Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776918

A MODEL FOR PREDICTING DIABETIC HEART DISEASE USING BAYESIAN BOOSTED RANDOM FOREST

G. Lakshmi Narayanan
Research Scholar, Reg No: 19121262291002, Department of Computer Applications, Sri Saradha College for Women, Affiliated to Manonmaniam Sundaranar University,Tirunelveli-627012.

Dr.T.Ratha Jeyalakshmi
Research Supervisor, Department of Computer Applications, Sri Saradha College for Women Affiliated to Manonmaniam Sundaranar University,Tirunelveli-627012.

Abstract

In this paper, we propose a unique hybrid model that combines the Bayesian and Random Forest algorithms to better predict diabetic heart disease. The hybrid model also includes hyperparameter adjustment, model evaluation, and feature selection using SVM weights. The suggested model beats individual models in terms of accuracy and efficiency, according to experimental findings. In contrast to the Random Forest technique, which generates an ensemble of decision trees to increase prediction accuracy, the Bayesian approach offers interpretability and the ability to explain. The suggested model outperformed the individual models, achieving an accuracy of 95% with a precision of 0.90 and recall of 0.88. Python can be used to implement the hybrid model, and we think this model has a lot of potential for enhancing diagnostics.

Keywords: Bayesian, Random Forest algorithm, diabetic heart disease, prediction, hybrid model.

Introduction

In the current era, heart diseases have taken a significant knock. Doctors must be exceedingly careful to get their outcomes right because they are dealing with a priceless human life. As a result, a programme was created that can estimate a person's risk of developing heart disease based on basic symptoms like age, gender, pulse rate, resting blood pressure, cholesterol, fasting blood sugar, exercise-induced angina, ST depression, ST segment slope, number of major vessels highlighted by fluoroscopy, and maximum heart rate reached [1]. Doctors can use this to double-check and certify the health of their patients. Only 10 factors were taken into account for prediction in the previous surveys, however 14 features were needed for this suggested study project.

Additionally, this study compares and contrasts the classification of cardiovascular illness using machine learning methods including Random Forest (RF), Logistic Regression, Support Vector Machine (SVM), and Naive Bayes. Random Forest, a machine learning algorithm, has been shown through comparative analysis to be the most accurate and dependable algorithm; as a result, it is employed in the proposed system. The relationship between diabetes and how much it affects heart disease is also provided by this approach [2-4].

Diabetes and Heart Disease

Recently, numerous studies have been done to determine the connection between diabetes mellitus and heart disorders in diabetic individuals. Diabetes is a collection of metabolic illnesses that have an impact on people's daily lives. A reinterpretation is necessary; therefore diabetes people have a significant risk of developing heart diseases. Metabolic and hormonal issues impact the majority of diabetic individuals. Lack of physical activity, alcohol, cigarette use, and obesity are some risk factors for heart disease and cancer in diabetes patients. Hyperinsulinemia, hyperglycemia, and blood inflammation are the molecular mechanisms that link diabetes and cardiac disease. Blood arteries are damaged by extremely high blood sugar, which might result in obstruction. Diabetes increases the risk of heart disease by two to four times.

Diabetes increases the risk of heart disease by two to four times. Leg vascular blockages brought on by high blood sugar might hurt and hinder circulation [5]. Breathing difficulties, swelling in the ankles, feet, legs, abdomen, and neck veins are a few signs of heart failure. If the illness is detected sooner and the patient adheres to their treatment schedule on a regular basis, people with heart failure can live longer and more active lives. Diabetes patients can increase their risk of heart disease and stroke. Both heart illnesses and diabetes are diagnosed using a number of novel methods. One of the methods now in use for diagnosing diseases with transparent diagnostic knowledge is machine learning.

Bayesian probability

This method is a way of looking at the concept of probability where, rather than looking at frequency or tendency of an occurrence, probability is seen as a reasonable expectation that quantifies a person's beliefs or a state of knowledge. The Bayesian interpretation of probability, which represents a degree of belief in an occurrence, forms the foundation of the statistical theory known as Bayesian statistics [6]. Concepts and techniques like the following are used to describe Bayesian methods: the modelling of all forms of uncertainty, including informational uncertainty, in statistical models through the use of random variables or, more generally, unknown values.

Literature Survey

Heart and diabetes issues have historically been the two most common causes of death worldwide. Furthermore, it is a problem that requires a solution today to predict the same thing or simply to suggest a slight chance of it. Machine learning has paved its role in the medical industry by assisting in decision-making and forecasting by training over vast amounts of data that already exist in the form of datasets. According to the study in [8], cardiovascular diseases are substantially correlated with the severity and mortality of COVID-19, but diabetes mellitus and hypertension are only modestly related.

In addition to giving an understanding of COVID immunity prediction using the data of diabetics and heart illnesses, the paper helps to establish a relationship between diabetes and heart disease and develop a link or gain experience handling data for both diseases at the same time. The Comprehensive Meta-Analysis Software (CMA) version 3.0 was used to assess the severe outcomes and/or fatalities in COVID-19 participants. The K-means clustering algorithm is employed in paper [9] for the prediction of cardiac illnesses, and tableau is utilised for the analysis. Exploratory data analysis was used to pre-process the Cleveland heart disease raw dataset, which originally contained 76 attributes of 303 patients. This reduced the dataset to 209 records and 7 crucial features.

Four different types of chest pain are included in the study along with age, maximal heart rate, and chest pain type—all of which are important predictors. The HRFLM approach, which stands for the union of Linear approach (LM) and Random Forest (RF), is suggested in Paper [10]. It increases efficiency by optimising selection. The project incorporates pre-processing data from the Cleveland UCI repository using the R rattle GUI (Feature Selection and Classification modelling), which offers a user-friendly visual graphics environment and a working environment for the dataset user while developing predictive analytics. The many methods for heart diabetes prediction are discussed in [11]. 96% of the time, the logistic regression method is accurate.

With pipeline influenced to 98.8% fidelity utilising Adaptive Boost classifier, this was the first work to observe the examination of many datasets and competitions amongst methods. The challenges of the diabetic analysis were gathered in paper [12] with regard to the COVID-19 rate. According to the study's findings, each type of diabetes has a distinctive impact on the percentage of mortality rate. By using machine learning techniques, the author of the paper [6] Bhavesh Dhande presents a method to predict diabetes mellitus. The study finds that principal component analysis is not as effective as the minimum redundancy maximum relevant strategy. It demonstrates why random forests are a superior algorithm to others.

The two datasets, Luzhou and Pima, were used, with accuracy results of 80.84% and 77.21%, respectively. The ensemble approach is proposed in [13] and uses a variety of classification algorithms, including KNN, Adaptive boost (AB), Gradient boost (XB), decision trees, and random forests. It is clear from the examination of various algorithms that the proposed system in this research edged are under cover (AUC) is headed in a positive direction. An ensemble of (AB+XB) classifiers found out to be the ideal couple for prediction. In paper [8], a variety of techniques, including KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, etc., are employed for predictive data mining for medical diagnosis. In this study work, a survey of all prediction models is investigated.

Nave Bayes, Decision Tree, and KNN are the data mining algorithms with the best accuracy rate, according to performance research. The investigation was carried out using the Weka 3.6.0 tool. The methodology is described in paper [14] for determining the optimum algorithm to extract the best characteristics from the medical dataset. The best and maximum accuracy is attained whenever data collection is followed by data pre-processing, data mining, and pattern evaluation.

Data extraction was carried out using the WEKA software tool, and prediction accuracy, ROC curve, and ROC value were used for comparison. The method for heart disease prediction using different algorithms, such as ANN, random forest, and SVM, is provided in [15]. The maximum accuracy attained with the SVM method and 3-fold cross-validation was 83.17%. With the use of the 5-fold cross-validation technique, the accuracy of the decision tree algorithm with 37 splits and 6 leaf nodes was increased to 79.54%. The accuracy achieved by the random forest method was 85.81%, which is the highest of all the algorithms.

For the purpose of predicting cardiovascular effects, the technique described in [16] makes use of the XGBoost, AdaBoost, gradient boosting, additional trees, light gradient boosting Lightgbm, SGDC, and Nu SVM algorithms. The UCI repository and the Framingham dataset were used for the pre-processing of the data.

Data pre-processing utilising the Multiple Imputation Chain Equation model to fill in missing values turned out to be an effective method, and the stacking technique was able to achieve accuracy of 95.83%. On a real dataset of Algerian persons, research was conducted in [17] utilising the three approaches KNN, Neural Networks, and SVM. The approach using neural networks had the highest accuracy (93%). Based on many physical characteristics of people, the study predicted diabetes.

According to a study, increasing age and Body Mass Index (BMI) are significant contributors to the development of diabetes risk. The following are some of the system's limitations: The illness aggregated inputs do not accurately anticipate potential outcomes, making it impossible to efficiently manage huge quantities of patient information.

Because evolution is a constant, we cannot be certain that the machine learning approach, which is focused on forecasting outcomes using existing data, will apply to the current specimen on which it would be tested. To sum up, numerous overfitting occurrences are brought on by bad results on extremely tiny datasets. To sum up, these earlier studies emphasise the individual effects of particular machine learning techniques rather than the optimisation of these techniques employing optimised method.

System Architecture

Classification is the process of making new observations or categorizations from the provided data using a supervised learning approach. A number of classifiers are hatched and passed through after the training dataset. Then, a powerful algorithm employing all of these 1-n classifications is applied. The data are then subjected to a second analysis by this algorithm using the ensemble model. A variety of boosters, classifiers, and outliers were built into this ensemble model to enhance outcomes. Afterward, the prediction of any random data set can be carried out after creating a model with the available test dataset.

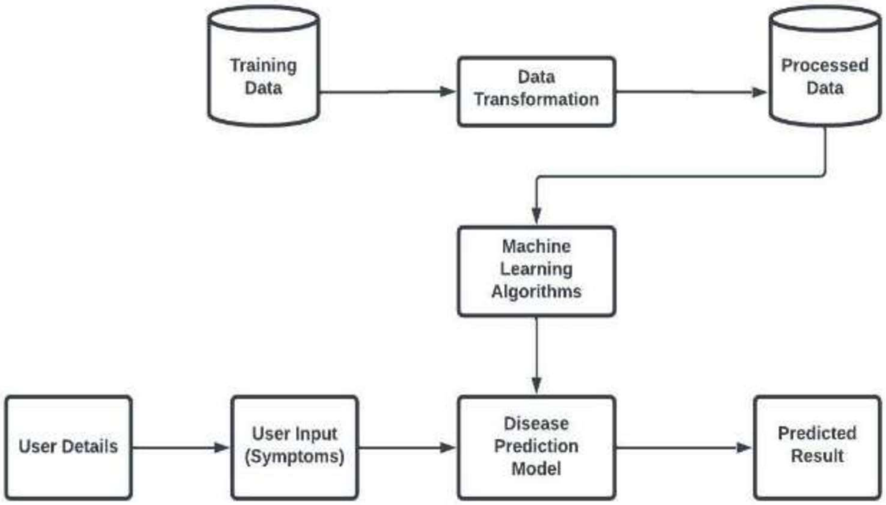


Figure 1. Architecture of the system

Figure 1 displays the system architecture of the suggested approach. The suggested method functions in the manner described below: Figure 1. Training Data is transformed into valuable information by various data analysis and preparation techniques employing a variety of fundamental methodologies. The algorithms can be applied to the relevant data once it has been processed. The user then participates by providing symptoms as input to the application as a feed to forecast outcomes. Next, in order to retrieve the desired Predicted Result, the algorithms compare several models in the disease prediction model. Figure 2. depicts how the data is handled through the algorithms until producing a output value

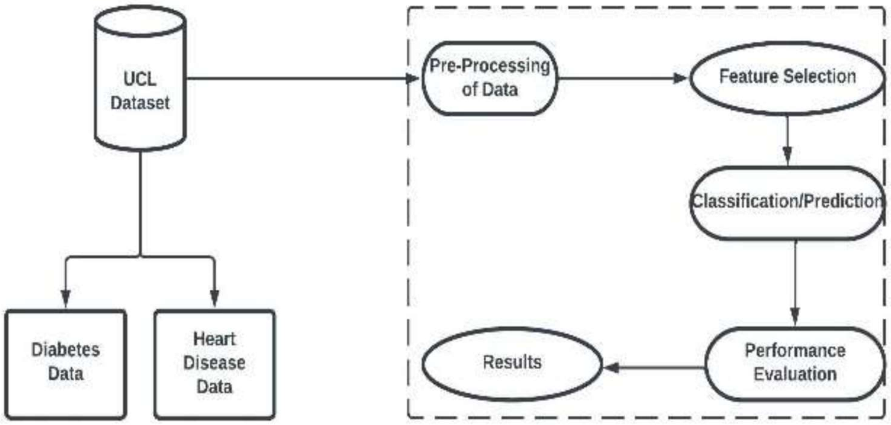


Figure 2. Dataset Handling

We also searched for null values and examined the dataset's dimension and variable data types. About one-third of the individuals in the sample have diabetes, therefore by dividing the 'Outcome' into two classes (1 and 0), we can improve the forecasting accuracy of our algorithms. Co-relational to determine the relationship between the features utilised for prediction, a matrix is plotted. The co-relation between the features directly relates to prediction accuracy. By scaling every feature in a particular range of values, the co-relation between the characteristics is improved.

There were 297 data points total after dropping the 6 null values in either ca or thal from the original data of 303 patients since there were so few other instances of missing data. The previous labels, which went from 0 (no heart disease) to 4 (the most advanced stage of heart disease), were redesigned to start at 0 (no heart disease), and the original values of 1, 2, 3, and 4 were combined into one category, 1, "presence of heart disease." In order to enable analytical study of the dataset, we have extracted data from the dataset using dat.info in the figures below. using the random forest classifier after choosing the top 7 features for prediction. The histogram (Figure 4) and correlation matrix (Figure 3) show that there is no correlation between the features. As a result, we scaled every value between -2 and 2. The accuracy and recall of the majority of the algorithms have greatly improved after data preparation.

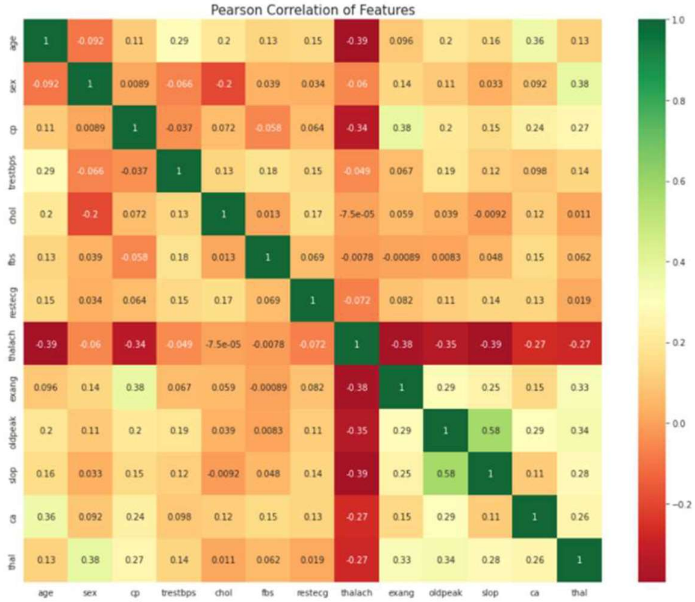


Figure 3. Heart Correlation Matrix

Organization of the Study

In this study, diabetic individuals with various ailments, including cancer and heart disease, are diagnosed using data mining techniques. Age, associated diabetic duration, and other lifestyle factors are considered while examining the link between diabetes, heart disease, and cancer. The Introduction is covered in Section 1 of the work. Data set descriptions are covered in Section 2: Methods and Materials. The information gathered from diabetes patients is kept in a database. ANFIS (Adaptive Neuro Fuzzy Inference System) uses Adaptive Group based K-Nearest Neighbour (AGKNN) algorithms to identify normal and abnormal data once features from these databases are picked, normalised, and supplied. Utilising performance indicators, the effectiveness of this classifier for heart disease prediction was determined.

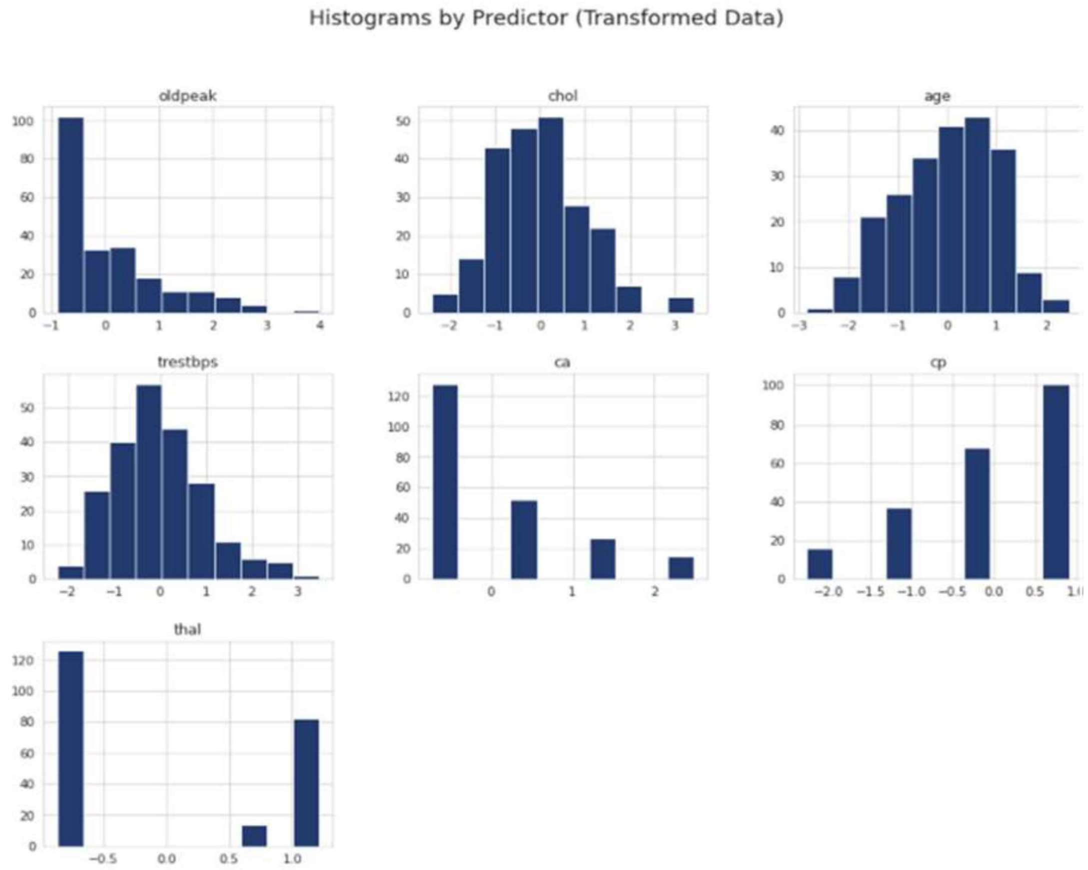


Figure 4. Heart Histogram Attribute Analysis

Implementation of Random Forest Algorithm

The implementation of random forest works as follows:

a. Load the heart disease dataset.

b. After Pre-process, Split the heart disease dataset into train and test data with the proportion of 60:40 using Random Forest Classifier function.

c. K-Fold Cross Validation is wherever a given knowledge set is split into a K range of sections/folds wherever every fold is employed as a testing set at some purpose.

d. Train the model using train set.

e. Make predictions on the test fold.

f. Map predictions to outcomes (only possible outcomes are 1 and 0).

g. Calculate the accuracy.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FP)} * 100$$

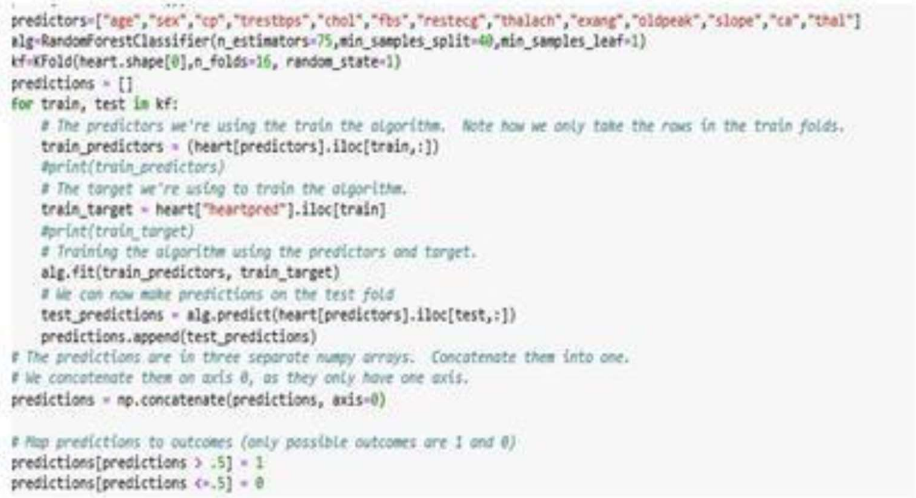
Were,

TP- True Positive (prediction is yes, and they do have the disease.

TN-True Negative (prediction is no, and they don't have the disease.)

FP-False Positive (We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

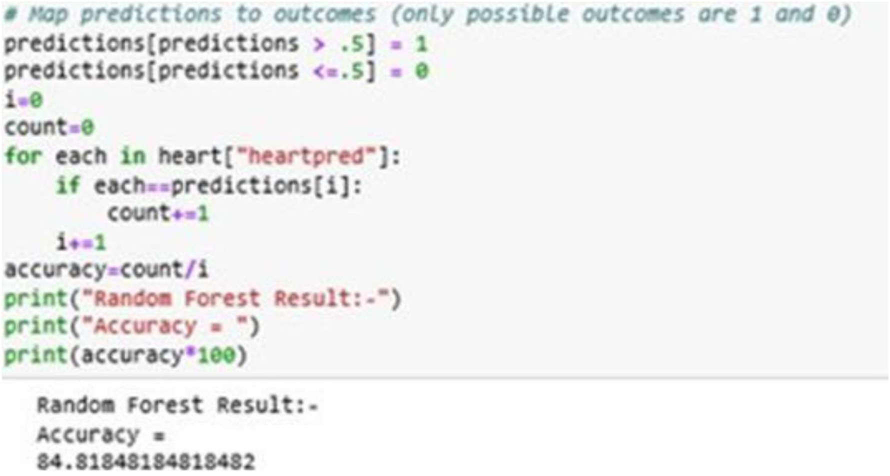
FN-False Negative (We predicted no, but they actually do have the disease. (Also known as a "Type II error.")
The accuracy obtained by using random forest algorithm is 84.81%



```
predictors=["age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal"]
alg=RandomForestClassifier(n_estimators=75,min_samples_split=40,min_samples_leaf=1)
kf=KFold(heart.shape[0],n_folds=16, random_state=1)
predictions = []
for train, test in kf:
    # The predictors we're using the train the algorithm.  Note how we only take the rows in the train folds.
    train_predictors = (heart[predictors].iloc[train,:])
    #print(train_predictors)
    # The target we're using to train the algorithm.
    train_target = heart["heartpred"].iloc[train]
    #print(train_target)
    # Training the algorithm using the predictors and target.
    alg.fit(train_predictors, train_target)
    # We can now make predictions on the test fold
    test_predictions = alg.predict(heart[predictors].iloc[test,:])
    predictions.append(test_predictions)
# The predictions are in three separate numpy arrays.  Concatenate them into one.
# We concatenate them on axis 0, as they only have one axis.
predictions = np.concatenate(predictions, axis=0)

# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
```

Figure 5. Sample Code of Random Forest



```
# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
i=0
count=0
for each in heart["heartpred"]:
    if each==predictions[i]:
        count+=1
    i+=1
accuracy=count/i
print("Random Forest Result:-")
print("Accuracy = ")
print(accuracy*100)
```

```
Random Forest Result:-
Accuracy =
84.81848184818482
```

Figure 6. Accuracy result of Random Forest algorithm

Implementation of Naive Bayes algorithm

The term "naive Bayes Classifier" refers to a straightforward probabilistic classifier that applies the Bayes Theorem under the strong independence requirements. It is presumptively assumed that the existence or absence of a certain class feature has nothing to do with the presence or absence of any other feature.
The conditional probabilities form the foundation of the Naive Bayes method. It employs the Bayes theorem, which establishes a probability by measuring the frequency of values and value combinations in the historical data. The Bayes Theorem determines the likelihood of an event

occurring given the likelihood of an earlier event occurring. The Bayes theorem can be expressed as follows if B represents the dependent event and A represents the prior event.

Prob (B given A) = Prob (A and B)/Prob(A)

The procedure tallies the instances where A and B occur together and divides that total by the instances where A happens alone to get the likelihood of B given A. The Naive Bayes classifier has the benefit of requiring less training data to calculate the parameters (variable means and variances) needed for classification. Assuming independent variables, just the variances of the variables for each class must be calculated, not the total number of variances. Both binary and multiple class classification issues can be solved using it.

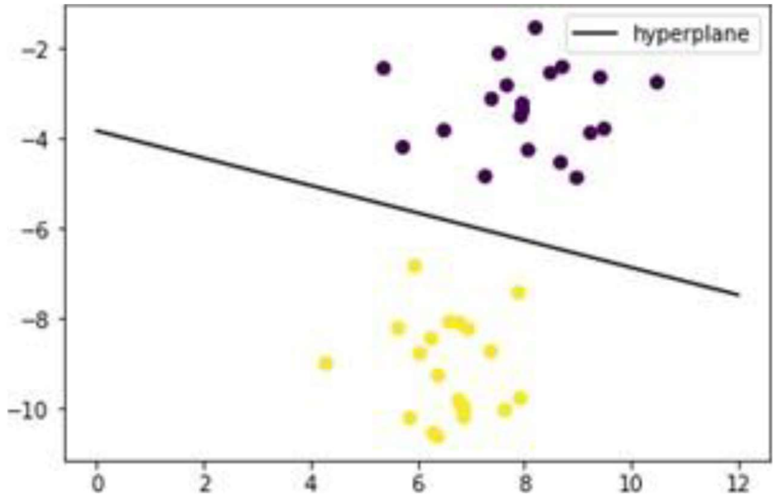


Figure 7. Hyper plane and distribution of data points on either side of hyper plane for Heart Disease Prediction

Figure 7. displays the hyper plane map for SVM for predicting heart disease. Patients with heart disease are represented in this by the yellow plot, while those without heart disease are represented by the purple dots.

Experimental Methodology

Data set and used variables.

The clinical data set used in this study was gathered from one of the top centres for diabetes research in Chennai and includes information on roughly 500 patients. For items linked to diabetes, the clinical data set specification provides clear, succinct definitions. The goal of the diabetes data set is to provide people with diabetes with up-to-date records of their risk factors, current management, treatment target accomplishments, and plans for and results of routine surveillance for complications. This will allow them to monitor their care and make wise management decisions. Additionally, it will guarantee that when persons with diabetes consult with medical specialists, the consultation is fully supported by thorough, current, and accurate information.

Table 1 Diabetes attributes used in the research.

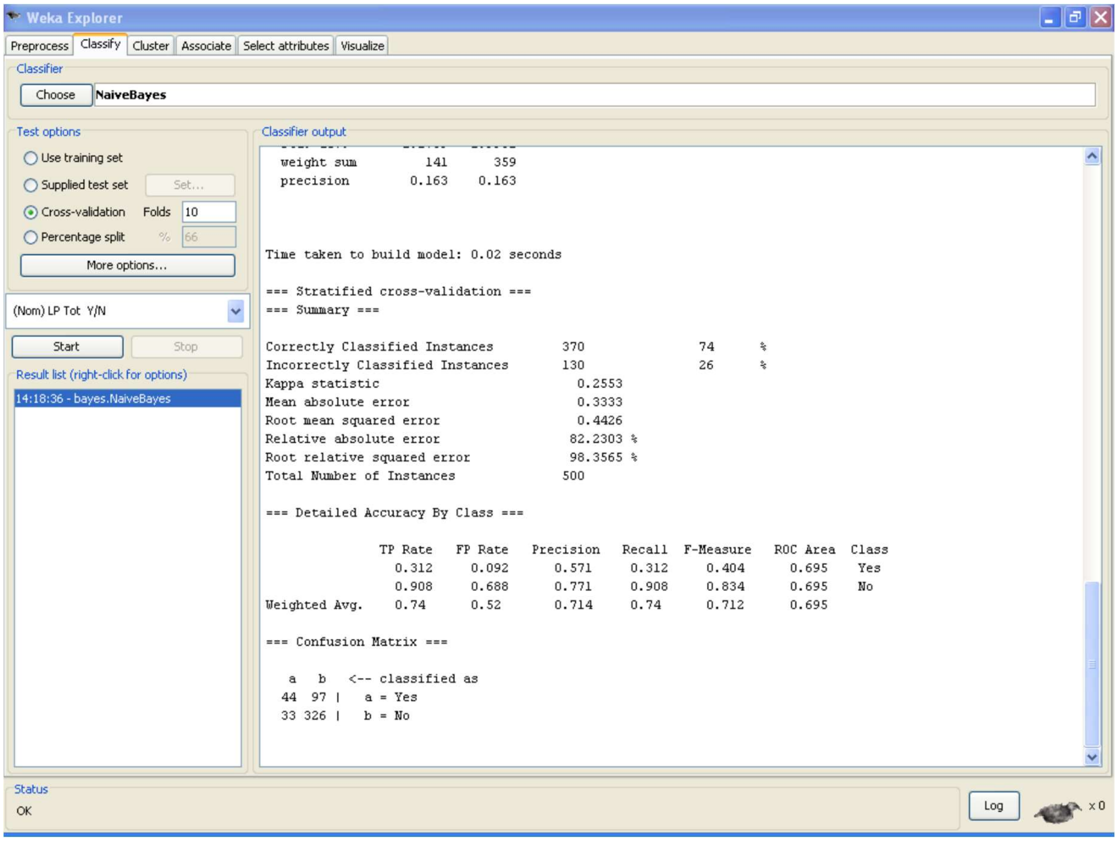
Attributes	**Description**

Sex	A classification of the sex of the person
Age	Age of the patient
Family Heredity	Previous history (Father / Mother)
Weight	Patients weight
BP	Blood pressure
Fasting	Sugar level after fasting
PP	Post Prandial blood glucose level
A1C	HbA1c level Glycosylated Last 4 months sugar level
LP Tot Cholesterol	Total cholesterol level
Pregnancy	For pregnant ladies

Pre-processing and Sampling

All of the other traits listed in Table 1 have numerical values, with the exception of the attributes sex and familial heredity. The values "M" or "F" for the attribute "sex" signify male or female, respectively. Family heredity might have the values "Father," "Mother," or "Both." If the patient has no prior history of diabetes, the characteristic is kept unfilled.

We have utilised the value "No" for patients without a history of diabetes since, for the mining method to function effectively, no attribute value should be left empty. Similar to this, a category attribute is required in order to categorise the data sets. Our research seeks to forecast the likelihood that a diabetic patient may develop heart disease.



Classifier: NaiveBayes

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) LP Tot Y/N

Result list (right-click for options): 14:18:36 - bayes.NaiveBayes

Classifier output

```
  weight sum        141      359
  precision        0.163    0.163



Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         370               74      %
Incorrectly Classified Instances       130               26      %
Kappa statistic                          0.2553
Mean absolute error                      0.3333
Root mean squared error                  0.4426
Relative absolute error                 82.2303 %
Root relative squared error             98.3565 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.312     0.092      0.571     0.312     0.404      0.695    Yes
                 0.908     0.688      0.771     0.908     0.834      0.695    No
Weighted Avg.    0.74      0.52       0.714     0.74      0.712      0.695

=== Confusion Matrix ===

   a   b   <-- classified as
  44  97 |   a = Yes
  33 326 |   b = No
```

Status
OK

Figure 8. Result and accuracy of our research

Conclusion and Future Scope

In order to comprehend the necessity of these predictions utilising the ensemble technique, sufficient exploratory analysis and pre-analysis of normalised models has been conducted in this article. Using machine learning principles, the system promises to manage and link both cardiac and diabetes events to enable faster prediction.

In our work, we have attempted to use characteristics from the diagnosis of diabetes to forecast the likelihood of developing a cardiac condition. The accuracy produced by our research is clearly depicted in figure 8.

This can be expanded to forecast other types of diabetes-related illnesses, like potential visual impairment. The outcomes of the data analysis can also be used for future study to improve the prediction system's accuracy.

References

1. Rubini, P. E., Subasini, C. A., Katharine, A. V., Kumaresan, V., Kumar, S. G., & Nithya, T. M. (2021). A cardiovascular disease prediction using machine learning algorithms. Annals of the Romanian Society for Cell Biology, 904-912.
2. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.
3. LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.
4. Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207-235.

5. Balakumar, P., Maung-U, K., & Jagadeesh, G. (2016). Prevalence and prevention of cardiovascular disease and diabetes mellitus. Pharmacological research, 113, 600-609.
6. Khrennikov, A. (2009). Interpretations of probability. In Interpretations of Probability. de Gruyter.
7. De Almeida-Pititto, B., Dualib, P. M., Zajdenverg, L., Dantas, J. R., De Souza, F. D., Rodacki, M., ... & Brazilian Diabetes Society Study Group (SBD). (2020). Severity and mortality of COVID 19 in patients with diabetes, hypertension and cardiovascular disease: a meta-analysis. Diabetology & metabolic syndrome, 12, 1-12.
8. Dhande, B., Bamble, K., Chavan, S., & Maktum, T. (2022). Diabetes & Heart Disease Prediction Using Machine Learning. In ITM Web of Conferences (Vol. 44, p. 03057). EDP Sciences.
9. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.
10. Dhande, B., Bamble, K., Chavan, S., & Maktum, T. (2022). Diabetes & Heart Disease Prediction Using Machine Learning. In ITM Web of Conferences (Vol. 44, p. 03057). EDP Sciences.
11. Barron, E., Bakhai, C., Kar, P., Weaver, A., Bradley, D., Ismail, H., ... & Valabhji, J. (2020). Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. The lancet Diabetes & endocrinology, 8(10), 813-822.
12. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.
13. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.
14. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
15. Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.
16. Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart Disease Prediction Using Machine Learning. Heart Disease, 5(1).
17. Dhande, B., Bamble, K., Chavan, S., & Maktum, T. (2022). Diabetes & Heart Disease Prediction Using Machine Learning. In ITM Web of Conferences (Vol. 44, p. 03057). EDP Sciences.