%

Table-Comparison of classifiers

According to the performance metrics mentioned before, adaboost is the top model with a score of 97.15 percent.

All of this points to the fact that each model is accurate.Among these methods, logistic regression comes in at 96.05%, random forest at 96.77%, multi-layer perceptron at 95.6%, adaboost at 97.15%, and XGboost at 97.01%.

In order to create the learning curve, we plot the performance indicators' scores against the number of training samples.

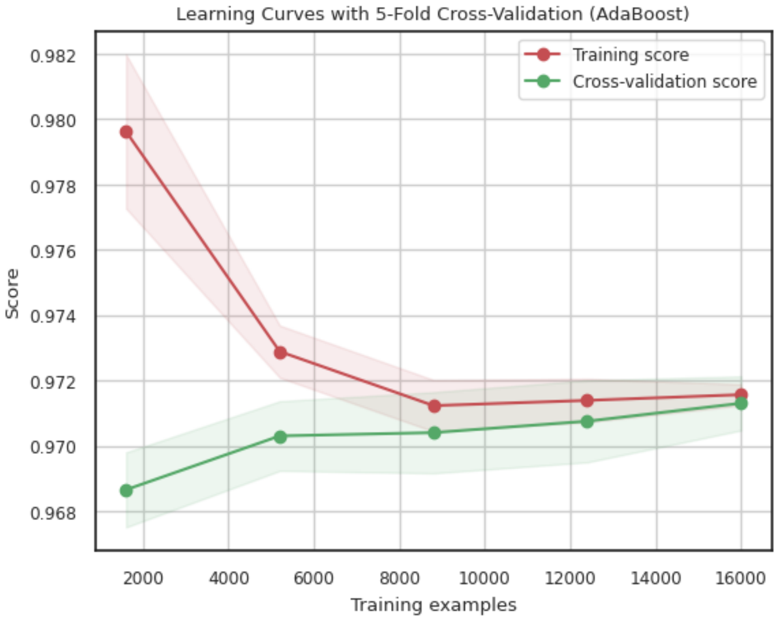

Fig.Learning Graph of AdaBoost

The selected model is tested using K-fold cross validation to see how well it handles the concealed data.Accuracy measures such as standard deviation and mean are computed.Averaging a precision of 97.13%, the standard deviation is 0.08%.The results show that the adaboost model performs well with concealed data.

Performance metrics are computed and tabulated once the models are tested.The best model was determined to be adaboost with a 97.15% accuracy rate after comparing it to others.

5. Conclusion and future enhancement

Several models are tested in order to find the most accurate one for diabetes prediction.After analyzing all of the models' performance measures, the best one is selected to serve as the adaboost.The adaboost model is validated using k-fold to see whether it is a generalized model.The model performs well on unseen data, with a mean accuracy of 97.13% and a standard deviation of 0.08%.The healthcare sector may benefit from early detection and preventative interventions for people with diabetes thanks to the diabetes prediction algorithm that uses machine learning

This research helps the healthcare sector by saving time and difficulties involved in the traditional system. The traditional system involves recording the patient data. The patient undergoes several test. They should wait until the reports are generated. The report shows the result with which the doctors conclude whether the patient is diabetic or not. The reports generated might be inaccurate because several reasons. The reasons include the performance of the instrument, environment and the person who keeps track of the record. Using Machine learning technique, we can overcome this issues.

This approach can be further enhanced by combining it with different models. Ensemble models and voting classifier can be used to select the best model which produces the accurate output. This approach can be extended and applied to various disease prediction which will be helpful to the patient and healthcare industry. Real time dataset can also be collected and the various algorithms can be evaluated using the performance evaluators. The early detection of disease plays a vital role in proceeding to treatments and prevention. Hyperparameter tuning can be done in each model to bring it as a more generalized model.

This research concludes that among various classifiers Adaboost was found to be the best through prior processing of data, encoding the data, reducing the dimensions of it. The reduction in the dimensions eliminates the inconsistencies in the data.

References

1.H. E. Massari, Z. Sabouri, S. Mhammedi and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," in Journal of ICT Standardization, vol. 10, no. 2, pp. 319-337, 2022

2.M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020

3.U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022

4.N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," in IEEE Access, vol. 9, pp. 103737-103757, 2021

5.E. A. Pustozerov et al., "Machine Learning Approach for Postprandial Blood Glucose Prediction in Gestational Diabetes Mellitus," in IEEE Access, vol. 8, pp. 219308-219321, 2020

6.R. Marzouk, A. S. Alluhaidan and S. A. El_Rahman, "An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System," in IEEE Access, vol. 10, pp. 105657-105673, 2022

7.S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques," in IEEE Access, vol. 7, pp. 1365-1375, 2019

8.A. Site, J. Nurmi and E. S. Lohan, "Machine-Learning-Based Diabetes Prediction Using Multisensor Data," in IEEE Sensors Journal, vol. 23, no. 22, pp. 28370-28377, 15 Nov.15, 2023

9.K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," in IEEE Access, vol. 12, pp. 16783-16793, 2024

10.D. K. Behera, S. Dash, A. K. Behera and C. S. K. Dash, "Extreme Gradient Boosting and Soft Voting Ensemble Classifier for Diabetes Prediction," 2021 19th OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2021

11.M. T. Islam, M. Raihan, N. Aktar, M. S. Alam, R. R. Ema and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020

12.H. A. Al-Jamimi, "Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction," in IEEE Access, vol. 12, pp. 62215-62233, 2024