A Machine Learning-based Diagnostic Framework for Heart Disease Prediction

Shahanur Rahman
Student
Department of Information and Communication Engineering, Pabna University of Science and Technology

Md. Ebrahim Hossain
Assistant Professor
Department of Computer Science and Engineering, Leading University, Sylhet

Taskin Noor Turna
Assistant Professor
Department of Information and Communication Engineering, Pabna University of Science and Technology

ABSTRACT
Heart disease continues to be one of the primary causes of death around the world, and early identification is crucial for lowering death rates and patient outcomes. While traditional diagnostic techniques are effective, but they often take large amount of time, financial investment, and expert analysis. This thesis investigates the capabilities of machine learning algorithms in predicting heart disease, providing a means for diagnostic support that is efficient, precise, and widely accessible. We have utilized several machine learning models, including Logistic Regression, Support Vector Machines, Naïve Bayes, Random Forest, Gradient Boosting, ANN, XGBoost and KNN, on a dataset consisting of patient health records featuring essential factors like age, cholesterol levels, blood pressure, and various lifestyle elements. The research encompassed data preprocessing, feature selection, and model optimization to improve prediction accuracy. To identify the key features of heart disease, seven performance metrics (Accuracy, Classification error, Prediction Time, Precision, Sensitivity, F-Measure and Specificity) are employed, which provide better insight into the behavior of various feature-selection combinations. By analyzing the seven matrices values of the eight models, we have chosen three models (Logistic Regression, Gradient Boosting and Random Forest) from them and we propose a novel method (LGR Model) by combining these three models for getting higher accuracy. The accuracy of the proposed model is 88%.

General Terms
Heart Disease Prediction

Keywords
Logistic Regression, Gradient Boosting, Random Forest, Prediction, Heart Disease

1. INTRODUCTION
For many years, Cardio-vascular disease, also referred to as heart-based illness, has been the world's leading cause of mortality and encompasses a variety of heart-related disorders. Its development is linked to numerous risk factors, highlighting the urgent need for accurate, reliable, and effective methods for early diagnosis to ensure timely management of the condition [1]. Heart failure, arrhythmias, coronary artery disease, valvular heart disease, and other heart-related conditions that affect the structure and function of the heart are all included in the wide category of cardiovascular disease. According to WHO estimates, heart-based disease kills 17.9 million people annually, making it the world's top cause of death [2]. The rising prevalence of risk factors such high blood pressure, diabetes, obesity, elevated cholesterol, smoking, and sedentary lifestyles has further worsened this concerning trend.

The complexity of heart disease lies in its multifactorial nature. Environmental variables and genetic predispositions can work together to influence it. Early detection and risk assessment are crucial, as timely intervention can significantly reduce morbidity and mortality. However, traditional diagnostic methods often rely on subjective assessments, which can lead to delayed diagnoses and treatment.

Recent advancements in medical technology and data analytics have opened up new possibilities for enhancing heart disease prediction. Machine based learning algorithms, which can analyze massive amounts of data and spot complex patterns, have a lot of potential in this field. These algorithms can more precisely determine a person's risk of heart disease by looking at variables including age, gender, medical history, and lifestyle.

In paper [3], Senthilkumar Mohan et al. (2019) developed a hybrid model (HRFLM) combining Random Forest and a Linear Model, achieving 88.7% accuracy in heart disease prediction by identifying key variables through machine learning techniques. In [4], Jaymin Patel et al. (2015) used data mining and machine learning techniques in WEKA to evaluate Decision-Tree algorithms like Random Forest, J48, and Logistic Model Tree for heart disease prediction using the Cleveland UCI dataset. Paper [5], conducted a comprehensive study using nine machine learning classifiers with preprocessing, hyperparameter tuning, and K-fold cross-validation to predict heart disease, evaluating performance based on sensitivity, specificity, F-measure, and accuracy. In [6], authors developed a machine learning-based framework using algorithms like Random Forest, KNN, Decision Trees, and Logistic Regression to predict various cardiac disorders, training the model on UCI dataset and validating it with new data.

2. SYSTEM MODEL
These days, machine-based learning techniques are widely used to forecast cardiac illness. Here, the effectiveness of several prediction methods is evaluated using a range of machine-learning algorithms. Each method provides individual prediction results, and their performances are compared to determine which one offers the best accuracy with the least error. This study predicts cardiac-disease using machine-learning methods. The process, which composed of many stages including data-preparation, data training and testing, and prediction techniques, is succinctly summarized by the system model. The system model illustrates a step-by-step process for building a heart disease prediction model. The process begins with starting the system and collecting a heart disease dataset

that contains relevant health indicators and patient data. Important features or variables related to heart disease are then extracted from this dataset to ensure that only the most significant factors are used in the model. After selecting the features, the data undergoes preprocessing, which includes tasks such as resolving any missing values, cleaning, and normalizing them in order to get them ready for model training. Following pre-processing, the data is partitioned into two subsets: testing-data and training-data. Machine-learning models, especially classifiers like K-Nearest Neighbors (KNN), Logistic-Regression, Random-Forest Classifier, SVM, Naïve-Bayes, Gradient-Boosting, and XGBoost, are trained using the training-data. Following training, the classifier usage to generate predictions on the testing-data. The classifier produces a prediction that assigns a "High Risk of CHD" or "Low Risk of CHD" classification to the data. Based on the input data, this method is intended to allow for the precise prediction of cardiac illness. The phases of the system model are illustrated below:

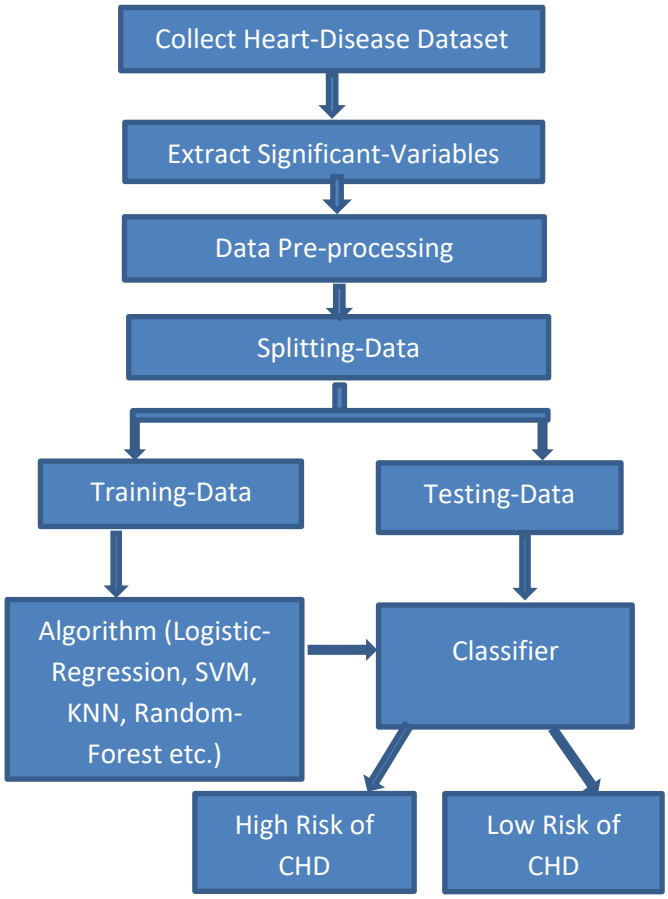

Fig 1: System Model

2.1 Dataset and Attributes

The dataset comes from an ongoing cardiovascular study of residents in Framingham, Massachusetts and I have collected this dataset from the Kaggle website [7]. The classification job's objective is to forecast a patient's 10-year risk of coronary heart disease (CHD). This data-set contains over 4,000 records and includes 15 attributes with information about the patients.

Table 1: Dataset Description

Serial No.	Attributes Name	Attributes Descriptions
1	Sex	Binary variable indicating the sex of the patient, where 1 represents male and 0 represents female.
2	Age	The age of the patient in years.
3	CurrentSmoker	Binary variable indicating whether the patient currently smokes (0 for non-smokers and 1 for smokers).
4	CigsPerDay	If the patient smokes, how many cigarettes do they smoke each day?
5	BPMeds	Whether the patient is currently taking blood pressure medicine is suggested by a binary variable (1 for yes, 0 for no).
6	prevalentStroke	The patient's history of stroke is suggested by a binary variable (1 for yes, 0 for no).
7	prevalentHyp	A binary variable with 1 denoting a history of hypertension and 0 denoting none.
8	diabetes	Whether the patient has diabetes is suggested by a binary variable (1 for yes, 0 for no).
9	totChol	The total cholesterol level of the patient (measured in mg/dL).
10	sysBP	The patient's systolic blood-pressure (measured-in mmHg).
11	diaBP	The patient's diastolic blood-pressure (measured-in mmHg).
12	BMI	Which calculates body fat based on height and weight.
13	heartRate	The patient's heart-rate (measured in beats-per-minute).
14	glucose	The patient's glucose-level (measured in mg/dL), typically used to diagnose diabetes.
15	TenYearCHD	Risk of CHD within the next ten years is suggested by a binary variable (1 being high risk and 0 being low risk). This is the target variable for prediction.

2.1.1 Output Data Distribution

According to the dataset, it has contained exactly 4240 patient information. After handling and removing the NULL values, the dataset contains 3,751 patient records, which are used to predict whether an individual is at risk for TenYearCHD. The output class attribute has two values: 0 and 1, representing whether a person is affected by TenYearCHD or not. The results show that 572 individuals are affected by TenYearCHD, while 3,179 are not. Below given a diagrammatic representation of the TenYearCHD informations:

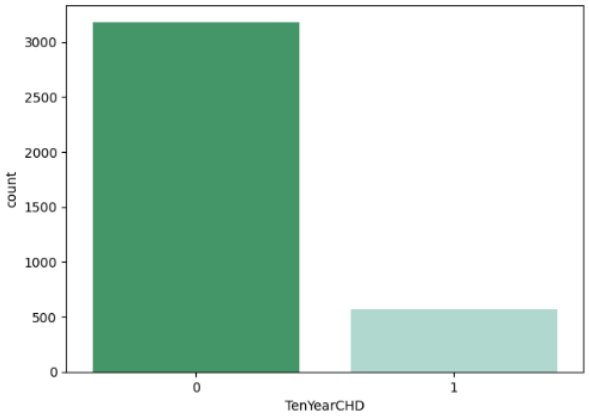

Fig 2: Calculating the number of CHD patients (0 = Not Affected; 1 = Affected)

2.2 Extract Significant Variables

"Extract Significant Variables" refers to the process of identifying and selecting the most important or relevant-features (variables) from a data-set that have the greatest influence on the outcome or target-variable. When used to a prediction model, this indicates selecting the attributes that contribute the most to predicting the target value (e.g., whether a person has a risk of CHD in the data-set). This step helps to:

- Simplify the model by eliminating aspects that are superfluous or unnecessary.
- Enhance the model's functionality by concentrating on the most important elements.
- Reduce the number of variables taken into consideration to improve the model's interpretability.

Common techniques for extracting significant variables include statistical tests, correlation analysis, or machine-learning algorithms like decision-trees and feature importance methods.

2.3 OUTPUT

The output of the classifier represents the prediction result. It typically falls into two categories:

- High Risk of CHD: Individuals identified as having a significant likelihood of developing heart- disease based on the predictive model.
- Low Risk of CHD: Individuals determined to have a lower likelihood of developing heart-disease in line with the study of the model.

The output allows for a straightforward interpretation of the results, supporting the decision-making process for additional medical assessment or treatment.

3. RESULT AND DISCUSSION

3.1 Input Data Analysis

Through the Kaggle platform, we were able to access the data-set, which comes from ongoing cardiovascular research involving residents of Framingham, Massachusetts. Predicting a patient's 10-year risk of CHD is the aim of the categorization job. It includes data on over 4,000 patients, with 15 attributes such as sex, age, diabetes, BMI, systolic and diastolic blood pressure (sysBP and diaBP), heart rate, and more. The following figures show the distribution of input data.

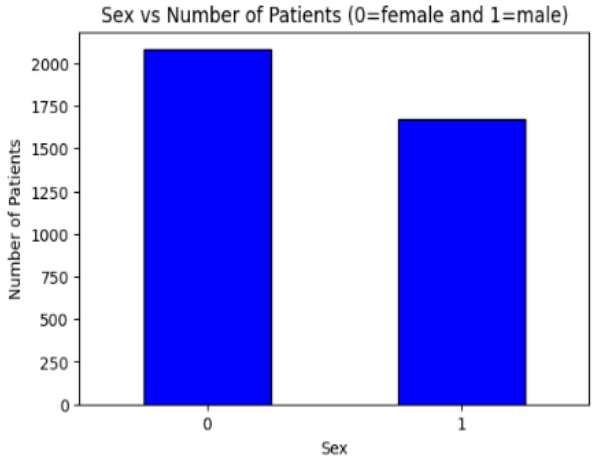

Fig 3: Sex vs Number of Patients

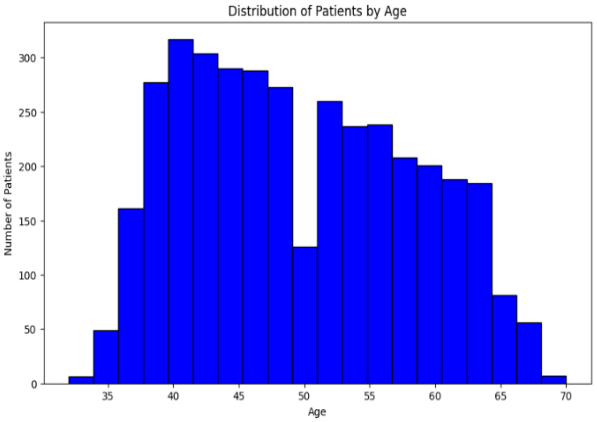

Fig 4: Age vs Number of Patients

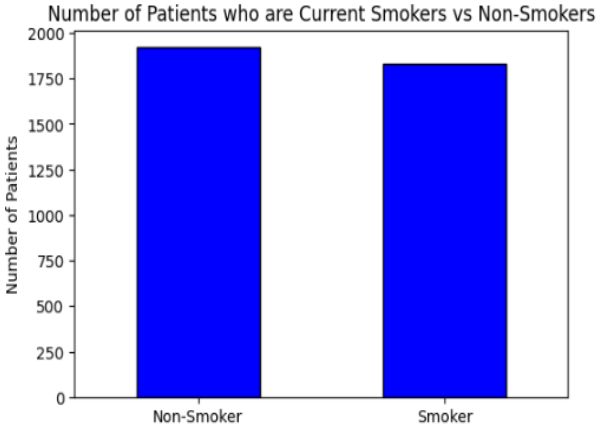

Fig 5: Smoker vs Non-Smoker

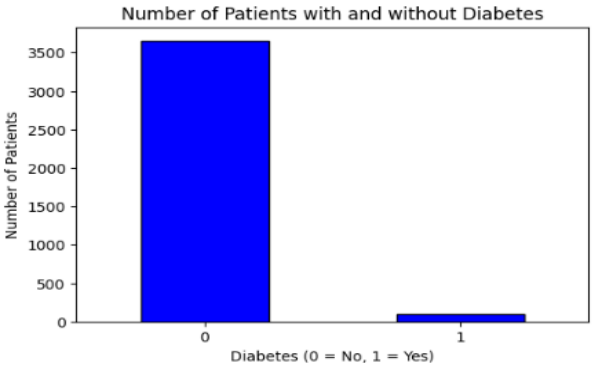

Fig 6: Diabetes status vs Number of Patients

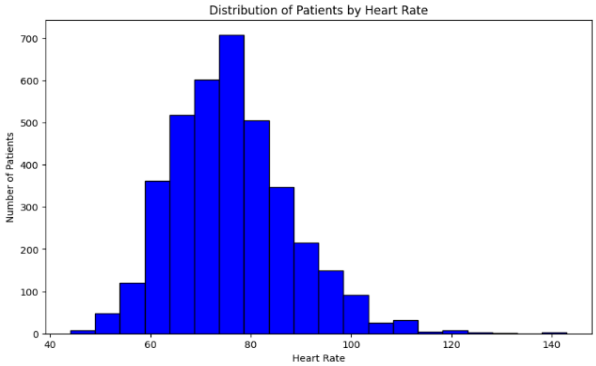

Fig 7: Distribution of patients by Heart rates

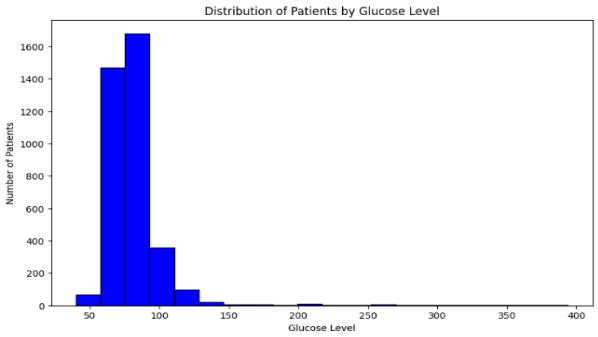

Fig 8: Distribution of patients by Glucose Level

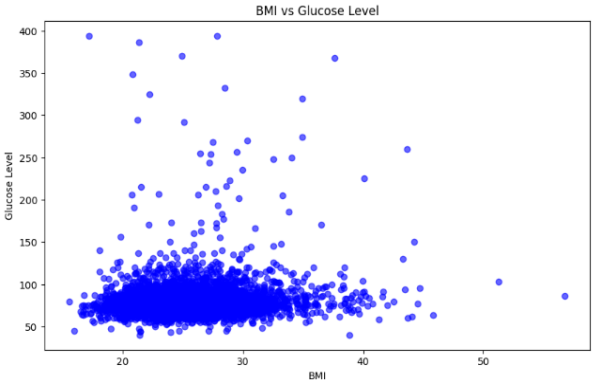

Fig 9: BMI vs Glucose Level

3.2 Accuracy Analysis

Accuracy-analysis is the process of assessing a machine-learning model's performance by comparing its predictions with the actual outcomes. In classification tasks, accuracy is calculated by dividing the total number of predictions by the number of accurate predictions made by the model. It is a frequently used metric to evaluate model performance. The following is the accuracy formula:

$$\text{Accuracy} = \frac{\text{The quantity of accurate forecasts}}{\text{Total forecasts}}$$

Below given the accuracy table for the various types of machine learning techniques which we have utilized in our project.

Table 2: Comparison of accuracy for various models

SL. No.	Name of the Model	Accuracy
1	Logistic-Regression	0.85
2	Naïve-Bayes	0.81
3	SVM	0.84
4	KNN	0.83
5	Random-Forest	0.84
6	Gradient-Boosting	0.84
7	ANN	0.83
8	XGBoost	0.83

Below given the accuracy curve for these models so that we can clearly understand the accuracy differences between these models:

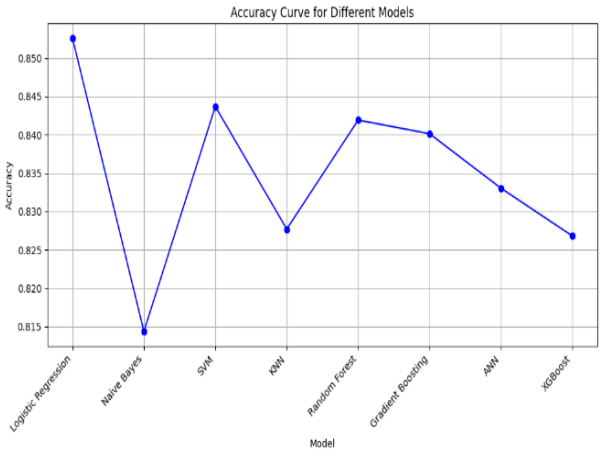

Fig 10: Accuracy Curve for various models

By seeing the accuracy table and curve, we can conclude that, out of the eight models, the model with the best accuracy is the logistic regression model, whereas the accuracy of the Naïve-Bayes model has been the lowest.

3.3 Classification Error

By evaluating the percentage of inaccurate predictions, classification error is a statistic used to evaluate a classification model's performance. In essence, it shows the proportion of cases in which the actual class and the model's projected class are different.

The formula for classification error is:

$$\text{Classification Error} = \frac{\text{The quantity of incorrect forecasts}}{\text{Total forecasts}}$$

Another way to express this is:

$$\text{Classification Error} = 1 - \text{Accuracy}$$

Table 3: Comparison of Classification error for various models

Name of the Model	Classification Error
Logistic-Regression	0.15
Naïve-Bayes	0.19
SVM	0.16
KNN	0.17
Random-Forest	0.16
Gradient-Boosting	0.16
ANN	0.17
XGBoost	0.17

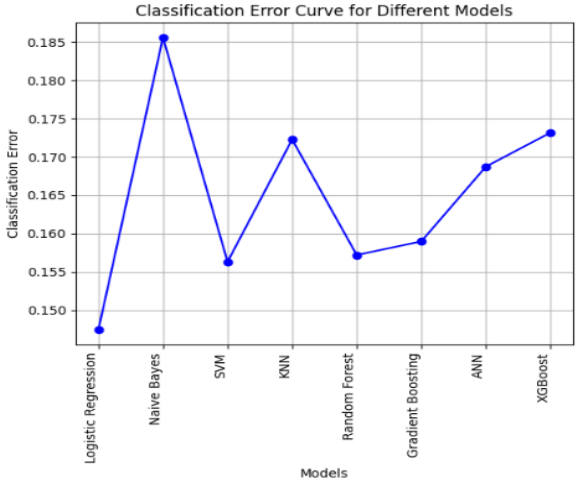

Fig 11: Classification Error Curve for various models

By seeing the classification error table and curve, we can conclude that, out of the eight models, the logistic regression model has yielded the lowest classification error, while the Naïve Bayes model has resulted in the highest classification error.

3.4 Various measurement for Eight Models

Table 4: Various measurement for Eight Models

Model	Accuracy	Error	Prediction Time	Precision	Sensitivity	F-Measure	Specificity
Logistic-Regression	0.85	0.15	0.0050	0.83	0.85	0.84	0.99
Naïve-Bayes	0.81	0.19	0.0009	0.78	0.81	0.80	0.93
SVM	0.84	0.16	0.1499	0.78	0.84	0.81	0.99
KNN	0.83	0.17	0.1237	0.77	0.83	0.80	0.96
Random-Forest	0.84	0.16	0.0200	0.79	0.84	0.82	0.99
Gradient-Boosting	0.84	0.16	0.0028	0.79	0.84	0.81	0.98
ANN	0.83	0.17	0.0013	0.79	0.84	0.81	0.97
XGBoost	0.83	0.17	0.0038	0.78	0.83	0.80	0.95

For the job of predicting cardiac illness, the table contrasts the performance of eight machine-learning models using a number of variables, including accuracy, error rate, prediction time and precision.

By analyzing the several matrices values of the eight models we have chosen three models namely- Logistic-Regression, Gradient-Boosting and Random-Forest. By combining the performances of these three models we have created a proposed model (LGR Model). When compared to all eight models, the accuracy of this suggested model is superior. The accuracy of the proposed model is 0.88.

3.5 Proposed Model: Stacked Ensemble

To utilize the strengths of Logistic-Regression, Random-Forest, and Gradient-Boosting models, we can combine these models into an ensemble learning approach. A common method to achieve this is stacking (stacked generalization). Here's the proposed model and explanation:

The process begins by collecting the activity data, which includes the necessary features and target labels for predicting coronary heart disease (CHD) risk. Once the data is acquired, it undergoes preprocessing, where features are scaled or normalized, missing data are handled, categorical variables are encoded, and unnecessary columns are eliminated. This stage guarantees that the data is clean and training-ready.

Following preprocessing, the cleaned data set is used to train the base models, including Logistic-Regression, Random-Forest, and Gradient-Boosting. Every basic model is trained separately on the training dataset and configured to output probabilities rather than direct class predictions, which is essential for ensemble methods.

The next step involves generating meta-features. The following layer, called the meta-model, uses the outputs (predicted

probabilities) of the underlying models as input characteristics. These meta-features are typically generated using cross-validation or a separate validation set to prevent data leakage and ensure robust training.

The meta-model, often a simple Logistic Regression or another classifier, is then trained using the meta-features. Its function is to create the final forecast by combining the results of the base models. The meta-model offers a more accurate overall forecast by taking into account the advantages and disadvantages of each base model. Lastly, predictions are made on the test dataset or fresh incoming data using the learned meta-model. Based on the predicted probabilities or specified thresholds, the outcomes are classified as either "High Risk of CHD" or "Low Risk of CHD."

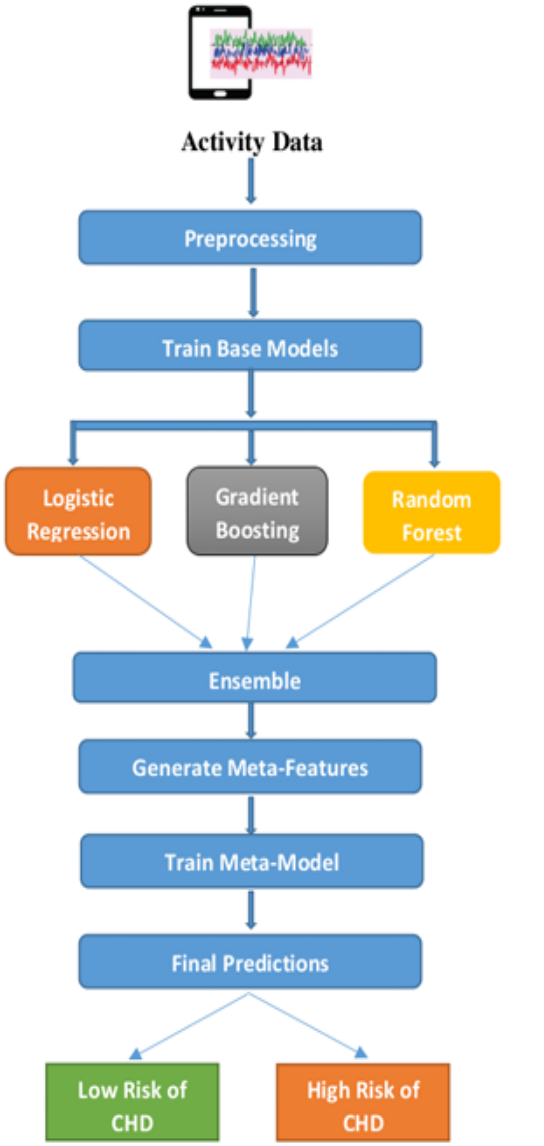

Fig 12: LGR Model (Proposed)

3.6 Input from the user

We have taken input from the user which was same as the 3rd row of the dataset and the dataset output is 0 (i.e. Low risk of CHD) in this particular row. Our model provide the same output as the output of the dataset. Below given the inputs which we were taken from the user:

Input:

Enter values for the following features:

Sex (1 for male, 0 for female): 0

Age: 46

Current Smoker (1 for Yes, 0 for No): 0

Cigarettes per Day: 0

On BP Meds (1 for Yes, 0 for No): 0

Prevalent Stroke (1 for Yes, 0 for No): 0

Prevalent Hypertension (1 for Yes, 0 for No): 0

Diabetes (1 for Yes, 0 for No): 0

Total Cholesterol: 250

Systolic Blood Pressure: 121

Diastolic Blood Pressure: 81

Body Mass Index (BMI): 28.73

Heart Rate: 95

Glucose Level: 76

Output:

Predicted outcome: Low risk of CHD

Besides, we have taken another input from the user which was same as the 8th row of the dataset and the dataset output is 1 (i.e. High risk of CHD) in this particular row. Our model also provide the same output as the output of the dataset. Below given the inputs which we were taken from the user:

Input:

Enter values for the following features:

Sex (1 for male, 0 for female): 0

Age: 63

Current Smoker (1 for Yes, 0 for No): 0

Cigarettes per Day: 0

On BP Meds (1 for Yes, 0 for No): 0

Prevalent Stroke (1 for Yes, 0 for No): 0

Prevalent Hypertension (1 for Yes, 0 for No): 0

Diabetes (1 for Yes, 0 for No): 0

Total Cholesterol: 205

Systolic Blood Pressure: 138

Diastolic Blood Pressure: 71

Body Mass Index (BMI): 33.11

Heart Rate: 60

Glucose Level: 85

Output:

Predicted outcome: High risk of CHD

4. CONCLUSION

For predicting heart disease, we have used eight machine-learning model namely: Logistic-Regression, Support-Vector Machine (SVM), K-Nearest Neighbor (KNN), Random-Forest,

Gradient-Boosting, Artificial Neural Network (ANN), XGBoost and Naïve-Bayes. To identify the key features of heart disease, seven performance metrics (Accuracy, Classification-error, Prediction-Time, Precision) are employed, which provide better insight into the behavior of various feature-selection combinations. By analyzing the four matrices values of the eight models, we have chosen three models (Logistic-Regression, Gradient-Boosting and Random-Forest) from them and we propose a novel method (LGR Model) by combining these three models for getting higher accuracy. The accuracy of the proposed model is 0.88. Additionally, we used user input, and the model's output produced the same outcome as the dataset.

5. REFERENCES

[1] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345.

[2] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[3] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[4] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.

[5] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, 2022(1), 1410169.

[6] Yadav, A. L., Soni, K., & Khare, S. (2023, July). Heart diseases prediction using machine learning. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[7] Coronary Heart Disease Prediction in Ten Years. (2023, December 10). Kaggle. https://www.kaggle.com/datasets/palakdoshijain/coronary-heart-disease-prediction-in-ten-years?resource=download