Prediction General Cardiovascular Risk rating Profile

Rimpi Datta[1],A Kiran Kumar[2], Protyay Banerjee[3], Devarati Saha[4], Sruti Bihani[5], Gourab Pal Chowdhury[6]

ABSTRACT

Cardiovascular Diseases (CVDs) are the principal motive for a huge number of loss of life in the world over the previous couple of decades, about one character dies according to the minute because of heart sickness. So, there's a need for dependable and correct gadgets to diagnose such diseases in time for the correct remedy. Data technology was implemented to automate the evaluation of big and complex records inside the field of health care. The proposed work predicts the possibilities of Heart Disease and classifies the threat stage by way of imposing distinctive statistics mining strategies which include Naive Bayes, Decision Tree, and Logistic Regression. Data preprocessing and function selection steps have been done before constructing the models. The fashions were evaluated based totally on the accuracy, precision, bear in mind, and F1 score. The correct prediction of coronary heart ailment can prevent existing threats. The most important goal of this paper is to build an ML model for coronary heart disease prediction primarily based on the related parameters.

Keywords: Cardiovascular Diseases (CVDs)**;** Data science**;** Decision Tree and Logistic Regression; heart disease prediction

1. INTRODUCTION

The excessive changes within the life via the final decades have moved the human societies from farming meals and energetic lives to speedy meals and inactive lifestyles internationally. The excessive consumption of cigarettes is enhancing the danger of cardiovascular illnesses (CVDs) [2]. CVD can be a sort of sickness that consists of heart and blood vessels. According to the World Health Organization, every 12 months 12 million deaths arise worldwide because of Heart Disease. One of the main causes of illness and mortality in a number of the international's populace is heart ailment. It is likewise known as a silent killer because it causes demise without causing evident signs and symptoms.

This look at trying to are expecting future coronary heart illness [5] via comparing affected person records and using a system-mastering set of rules to categorize whether they have coronary heart ailment or no longer. In this situation, machine learning strategies can be extraordinarily useful. Even though coronary heart sickness can take many exclusive bureaucracies [1], there are a few key threat factors that could determine whether or now not someone is a danger of heart ailment. We may say that this approach may be thoroughly adapted to accomplish the prediction of heart sickness by way of accumulating statistics from numerous assets, classifying them beneath suitable headings, after which studying to extract the needed information

2. THEORETICAL BACKGROUND

The Artificial Heart Disease Prediction System was created in 2008 by Sellappan Palaniappan and Rafiah Awang using numerous data mining techniques. They created the Prediction System in 2008 using Decision Trees, Nave Bayes, and Neural Networks (a variety of data mining methodologies).[3] The system was designed on the.NET framework and was entirely Web-based, user-friendly, and scalable.

In 2011, they built a heart disease forecast system in which they first identified 13 key clinical features of patients and then created an artificial neural network algorithm for diagnosing heart disease based on those clinical factors. Input clinical data part, ROC curve display section, and prediction performance display section (execution time, accuracy, sensitivity, specificity, and forecast outcome) were all included in the HDPS [6] system.

3. RESEARCH METHODOLOGY

The methodology section outlines the plans and methods for conducting the survey. This includes survey universes, survey samples, data and data sources, survey variables, and analysis frameworks. The details are as follows,

3.1 Data Collection

The dataset for the project is the records of the patients from a certain hospital.
The dataset contains columns which are:

- **sex**: male(0) or female(1);(Nominal)
- **age**: age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- **currentSmoker**: whether or not the patient is a current smoker (Nominal)
- **cigsPerDay**: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- **BPMeds**: whether or not the patient was on blood pressure medication (Nominal)
- **prevalentStroke**: whether or not the patient had previously had a stroke (Nominal)
- **prevalentHyp**: whether or not the patient was hypertensive (Nominal)
- **diabetes**: whether or not the patient had diabetes (Nominal)
- **totChol**: total cholesterol level (Continuous)
- **sysBP**: systolic blood pressure (Continuous)
- **diaBP**: diastolic blood pressure (Continuous)
- **BMI**: Body Mass Index (Continuous)
- **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, are considered continuous because of a large number of possible values.)
- **glucose**: glucose level (Continuous)
- **10-year risk of coronary heart disease CHD** (binary: “1” means “Yes”, “0” means “No”) - Target Variable.

The very first step in approaching the machine learning problem is data collection.
After data collection, the data is imported into the Jupyter notebook via the Pandas library of python

3.2 Exploratory Data Analysis

It is the method during which the information is analyzed to induce the summarization of the most characteristics of the information. It helps us to urge more insights into the information and the way one column is correlated with the opposite columns. It also helps us to see if the statistical techniques for records analysis are correct or not.
The tools which are used for data visualization are Seaborn and matplotlib the visualization techniques which we use are-

3.2.1 Univariate Analysis

In univariate analysis, the summary statistics of all the features are performed separately and we can analyze each feature by their distributions whether it is normal, left-skewed or right-skewed

3.2.2 Multivariate Analysis

In the statistical procedure, we are able to determine how each feature is correlated with one another and also how each feature is expounded to the target feature, if two features are highly correlated with one another or with the target variable then anybody of the features is often dropped or deleted because it won’t help much because both of them are equally contributing to the target feature [3].

The libraries like seaborn and matplotlib make us easier to try and do the exploratory data analysis and data visualization

3.2.3 Feature Engineering

It is a very important part of machine learning where we check the number of missing values within the data and also the decisions like removing the missing values or filling the missing values with mean, and also the median is taken during this part. So as to form the model work on unknown data, it's very necessary to coach the information well and it may also produce new features with the goal of speeding up data transformations and increasing the accuracy. Feature engineering is extremely much essential in machine learning as a terrible feature within the data will impact the model. By the method of feature engineering, new artificial features are designed into an algorithm, this is often because the new artificial features can improve the model performance.

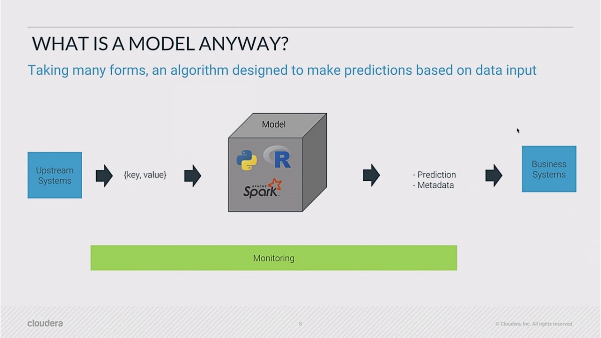
3.2.4 Outlier Detection

An outlier may be a value that lies far-flung from the opposite values in a very random sample of the population. Because of these outliers the mean of the sample gets deviated leading to a skewed distribution and thanks to this the model Performance also gets affected. However, in some cases, outliers also play an important role in the model’s prediction (an example of such a case is credit card fraud detection). The simple regression model [8] is more at risk of outliers. There are some methods to handle the outliers:

Removal- because the name suggests during this method the outliers are removed, but in some cases, if the outliers are removed there'll be plenty of knowledge that can be lost which is able to end in a decrease in the model accuracy.

3.2.5 Modeling

In this step first the data is split into training and testing sets and then the model is trained using the train set and the testing of the model is done using the test set. In this project, several algorithms are used for predictions like Logistic regression, Decision tree classifier, Random forest classifier, Neighbors Classifier, and Gradient Boosting Classifier. Out of which we have chosen the Gradient boosting classifier as it is more accurate than the rest of the other models with an accuracy of about 84.03%



WHAT IS A MODEL ANYWAY?

Taking many forms, an algorithm designed to make predictions based on data input

4. EVALUATION

After training the model using the plaything and testing the model using the test set, the accuracy is calculated. Which is the difference between the predictions and therefore the original values. This accuracy of the model is often increased by evaluating the model using hyperparameter tuning. Hyperparameter tuning refers to selecting the simplest parameters for the model so the accuracy of the model increases but it doesn't over fit hyperparameter [8] tuning may be wiped out two methods:

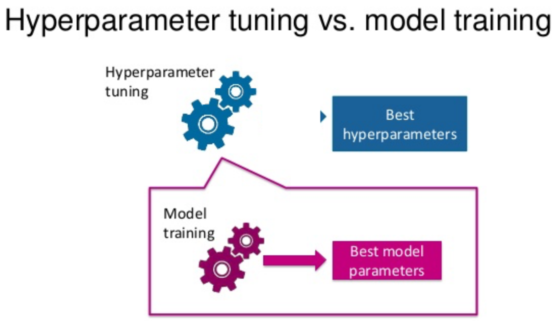
4.1 GridSearchCv

It facilitates looping thru predefined hyper parameters and healthy your estimator (model) for your education set. So, in the end, you can select the best parameters from the listed hyperparameter.

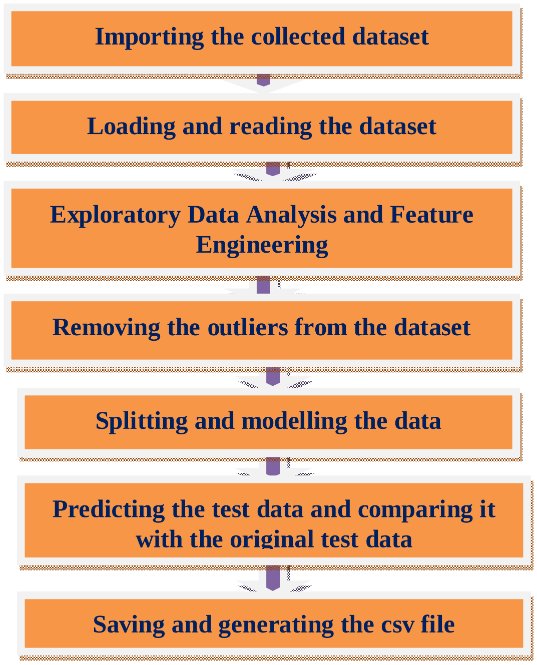
4.2 RandomizedSearchCV

In RandomizedSearchCV, instead of providing a discrete set of values to explore each hyperparameter, we provide a statistical distribution or list of hyperparameters [8]. The values of various hyperparameters are randomly extracted from this distribution.

Fig2: Model



The Flowchart of the above processes is given below


Importing the collected dataset
Loading and reading the dataset
Exploratory Data Analysis and Feature Engineering
Removing the outliers from the dataset
Splitting and modelling the data
Predicting the test data and comparing it with the original test data
Saving and generating the csv file

5. RESULT ANALYSIS

Tools: Anaconda3 (version 1.9.0)
Jupyter Notebook (version 6.4.5)
MS Excel (2016)
MS word (2016)
Platform: Microsoft Windows 10

Hardware:

Minimum

HDD	100 GB
Processor	Pentium 4
Memory	4 GB

Used

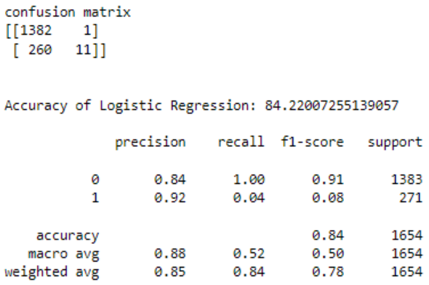
HDD	1 TB
Processor	AMD Ryzen 5 3400G
Memory	8 GB

There are five algorithms are used in this project.

5.1 Logistic Regression

Logistic Regression is very commonly used to solve classification problems. It is used to predict a categorical target variable from a group of independent variables.

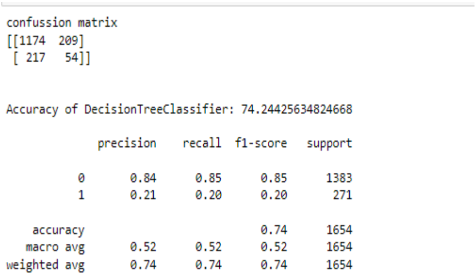
Here in this project by training the model of 'LR', its accuracy is near about 84.22%.



```
confusion matrix
[[1382    1]
 [ 260   11]]


Accuracy of Logistic Regression: 84.22007255139057

              precision    recall  f1-score   support

           0       0.84      1.00      0.91      1383
           1       0.92      0.04      0.08       271

    accuracy                           0.84      1654
   macro avg       0.88      0.52      0.50      1654
weighted avg       0.85      0.84      0.78      1654
```

5.2 Decision Tree Classifier

A tree-structured classifier, the Decision Tree algorithm is used to solve both regression and classification problems. The decisions or assessments are based on the characteristics of the dataset and its accuracy achieved by 74.24%

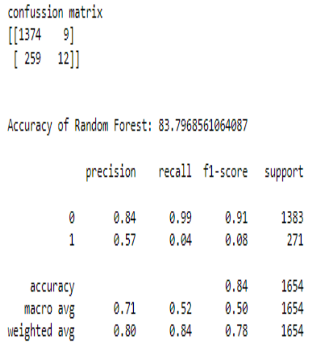


```
confussion matrix
[[1174  209]
 [ 217   54]]


Accuracy of DecisionTreeClassifier: 74.24425634824668

              precision    recall  f1-score   support

           0       0.84      0.85      0.85      1383
           1       0.21      0.20      0.20       271

    accuracy                           0.74      1654
   macro avg       0.52      0.52      0.52      1654
weighted avg       0.74      0.74      0.74      1654
```

5.3 Random Forest Classifier

The most flexible and easy-to-use algorithm is Random Forest, which includes bagging to improve the performance of the Decision Tree.

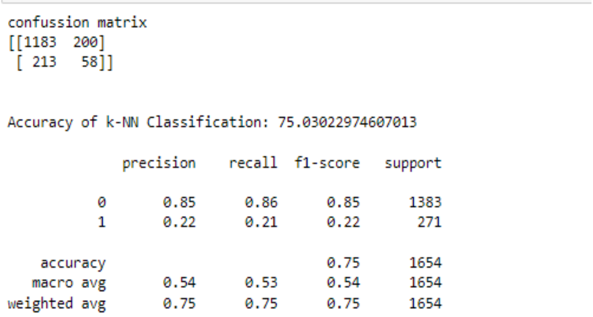
Training the 'RF' model got an accuracy of 83.19%.



```
confussion matrix
[[1374    9]
 [ 259   12]]


Accuracy of Random Forest: 83.7968561064087

              precision    recall  f1-score   support

           0       0.84      0.99      0.91      1383
           1       0.57      0.04      0.08       271

    accuracy                           0.84      1654
   macro avg       0.71      0.52      0.50      1654
weighted avg       0.80      0.84      0.78      1654
```

5.4 K-Nearest Neighbors (Knn) Algorithm

The KNN algorithm, which used for both classification as well as regression predictive problems.ch is used in both classification as well as regression predictive problems, predicts the values of new data points using 'significant positive correlation,' which implies that the fresh data point will be allocated a value depending on how closely it resembles the points in the training dataset.

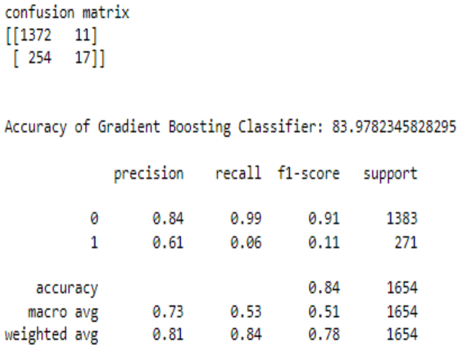


```
confussion matrix
[[1183  200]
 [ 213   58]]


Accuracy of k-NN Classification: 75.03022974607013

              precision    recall  f1-score   support

           0       0.85      0.86      0.85      1383
           1       0.22      0.21      0.22       271

    accuracy                           0.75      1654
   macro avg       0.54      0.53      0.54      1654
weighted avg       0.75      0.75      0.75      1654
```

5.5 Gradient Boosting Classifier

Gradient Boosting Trees can be used to solve regression and classification problems.It returns a prediction model in terms of an ensemble of poor prediction models, most commonly decision trees.

Here ‘GVC’ model is 83.97% accurate.



```
confusion matrix
[[1372   11]
 [ 254   17]]


Accuracy of Gradient Boosting Classifier: 83.9782345828295

              precision    recall  f1-score   support

           0       0.84      0.99      0.91      1383
           1       0.61      0.06      0.11       271

    accuracy                           0.84      1654
   macro avg       0.73      0.53      0.51      1654
weighted avg       0.81      0.84      0.78      1654
```

6. Accuracy

Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset.

Accuracy = (TP + TN) /(TP+FP+FN+TN)

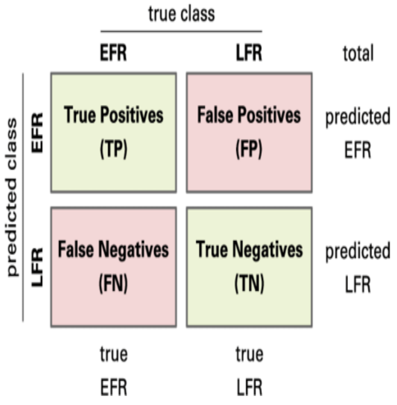
Where

True positives (TP): are cases in which the prediction is yes as they have the disease and they actually do.

True negatives (TN):as per prediction, they don't have the ailment.

False positives (FP): We projected that they would have the disease, but they don't.
False negatives (FN): We projected that they would not have the disease, yet they do.

Confusion Matrix- A confusion matrix is a table that shows how well a classification algorithm (or "classifier") performs on a set of test data for which the true values are known.



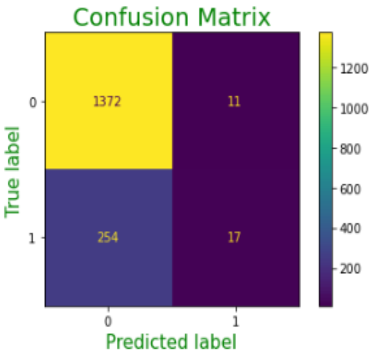
$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

$$\mathrm{CA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

$$\mathrm{F}_1 = \frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}}$$

Here confusion matrix is shown belowwith the total performance of the system.



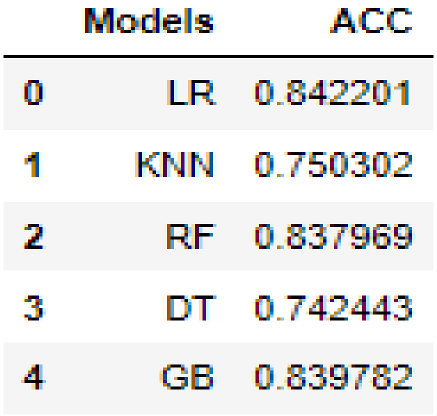
The ratio of correctly diagnosed patients with the disease (TP) to the total patients projected to have the disease (TP+FP) is the correctness of an algorithm.
Precision=TP/(TP+FP)=(1372)/(1372+11)=0.99

The recall matrix is the ratio of TP divided by the total number of patients who actually have the disease.
Recall=TP/(TP+FN)=(1372)/(1372+254)=0.84

F1 score represents the difference between recall and precision
F1Score=(2*precision*recall)/(precision+recall)=0.91

6.1 Accuracy Analysis

In the end, we conclude that the G-Boost is more accurate than other methods after using a machine learning approach for training and testing. The number count of TP, TN, FP, and FN is supplied [7], and using the equation of accuracy, a value has been determined. It is concluded that extreme gradient boosting is the best with 83.97% accuracy, and the comparison is presented below.

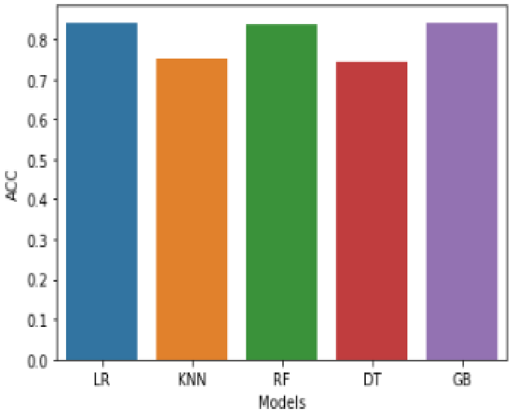


	Models	ACC
0	LR	0.842201
1	KNN	0.750302
2	RF	0.837969
3	DT	0.742443
4	GB	0.839782

6.2 Bar Graph

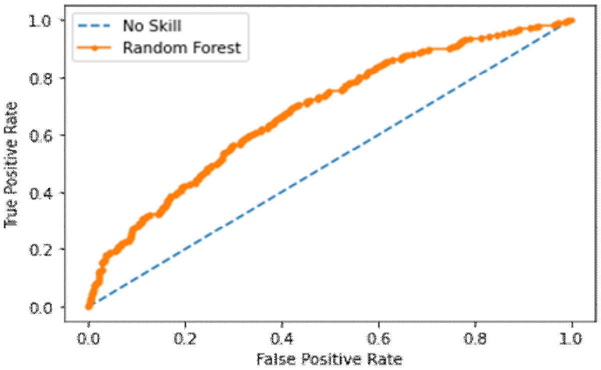
Here’s the summarized bar plot of the Models along with their accuracy.

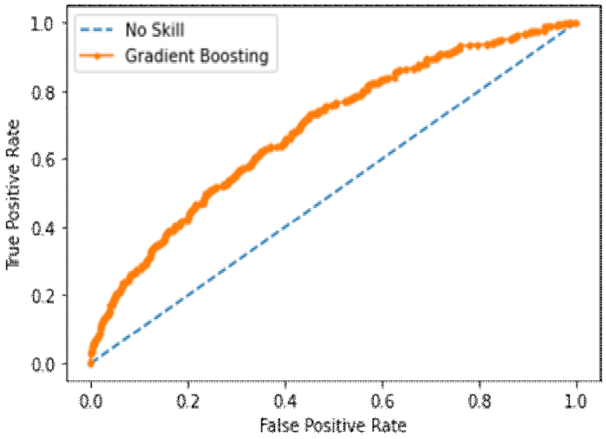
By this, we are taking the Gradient Boosting Classifier for our output prediction due to its high accuracy with less time complexity.

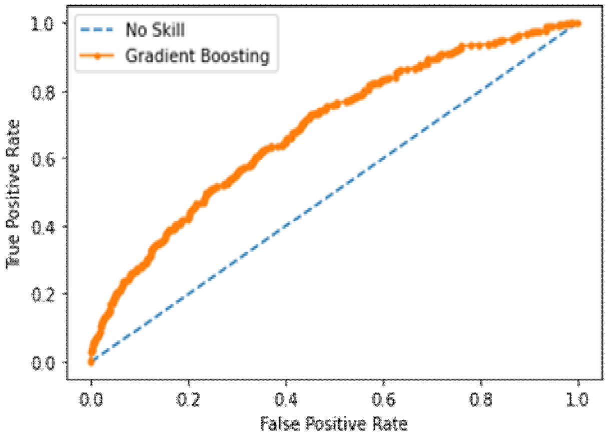


6.3 Roc Curve

The ROC curve depicts the trade-off between sensitivity and specificity in terms of accuracy. Classifiers with curves that are closer to the highest corner perform better [7]. A random classifier is supposed to give points that are diagonal (FPR = TPR) as a baseline.





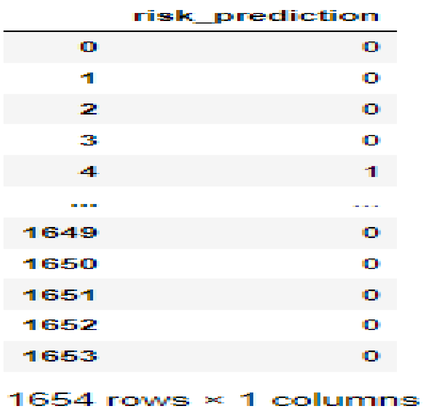


7. Prediction

As we know the GB classifier model has the highest accuracy score, so we used GB for predictions.

Here, for prediction, 0 →there is no risk of heart disease and

For prediction 1 →there is a risk of heart disease.



```
      risk_prediction
   0                0
   1                0
   2                0
   3                0
   4                1
 ...              ...
1649                0
1650                0
1651                0
1652                0
1653                0
1654 rows × 1 columns
```

After that, by using file handling we generate one CSV file and store all predicted data in that file named ‘risk_prediction.csv’.

8. CONCLUSION

Heart Disease is one of the major concerns for society today. With the increasing number of deaths due to heart diseases, it has become necessary to develop a system to predict heart diseases effectively and accurately. In this paper, we proposed different methods in which comparative analysis was done and results were achieved. It predicts people with cardiovascular disease by extracting the patient medical history that leads to fatal heart disease from a dataset that includes path patient's medical history such as diabetes sugar level, blood pressure, etc. The algorithms used in building the model for predicting heart disease are Logistic Regression, gradient boosting classifier, Decision tree classifier, Random Forest classifier, and K nearest neighbor (KNN). Various promising results are achieved and are validated using the accuracy and confusion matrix. The result of this study indicates that the Gradient Boosting algorithm is the most efficient algorithm with an accuracy score of 83.97% for the prediction of heart disease.

References

[1] Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: a survey. Artif Intell Rev. 2017;47(3):313

[2] International Journal of Computer Applications (0975 – 8887) Volume 176 – No. 11, April 2020 17 Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms Daniel Ananey-Obiri Department of Computational Science and Engineering North Carolina Agricultural and Technical State University Enoch Sarku

[3] IJCSMC, Vol. 8, Issue. 5, May 2019, pg.119 – 125 PREDICTING THE PRESENCE OF HEART DISEASE USING MACHINE LEARNING Akshay Jayraj Suvarna1; Arvind Kumar; Muthamma K

[4] Akbaş KE, Kivrak M, Arslan AK, Çolak C. Assessment of association rules based on certainty factor: an application on heart data set, in 2019 International artificial intelligence and data processing symposium (IDAP) (pp. 1–5). IEEE; 2019.

[5] A novel approach for heart disease prediction using strength scores with significant predictors by Armin Yazdani, Kasturi Dewi Varathan, Yin Kia Chiam, Asad Waqar Malik & Wan Azman Wan Ahmad on BMC Medical Informatics and Decision Making volume 21.

[6] Amin MS, Chiam YK, Varathan KD Identification of significant features and data mining techniques in predicting heart disease.

[7] Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms by Kaushalya Dissanayake.

[8] Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators by Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam, Hui Na Chua, Pranavanand