INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 515-517

e-ISSN : 2583-1062

Impact Factor : 5.725

PREDICTION OF DIABETES USING GRADIENTBOOST AND ADABOOST

S. Vijayalakshmi[1], Mrs. A. Karmehala[2]

[1]B. Sc., Department of Computer Science, Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

[2]M.C.A., M.Phil., Department of Computer Science, Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

ABSTRACT

Diabetes affects kidney disease, eyesight loss and heart disease in addition to being the worlds largest cause of mortality. By assisting with precise disease diagnosis and treatment decisions, data mining tools lighten the burden on specialiats in the medical field. Better treatment outcomes will arise from early diabetes prediction. In this scope, a publicly available diabetes dataset, which includes 16 features that are collected from 952 people, was used to create predictive models. I apply two machine learning algorithm such as Gradient Boost and AdaBoost. Python is used to train the suggested method, and an actual dataset obtained from Kaggle is used for analysis. Additionally, the confusion matrix and performance metrics are used to assess how well the suggested mechanism performs. The Gradient Boost model outperforms the other models, according to the comparison between the two..

Keywords- Diabetes, Gradient Boost, AdaBoost

1. INTRODUTION

Diabetes mellitus is a chronic illness that is spreading quickly and affecting people of all ages. This crippling illness can result in blindness, renal failure, amputation, heart failure, and stroke, among other grave problems. Our systems turn the energy from food into glucose or sugar. The pancreas secretes insulin, which is what permits glucose to enter our cells. On the other hand, diabetes results from insufficient insulin production by the pancreas or from the body's inability to utilise the insulin that is generated. Consequently, glucose does not enter the cells and stays in the circulation. Type 1, type 2, and gestational diabetes are the three main forms of diabetes mellitus. When the pancreas is unable to create enough insulin, beta cells become insufficient, which results in type 1 diabetes. Although it can strike at any age, children and teenagers are the most prevalent age group for this form of diabetes to be

diagnosed. Excessive thirst, dry mouth, unexplained weight loss, impaired eyesight, frequent urination, and other symptoms are common with type 1 diabetes. People who with type 1 diabetes need daily insulin injections to keep their blood glucose levels under control, and they are also much more likely to develop heart disease.

The ailment known as type 2 diabetes is typified by the cells' improper response to insulin. The patient develops insulin resistance as a result. Type 2 diabetes affects over 90% of people with diabetes globally. Even while type 2 diabetes is usually thought to be less severe than type 1, it can still lead to health issues, especially when it comes to the kidneys, eyes, nerves, and small blood vessels. This type of diabetes was once mainly seen in adults, but it is becoming more common in youngsters.

In contrast, gestational diabetes is the third major form that affects pregnant women who have never had diabetes before and results in elevated blood sugar levels. Although type 2 diabetes can develop months or even years after giving delivery, up to 10% of women with gestational diabetes may experience its disappearance. Compared to the mother, the infant is more likely to experience irregular growth, breathing difficulties at delivery, or a higher chance of developing obesity and diabetes later in life. Leading the charge in the diabetes pandemic are China and India.In summary, the following tasks have been completed in relation to this work .

Using a real dataset that is gathered from Kaggle, the suggested mechanism is trained using Python for Gradient Boost and AdaBoost,

2. LITERATURE SURVEY

Sisodia et. al. in [1] applied three machine learning methods i.e. decision tree (DT), naïve based (NB) and support vector machine (SVM) on PIDD in order to predict the diabetes. Naïve bayes classifier was found to be 76.30% accurate. Han Wu et. al. in [2] applied data mining techniques (i.e. improved kNN and logistic regression) to accurately predict up to 95.42% the risk to an individual of developing type 2 diabetes. The modification was done by selecting value of initial seed point experimentally. The initial seed point was selected by conducting 100 experiments in which they selected smallest value of 'within cluster sum of squared errors.'

Nongyao et. al. in [3] compared four classification techniques i.e. decision tree, ANN, logistic regression and naïve bayes. Further bagging and boosting were applied on all and random forest was also included. The maximum accuracy achieved by all was in between 84% and 86% .

Choi et. al. in [4] applied machine learning algorithms on patients having history of non-diabetes having cardiovascular risk. Five years data has been collected in form of EMR from Korea University Guro Hospital. Then, machine learning methods were applied with 10-fold cross validation. Highest accuracy was obtained in logistic regression model.

Meng et. al. in[5] compared logistic regression, artificial neural network (ANN) and decision tree (DT) for identifying the risk of diabetes and prediabetes based on 12 risk factors which included education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee drinking, preference to salty foods, physical activity, and consumption of fish. DT was found to provide best results among the three methods.

3. METHODOLOGY

A. GradientBoost–

Gradient boosting is a machine learning technique that combines multiple weak predictive models (typically decision trees) to create a powerful predictive model. Iteratively improving the overall prediction accuracy is facilitated by training new models iteratively while focusing on the errors made by the previous models.Until a predetermined number of models are trained or a predetermined performance level is attained, this iterative process is continued. Gradient boosting is well-known for its capacity to manage intricate datasets and generate remarkably precise predictions across multiple areas, including classification and regression assignments.

B. AdaBoost–

AdaBoost is a machine learning algorithm that combines multiple weak classifiers to create a strong classifier. It is primarily used for binary classification tasks but can be extended to multi-class problems. Iteratively trains weak classifiers on different data subsets, assigning higher weights to misclassified samples. AdaBoost is known for handling complex datasets and adapting to difficult instances.

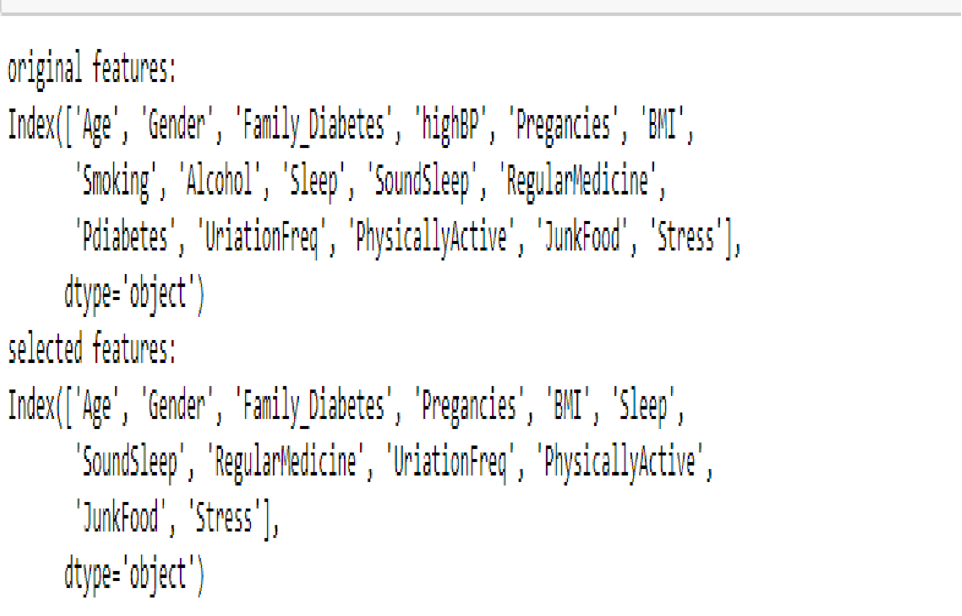
4. EXPERIMENT OF RESULT

A. Dataset Used

Downloaded the Diabetes dataset from Kaggle. This dataset contains 952 patient cases' worth of medical data. Numerically valued attributes are also included in the dataset; the value of a class corresponds to a diabetes test result, while the value of another class corresponds to a diabetes test result.There are 16 attributes and 952 instances in this collection. 80 % of the training data and 20 % of the testing data were separated into two sections of the sample. The proposed technique is implemented using Jupyter notebook and Python. The language Python is open-source.The packages Numpy, Pandas, Scikit-Learn, Matplotlib, etc. were utilised in this paper. The preferred language for data processing software is Python .

B. Preprocessing

Feature selection: During model building, this technique is used to choose the most pertinent attributes. It lessens the prediction model's intricacy. Using the Python programming language, I used feature selection with variance threshold in the study to create a model that only 12 features includes the most important features.



```
original features:
Index(['Age', 'Gender', 'Family_Diabetes', 'highBP', 'Pregancies', 'BMI',
       'Smoking', 'Alcohol', 'Sleep', 'SoundSleep', 'RegularMedicine',
       'Pdiabetes', 'UriationFreq', 'PhysicallyActive', 'JunkFood', 'Stress'],
      dtype='object')
selected features:
Index(['Age', 'Gender', 'Family_Diabetes', 'Pregancies', 'BMI', 'Sleep',
       'SoundSleep', 'RegularMedicine', 'UriationFreq', 'PhysicallyActive',
       'JunkFood', 'Stress'],
      dtype='object')
```

Fig. 1. Feature Selection

Once data preprocessing and feature selection are completed, the next step is model evaluation, which assesses the performance of the trained model. Evaluation metrics include the Confusion Matrix, which provides insights into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as well as accuracy, precision, recall, and F1 score. Table 1 shows that the precision, recall, and F1-score values of GradientBoost and AdaBoost.. Accuracy measures the overall correctness of predictions, while precision quantifies the quality of positive predictions. Recall assesses the model's ability to correctly identify positive instances, while the F1-score provides a balanced measure of a model's overall performance, considering both precision and recall . These evaluation metrics collectively offer valuable insights into the model's effectiveness and help guide further optimization efforts. Table 2 shows that the accuracy values of GradientBoost are greater than AdaBoost..

Table -1 Comparison between GradientBoost and AdaBoost

	GradientBoost	**AdaBoost**
Precision	**0.91**	**0.82**
Recall	**0.97**	**0.87**
F1-score	**0.94**	**0.85**

Table -2 Experiment Result

Algorithm	**Accuracy**
GradientBoost	**0.916**
AdaBoost	**0.88**

5. CONCLUSION

An ensemble can perform better and make more accurate predictions. One of the world's biggest health problems is diabetes. Results will be better if diabetes is predicted early. With the aid of ensemble approaches, a machine learning algorithm for early stage diabetes prediction is presented in this study.used the AdaBoost, and Gradient Boost approaches to predict the presence of diabetes. Python is used in the suggested mechanism's implementation. After using the Diabetes Dataset, the data was preprocessed, the training and testing sets were divided, and the accuracy of the ensemble algorithm was predicted.Compared to other models, the Gradient Boost model has a high accuracy rate of 92%. Large real-time datasets will be gathered and used in the future.

6. REFERENCE

[1] Sisodia, D., Sisodia, D. S. (2018) “Prediction of diabetes using classification algorithms.” Procedia computer science 132: 1578-1585.

[2] Wu, H., Yang S., Huang, Z., He, J., Wang, X. (2018) “Type 2 diabetes mellitus prediction model based on data mining.” Informatics in Medicine Unlocked 10: 100-107.

[3] Nai-arun, N., Moungmai, R. (2015) “Comparison of classifiers for the risk of diabetes prediction.” Procedia Computer Science. 69: 132-142.

[4] Choi, B.G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., Noh, Y. K. (2019) “Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks.” Yonsei medical journal 60 (2): 191-9.

[5] Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q., Liu, Q. (2013) “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.” The Kaohsiung journal of medical sciences 29 (**2): 93-9.**