Diabetes Prediction Using Ensemble Algorithms Among Women

S.Ramadevi[1], M.Shanmuga Eswari[2]
[1, 2] Dept of Computer Science
[1, 2] Sri Kaliswari College (Autonomous) Sivakasi

Abstract- *All times diabetes as diabetes mellitus.Diabetes is a sickness that happens while your blood glucose is just too excessive.In this paper,propose a diabetes prediction using ensemble strategies. Practice three ensemble techniques inclusive of Random Forest,Bagging and Adaboosting. Random Forest gives 86% accuracy, bagging gives 92% accuracy.Adaboost gives 88% accuracy.The bagging technique offers great accuracy evaluate than different strategies. The dataset is accumulated from PIMA Indian Diabetes.Bagging gives best ROC range 0.946.*

Keywords- Diabetes, Ensemble Algorithm, Diabetes Prediction, Machine Learning, Prediction.

I. INTRODUCTION

Diabetes is a chronic condition that affects the way the bodyproduces a hormone called insulin. The pancreas is the organ that produces insulin. Diabetes affects the pancreas ability to produce insulin causing it to make too much too little none at all insulins number one job is to move sugar out of the bloodstream into cells.Cells use sugar for energy or fuel so the body's cells can work properly carbohydrates are a type of sugar.when eat carbohydrates your body converts it through the digestive process into sugar and dumps it into the bloodstream insulin removes the sugar from the bloodstream into cells.if your pancreas does not produce enough insulin to process the sugar starts to build up in your blood. There are three main types of diabetes: type 1, type 2, and gestational diabetes (diabetes while pregnant).

1.1 Types of Diabetes

Type1: Type 1 diabetes is thought to be caused by an autoimmune reaction (the body attacks itself by mistake). This reaction stops your body from making insulin. Approximately 5-10% of the people who have diabetes have type 1. Symptoms of type 1 diabetes often develop quickly. It's usually diagnosed in children, teens, and young adults. If you have type 1 diabetes, you'll need to take insulin every day to survive. Currently, no one knows how to prevent type 1 diabetes.

Type2: With type 2 diabetes, your body doesn't use insulin well and can't keep blood sugar at normal levels. About 90-95% of people with diabetes have type 2. It develops over many years and is usually diagnosed in adults (but more and more in children, teens, and young adults). You may not notice any symptoms, so it's important to get your blood sugar tested if you're at risk. Type 2 diabetes can be prevented or delayed with healthy lifestyle changes, such as:

- Losing weight.
- Eating healthy food.
- Being active.

Gestational diabetes: It develops in pregnant women who have never had diabetes. If you have gestational diabetes, your baby could be at higher risk for health problems. Gestational diabetes usually goes away after your baby is born. However, it increases your risk for type 2 diabetes later in life. Your baby is more likely to have obesity as a child or teen and develop type 2 diabetes later in life.

II. LITERATURE SURVEY

In today's medical world, doctors can use it to quickly and accurately interpret diseases. Because of that we can use machine learning to prevent the death by making an artificial inteligent model that can predict diabetes disease and the method that be used is comparison between the KNN and Naive Bayes algorithms to see which algorithm suit the best for diabetes prediction. The study concluded by comparing two k-Nearest Neighbor algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in the dataset using supervised machine learning. According to the results of our experiments and evaluating alghorithm using Confusion Matrix, the Naive Bayes algorithm outperforms KNN[1]. In this paper, we propose a diabetes prediction model using data mining techniques. We apply four data mining techniques such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. In logistic regression, the accuracy is high, i.e., 82.46%, in comparison to other data mining techniques[2]. The main data mining algorithms discussed in this paper are EM algorithm, K means, C4.5 algorithm, Genetic algorithm and SVM. It is found that the genetic algorithm gives a better performance over five data

mining algorithm[3]. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. . Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms[4]. A set of operation was led to assess this accuracy regarding a set.of data.mining procedures including.Decision.Trees (j48), Naïve Bayes, and hybrid proposed method of decision-tree and SVM into diabetes disease diagnosis. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 94%[5].

III. METHODOLOGY

3.1 Ensemble Methods

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. Ensemble methods,including bagging,boosting,and stacking.we will explore random forests.

Bagging: Bagging, the short form for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models.

AdaBoost: Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series.Adaboost is a one type of boosting method.AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”.

Random Forest: The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “the random subspace method” generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features.

3.2 Dataset Used

PIMA diabetes dataset downloaded from Kaggle. This dataset includes the medical information for 768 cases of female patients. The dataset also includes eight numeric-valued characteristics, where the value of one class is treated as a diabetes test result of type 0 and the value of another class is treated as a result of type 1 diabetes testing.In this dataset includes 768 Instances and 9 attributes. The sample was divided into two parts, one with 80% of the data for training and 20% of the data for testing. Python and Jupyter notebook are used to execute the suggested mechanism. Python is a open-source language.In this paper,used the packages such as Numpy,Pandas,Scikit-Learn,Matplotlib,etc.. Python is the language of choice for data processing software.

3.3 Preprocessing

Data Imputation: Data imputation is a method for retaining the majority of the dataset’s data and information by substituting missing data with a different value.In this paper,there are many zero values in the dataset,so replacing with median values.

Label Encoding: It refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Feature Scaling or Standardization: It is a step of Data Pre Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

IV. RESULT

Accuracy

Accuracy is used in classification problems to tell the percentage of correct predictions made by a model. Accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions made.Calculate it by dividing the number of correct predictions by the total number of predictions.

Accuracy=Number of Correct Prediction/ Total Number of Prediction

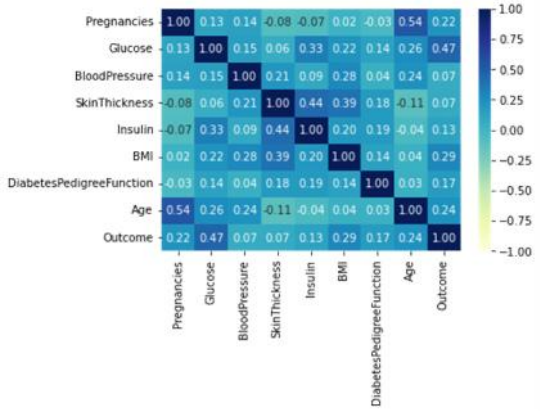

Fig.1: Heatmap for an attributes

Table1. Performance analysis for Proposed Algorithms

Techniques	**Accuracy**
Bagging	92%
Random Forest	86%
Adaboost	88%

From above comparison,.Bagging accuracy is high than other techniques.

Table2. Performance analysis for Existing Algorithms

Existing Algorithms	**Accuracy**
Logistic Regression	82.46%
Naïve Bayes	76.30%
Genetic algorithm	78.1%

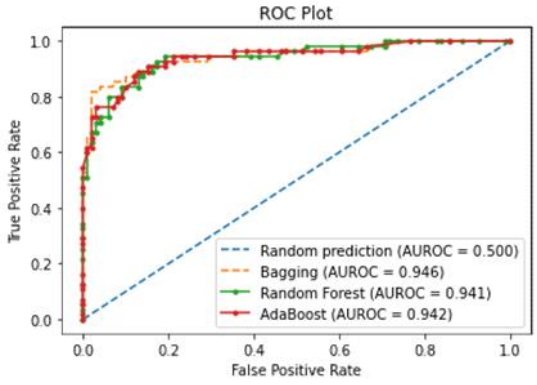

Fig2:Roc Curve for the ensemble techniques

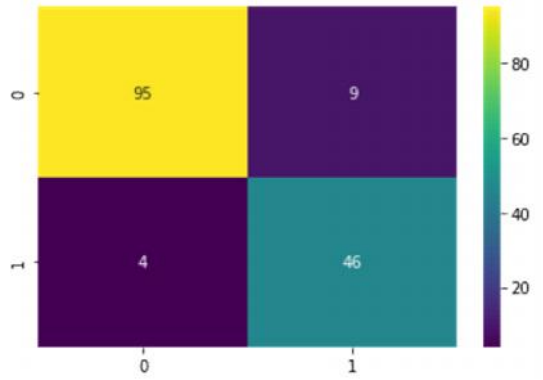

Fig3:Confusion Matrix for Bagging

5. Conclusion and Future Work

An ensemble can make better prediction and achieve better performance. Diabetes is a major health challenge in the world. Early prediction of diabetes will result in improved results. This paper presents a diabetes prediction among womens model with the help of ensemble techniques.Applied Bagging, AdaBoost and Random Forest techniques to predict diabetes . The proposed mechanism is implemented using Python.

To used PIMA Diabetes Dataset ,then preprocess the data,splitted training and testing data,and predicting the accuracy for used ensemble algorithm.In the Bagging model, the accuracy is high 92% as compared to other models. In the future,large real time dataset will be collected and implemented.

REFERENCES

[1] Muhammad Exell Febrian,Fransiskus Xaverius Ferdinan, Gustian Paul Sendani, Kristien Margi Suryanigrum,Rezki Yunanda, “Diabetes prediction using supervised machine learning”,Procedia Computer Science ,2023.

[2] R. Rastogi and M. Bansal,” Diabetes prediction model using data miningtechniques”,Measurement: Sensors,2023.

[3] N.Nandhini and K. Thangadurai,”Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus”, International Journal of Scientific & Engineering Research, 2016.

[4] Deepti Sisodiaa and Dilip Singh Sisodiab, “Prediction of Diabetes using Classification Algorithms”,Procedia Computer Science,2018.

[5] Mohammed Layth Zubairi Alkaragole and Sefer Kurnaz,“Comparision of datamining techniques for predicting diabetes or prediabetes by risk factors”,International Journal of Computer Science and Mobile Computing,2019.