Acta Scientific COMPUTER SCIENCES

Volume 6 Issue 3 March 2024

Research Article

Integration of Machine Learning Techniques for Heart Disease Prediction

Adebisi Abraham Owodunni*, Tareq Al-Jaber and Zhibao Mian

Computer Science (Artificial Intelligence and Data Science), University of Hull, Hull, England, United Kingdom

***Corresponding Author:** Adebisi Abraham Owodunni, Computer Science (Artificial Intelligence and Data Science), University of Hull, Hull, England, United Kingdom.

Received: February 02, 2024

Published: February 22, 2024

Abstract

As important as the heart is to humans, unfortunately, 43% of death is from heart disease [2] declared by Global Burden of Disease research. By 2030, deaths from cardiovascular disease will reach 23.6 million where heart disease takes the lead [3]. Annually, 10 million people die globally according to World Health Organization (WHO). There have been (pre)established conventional ways of detecting this disease in humans like angiography, electrocardiograms among others, which are not only expensive for the common man, but have been proven, but over 17 million individuals have lost their lives to lack of expertise, incapacitation with several side effects [4]. According to a WHO survey, only 67% of the time, doctors can accurately predict heart disease. Hence the need for non-invasive and a more efficient technique thereby leveraging on Data Science (Machine Learning - ML). This research makes use of ML techniques to classifying Heart Disease through the comparative way of their metrics to predict heart disease in individuals, ii. Investigate the most relevant features and the risk factors contributing to predicting heart disease, iii. Evaluate the performance of the developed models using appropriate metrics, iv. Provide insights and recommendations for healthcare professionals to improve early diagnosis and intervention strategies. These involve four classifiers: XGBoost, Random Forest (RF), Logistic Regression (LR), and Support Vector Machine, to classify and predict heart disease using the Framingham heart disease dataset. Different models were built after handling missing values and outliers in the dataset. Before balancing the dataset, the models built, LR and RF gave the best performance with an accuracy of 85% each. The dataset was later balanced/resampled, and important features selection was done using the XGBoost classifier, Sequential Feature Selection (SFS) and KBest methods respectively, and these improved the performance of the model. Ensemble techniques (AdaBoost and Bagging) were adopted and the AdaBoost model (RF classifier) performed as high as giving an accuracy of 93%. Hyperparameter tuning was done involving Randomized SearchCV and Grid SearchCV, but none outperformed the AdaBoost model's performance. Lastly, the balanced dataset was split into train and test datasets (ratio of 80:20), and a model was built/trained with the train dataset and then tested with the test dataset, this gave an accuracy of 93% as that of the AdaBoost model, but a better CV_score: 0.9110, R2_score: 0.7078, AUC curve: 0.98, RSME: 0.2701, MAE: 0.0730 with Random Forest classifier.

Keywords: Random Forest Classifier; Logistic Regression Classifier; Sequential Feature Selection; AdaBoost and Bagging; Support Vector Machine Classifier; XGBoost classifier

Abbreviations

RF: Random Forest; LR: Logistic Regression; SFS: Sequential Feature Selection; SVM: Support Vector Machine; CV: Cross Validation

Introduction

The heart is the most important organ in the human body. One of the many important functions is that as it beats, it sends blood which in turn sends oxygen and nutrients to all parts of the human body and then removes all unwanted wastes and carbon dioxide from the body. However, the major cause of morbidity and mortality is heart disease causing over 70% of fatalities [1]. In 2017, Global Burden of Disease research declared that 43% of every fatality is from heart disease [2]. In New Zealand, over 180,000 people have heart disease and claim one life every 90 minutes [14]. Li J. [10] analyzed that annual medical expenditures of 25% to 30% of all organizations were channeled to workers with heart disease. According to WHO, by 2030, deaths from Cardiovascular Diseases will reach 23.6 million where stroke and heart disease taking the lead [3]. Also, 10 million people die of heart disease annually, globally, by WHO. More than four of every five deaths from heart disease are caused by heart attacks or strokes, and one-third of these happen before the age of 70 [7].

Smoking, age, medical history, alcohol, ice, sugar consumption, excess body fat, or obesity (from high-income countries) like the US where 87% of deaths have been said to be as a result of chronic

heart disease [7]. In low and middle-income countries, attributes like undernutrition, habits, and unhealthy diets are said to be the cause of this disease [4]. Globally, economic stress from heart disease was predicted to Skyrocket to USD 3.7 trillion between 2010 and 2015 [5].

Angiography is one of the traditional ways of detecting heart disease, but it comes with disadvantages including high costs, a variety of side effects, and demands for significant technological expertise. Electrocardiograms and other technologies for critical checkups for hearts are also very expensive and almost unachievable for the common masses, and according to statistics, 17 million individuals have lost their lives to this incapacitation [4]. Non-invasive techniques can be used to get around the drawbacks of these conventional techniques by leveraging on the use of machine learning techniques for early detection of the disease and risks to lessen financial implications on all parties, individuals, society, and organizations as the case may be. Also, to save lives and avoid more calamitous consequences. According to a WHO survey, only 67% of the time, doctors can accurately predict heart disease.

In the Data Science space, Machine Learning algorithms are being used to deliver insightful information and guide decision-making in a variety of sectors [11]. According to [14], machine learning (ML) is an automated technique that computers employ to learn from data, find meaningful patterns, and reduce human intervention in the decision-making process.

Summary of key contribution of this research

- A precise coronary heart risk prediction system has been created using machine-learning algorithms.
- The performance of ensemble classifiers like AdaBoost, Bagging as well as single classifiers like XGBoost, Random Forest, Support Vector Machines (SVM), and Logistic Regression has been investigated both with and without hyperparameter optimization like Randomized SearchCV, and Grid SearchCV.
- Performance comparison between the suggested strategy and the most recent research on coronary heart disease risk prediction.

Related works and literature review

Experts, researchers, and the data science community have launched many research projects to predict and screen medical data for heart diseases. These predictions have been made in the past using a variety of machine-learning techniques. Evaluation of pertinent research publications.

Avinash Golande., *et al.* [6] used different ML algorithms for the classification of heart disease, Decision Tree gave the best accuracy but recommended that it can be better by combining different techniques and tuning of parameters.

Using a dataset with 13 variables, Dangari and Apte [8] prediction of heart disease was conducted. Smoking and obesity were two other characteristics that the writers mentioned. The results showed that the ANN classification technique had the highest predicted accuracy on the used dataset when compared to DT and NB.

Using a public dataset of 573 records, Karthiga., *et al*. [12] conducted research to successfully predict existing heart disease. The DT and NB classification methods were used by the authors to process the dataset. And replaced all missing variables with new ones using the MATLAB data analysis tool, and then they created accuracy results to assess the performance of the models. According to their results, DT gave higher accuracy than NB with the dataset under consideration.

Fahd [9] used just one model decision three with 10-fold cross-validation, after comparing among five different algorithms. They used the Rapid Miner tool which led to high accuracy of 93.19% as compared to Matlab and Weka tools.

Hasan and Bao [11] used the three feature selection approaches (Filter, Wrapper, and Embedding) to select the important features, then used five classifier models and compared their accuracies. The XGBoost coupled with the wrapper method has the highest accuracy of 73.74%.

Theresa and Thomas [13] conducted a survey that used various classification algorithms for heart disease prediction. Examination of the accuracy of the classifiers for different numbers of attributes using Naive Bayes, KNN, Decision trees, and Neural Networks as the classification algorithms. Decision Tree performed best.

Heart disease prediction was carried out by Nagaraj., *et al*. [15] utilizing Naive Bayes classification and SVM (Support Vector Machine). Squared Error's sum, Mean Absolute Error, and Root Mean Squared Error are the performance metrics employed in the investigation, and SVM outperforms Naive Bayes in their accuracies.

A model that employs the Naive Bayesian approach for categorizing datasets and the Advanced Encryption Standard algorithm for transporting data safely was proposed by Anjan., *et al*. [17] and achieved an accuracy of 89.77 using Naïve Bayes classifier which was the highest in 0.1 secondss.

Through the combination of Cat Swarm Optimization, Cuckoo Search, and Crow Search Algorithm, Mohamed., *et al*. [16] built a technique of meta-heurism named Parasitism-Predation Algorithms (PPA) for feature selection to increase the accuracy of the classifier. Upon running KNN classifier with cross-validation of 10 on the features that were chosen and retrieved from the Statlog heart dataset using, the PPA obtained an accuracy of 86.17% in 49.13 seconds.

Through the use of cross-validation of 10 folds also, Muhammad., *et al*. [18] trained a variety of machine learning classifiers on the most advantageous Cleveland dataset attributes to produce a predictive model for the early detection of heart disease. To identify the crucial and more correlated features, feature selection algorithms named Minimal Redundancy Maximal Relevance, Relief, Fast Correlation-Based Filter, and LASSO were used. The Extra Tree classifier improved accuracy by 94.41%.

Alam., *et al*. [20] used InfoGain, GainRatio, OneR, and Relief for selection of features, with cross-validation of 10, Random Forest gave the highest accuracy of 85.50%, sensitivity of 85.60 and AUC of 0.915.

Methodology

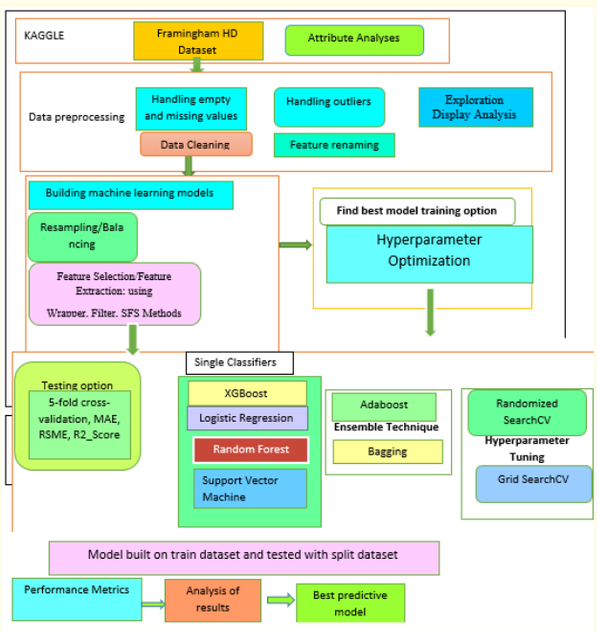


Figure 1: Above is showing the methodology of the research.

This research involves the extensive application of machine learning techniques to predict Heart Disease.

Data description

The research utilizes a dataset obtained from Kaggle called Framingham HD dataset. It contains 4,240 instances and 16 attributes. Approximately, it has a ratio of 57 to 43 females to males. Altogether, the dataset has 644 missing values. The name of a feature called 'male' was changed to 'sex'.

Data pre-processing and feature engineering

This involved elimination of noise and conversion to a suitable format for analysis. The dataset is unbalanced, has an 85% to 15% ratio of response of No to Yes (of Ten Year Cardiovascular Heart Disease - TenYearCHD), which is 0 to 1, this was later balanced up using the Smote resampling method. Exploration Display Analysis (EDA) was done using Pearson Correlation, which revealed weak correlations between independent features and the target feature. There was data transformation for all the numeric features for the relationship of the frequency of each feature with the target variable (TenYearCHD). Outliers were removed using the Multiple Interquartile Range (IQR) method and visualized using pairplot and boxplot accordingly.

Data cleaning

The Framingham heart disease has missing values of 644 as written above, these were filled with interpolation, forward, and backward filling methods; the application of any of these was dependent on the type of feature to be cleaned. Continuous features were done using the interpolation method and binary features were filled using forward and backward respectively. The data set has a wrong feature called 'male' instead of 'sex' as it comprises of the gender of the patients, and was corrected to 'sex'. The dataset has a ratio of 57% to 43% of females to males.

Selection of important features

Important features were selected using filter (with KBest), Wrapper (with XGBoost Classifier), and Sequential Feature Selector -SFS (using the forward selection).

Several machine learning algorithms will be considered for model development, building on the lapses of previous authors, including Decision Trees, Support Vector Machines, XGBoost, etc. Each algorithm will be trained and fine-tuned using a training dataset, and its performance will be assessed using cross-validation techniques.

Development of model and hyper-tuning

Multiple machine learning algorithms, including Random Forest, XGBoost, Logistic Regression, and Support Vector Machine (SVM) are used to train predictive models. The models are evaluated using various metrics such as confusion matrix, accuracy, F1-score, recall, ROC Curve, MAE, RMSE, R2-Score, and Cross-Validation involving 5 and 10 folds. The models were hyper-tuned with Randomized SearchCV and Grid SearchCV approaches. Also, ensemble techniques (AdaBoost and Bagging) were employed for the improvement of the performance of the models. In each model, the dataset was split into test and train. But separately, the whole dataset was split into train and test dataset, model built on it and results generated.

Discussion and Results

Data transformation for the continuous features was done by putting them into range and bins for the sake of visualizing the relationship between the independent and target/dependent variables (TenYearCHD).

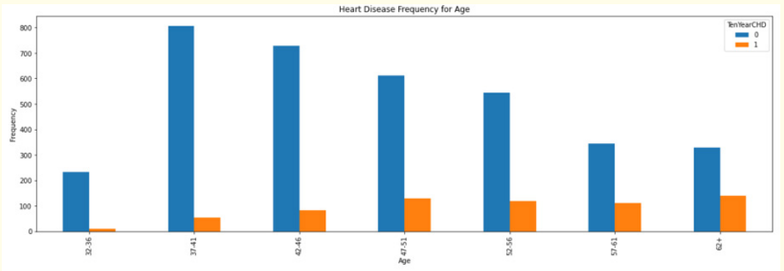


Figure 2: Frequency of Age against the TenYearCHD.

From this, it was inferred that there were more people between the ages of 37 to 41, although patients who were above the age of 62 had heart disease. From the other features' charts, there were more men with heart disease than women. People smoking 16 to 20 sticks of cigarettes per day are more and live more with heart disease. Patients with 201 to 205 total cholesterol were more and had more heart disease. Patients with 71 to 80 diastolic blood pressure are more, and the live more with heart disease for ten years. BMI values of 21 to 25 were higher and people here live more with heart disease. Patients with glucose values of 76 to 100 are more and they have more heart disease. Patients with no Blood Pressure medication are more with heart disease than the ones who use the medication. Patients with no Prevalent Stroke live more with heart disease. People with no hypertension are more but those with hypertension live more with heart disease. Lastly, people with no diabetes are more living with heart disease than those with no diabetes.

Exploratory display analysis

According to Pearson's correlation, the features have a very low correlation. The strongest (age) to the target variable shows a correlation value of 0.225. Which was still really poor.

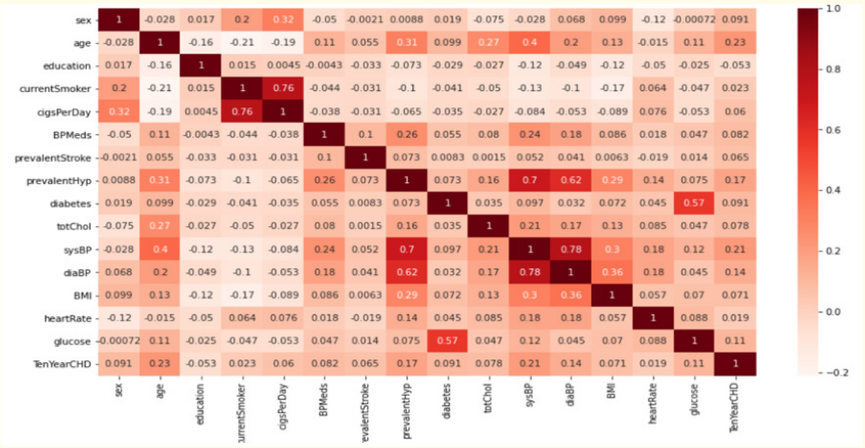


Figure 3: Exploratory Display Analysis of the dataset.

Outliers

The distribution of the whole dataset was visualized, Multiple Interquartile Range (IQR) technique was used to get rid of outliers after proven and properly calculated. The features with outliers are visualized before and after removal.

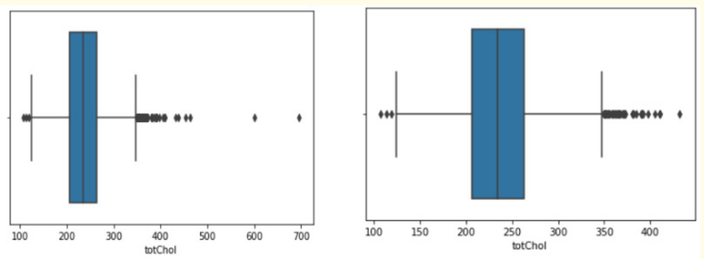


Figure 4: Shows Outliers in Cholesterol and when removed respectively.

The Framingham dataset is an unbalanced dataset with 85% to 15% of 'no' to 'yes' in the target variable (response to Ten years of Heart Disease). Four model classifiers (XGBoost, Random Forest, LogisticRegressiom, and Support Vector Machine) were used. Models were built on the unbalanced dataset and the results shown below were obtained (Classification Report).

Metrics \ Classifiers		XGBoost	Random Forest	Logistic Regression	SVM
Accuracy (%)		83	84	85	85
Precision	Class 0	0.86	0.85	0.85	0.85
	Class 1	0.37	0.43	0.75	0.75
Recall	Class 0	0.96	0.99	1.00	1.00
	Class 1	0.14	0.05	0.05	0.02
F1 Score	Class 0	0.90	0.91	0.92	0.92
	Class 1	0.20	0.08	0.09	0.04

Confusion Matrix Analysis:

Classifiers	XGBoost	Random Forest	Logistic Regression	SVM
True Positive (TP)	685	708	714	715
True Negative (TN)	18	6	6	3
False Positive (FP)	114	126	126	126
False Negative (FN)	31	8	2	1

Figure a: Showing classification report and confusion matrix results before balancing the dataset.

The Receiver Operating Characteristic curve/result for each classifier is shown below:

	XGBoost	Random Forest	Logistic Regression	SVM
Area Under Curve (AUC)	0.65	0.70	0.72	0.60

MAE, RSME, R2_Score and Average CV Score

	XGBoost	Random Forest	Logistic Regression	SVM
MAE	0.1665	0.1511	0.1467	0.1521
RMSE	0.4081	0.3888	0.3830	0.3900
R2 Score	-0.2926	-0.1736	-0.1388	-0.1809
Mean CV Score	0.8335	0.8478	0.8533	0.8488

Figure b: Shows AUC values and MAE, RSME, R2_Score and Average CV_Score of the classifiers.

From the above tables, the Logistic Regressor gave the best performance and metrics, although it has same accuracy as SVM but gave a better recall and F1_score. It classified the classes better also. The MAE, R2_Score and RMSE values are the smallest while the Mean CV score is the highest.

Selecting important features using kbest method – filter method

However, the results from the model were not encouraging, hence several attempts were made to improve the performance of the baseline model. From the unbalanced dataset, filter method (KBest) was used to select important features.

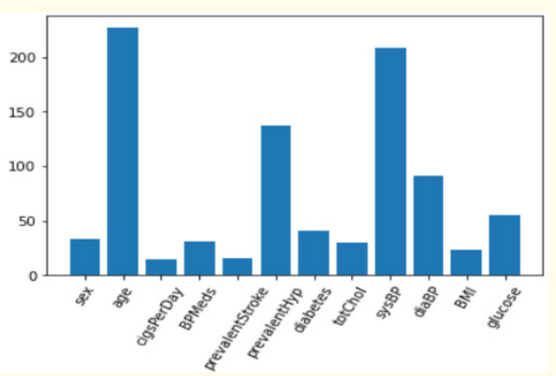


Figure 5: Important features selected from KBest approach.

A model was built using the important features from KBest, using the four classifiers. The best-performing classifier was the random Forest with 85% accuracy.

Classifier: Random Forest

	precision	recall	f1-score	support
0	0.85	0.98	0.91	1077
1	0.47	0.08	0.13	195
accuracy			0.85	1272
macro avg	0.66	0.53	0.52	1272
weighted avg	0.80	0.85	0.79	1272

Figure 6: Shows the results of the best classifier (RF) from the model run on KBest selected features.

But a drop in AUC value (ROC) to 68%. MAE: 0.1540, R2_Score: -0.1956, Mean CV Score: 0.8456, RSME: 0.3924.

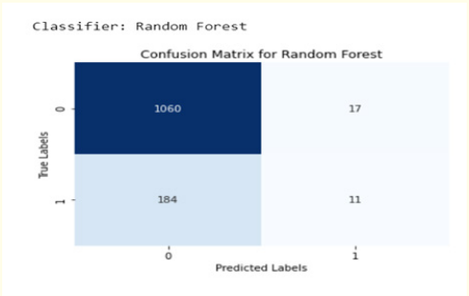
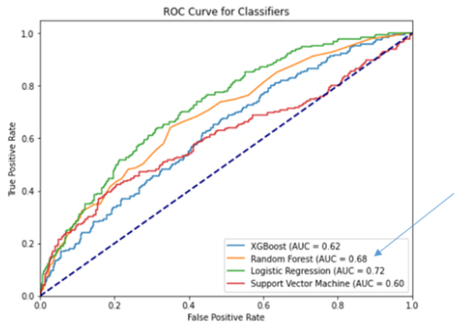


Figure 7: Showing True Positive of 1,067, True Negative of 10, False Positive of 185, False Negative of 10 for Random Forest for KBest selected features.



Random Forest Classifier for Important Features from KBest			
MAE	RMSE	R2_Score	Mean CV Score
0.1540	0.3924	-0.1956	0.8456

Figure 8: Shows AUC curve and metrics of Random Forest classifier for KBest model.

Domain knowledge

Using domain knowledge, 'education' was dropped from the features. Logically, from a perspective, 'education' or one's class of education does not matter regarding having heart disease or not, hence another model was built with the education feature dropped. Here, Logistic Regression gave the best accuracy, 86%.

Classifier: Logistic Regression

	precision	recall	f1-score	support
0	0.86	0.99	0.92	1077
1	0.70	0.11	0.19	195
accuracy			0.86	1272
macro avg	0.78	0.55	0.55	1272
weighted avg	0.84	0.86	0.81	1272

Confusion matrix

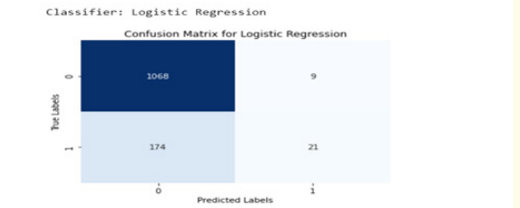
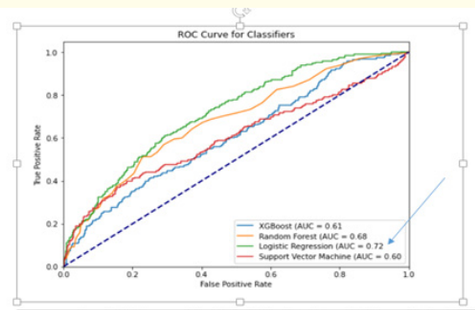


Figure 9: Classification report and confusion matrix for logistic regression.
TP = 1,068, TN = 21, FP = 174, FN = 9.

From above, logistic regression classified more labels than the random forest classifier from the previous model.

ROC curve



Logistic Regression Classifier for Domain Knowledge Model			
MAE	RMSE	R2_Score	Mean CV Score
0.1474	0.3839	-0.1443	0.8526

Figure 10: Shows the AUC value and other metrics for Logistic Regression for important features selected using domain knowledge.

The wrapper method

Using XGBoost classifier) was used to select important features and the features visualized below were selected.

Building a model with these features, Random Forest and Logistic Regression gave the best performance with an equal accuracy of 85%, but slightly, Logistic Regression classified better than the Random Forest according to the confusion matrix. Also, the precision and recall are better in Logistic Regression for classes 0 and 1.

From the above table of metrics, and comparison between the two best-performing classifiers, Logistic Regression performed better than the Random Forest and best for the selected features model for the Wrapper Method.

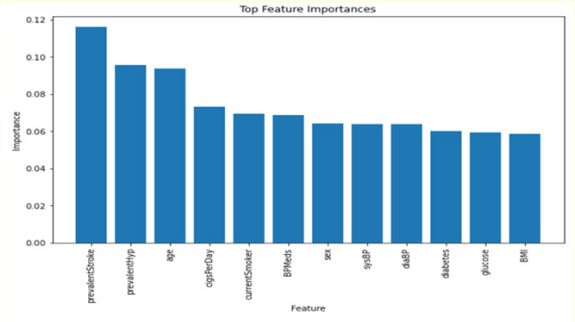


Figure 11: Shows important features from XGBoost classifier.

Metrics \ Classifiers		Random Forest	Logistic Regression
Accuracy (%)		85	85
Precision	Class 0	0.86	0.86
	Class 1	0.47	0.67
Recall	Class 0	0.98	0.99
	Class 1	0.08	0.09
F1 Score	Class 0	0.91	0.92
	Class 1	0.13	0.16

Confusion matrix for

	Random Forest	Logistic Regression
TP	1063	1068
TN	17	16
FP	178	179
FN	14	9

Figure c: Shows the classification report and confusion matrix of logistic regression and random forest classifiers from important features selected using XGBoost classifier.

MAE, RSME, R2_Score, and Average CV Score

Classifiers	MAE	RMSE	R2_Score	Mean CV Score
Random Forest	0.1531	0.3912	-0.1882	0.8483
Logistic Regression	0.1476	0.3842	-0.1461	0.8524

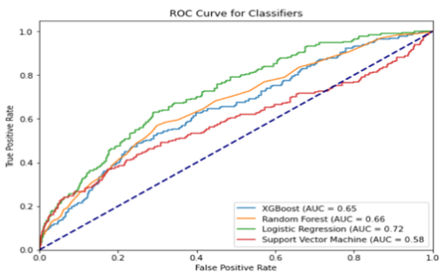


Figure d: Shows the metrics of Random Forest and Logistic Regression and AUC values.

Sequential feature selector (SFS – Wrapper Method)

Also, important features were selected using the Sequential Feature Selector (wrapper method) using the forward feature selection method. Again, the model classifier with the best performance here is logistic regression with accuracy of 85%.

From the SFS model, the best metrics were obtained from the Logistic Regression classifier with the above values.

Metrics \ Classifier		Logistic Regression
Accuracy (%)		85
Precision	Class 0	0.85
	Class 1	0.59
Recall	Class 0	0.99
	Class 1	0.05
F1 Score	Class 0	0.92
	Class 1	0.09

Figure e: Shows the classification report of Logistic Regression from SFS model.

Confusion Matrix

	Logistic Regression
TP	1070
TN	10
FP	185
FN	7

Other metrics:

Classifier	ROC Value	MAE	RMSE	R2_Score	Mean CV Score
Logistic Regression	0.68	0.1486	0.3855	-0.1535	0.8514

Figure f: Shows confusion matrix and other matrix of logistic regression.

RESAMPLING/Balancing of dataset

Furthermore, the dataset was balanced using 'Smote' and then, a model was built.

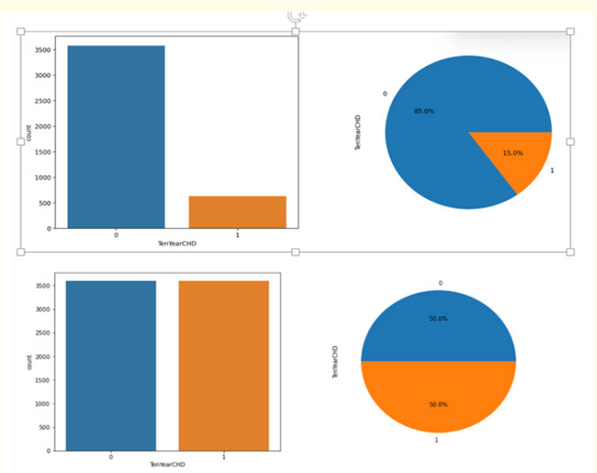


Figure 12: Shows the dataset before and after balancing.

The performance of the model increased numerically. The metrics in the tables below were obtained.

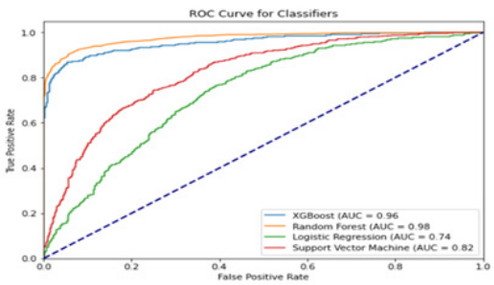
From the classification report, confusion matrix, ROC Curve, MAE, RSME, R2_Score, and CV_Score tables, the Random Forest Classifier showed the best metrics, it classified and performed better than all of the other classifiers. It gave the smallest MAE and RSME values and highest R2_score and CV Score.

Metrics \ Classifiers		XGBoost	Random Forest	Logistic Regression	SVM
Accuracy (%)		90	92	67	74
Precision	Class 0	0.89	0.92	0.69	0.77
	Class 1	0.92	0.92	0.66	0.71
Recall	Class 0	0.93	0.92	0.67	0.70
	Class 1	0.87	0.92	0.68	0.77
F1 Score	Class 0	0.91	0.92	0.68	0.73
	Class 1	0.89	0.92	0.67	0.74

Confusion Matrix

Classifiers	XGBoost	Random Forest	Logistic Regression	SVM
True Positive (TP)	690	690	497	522
True Negative (TN)	605	640	474	534
False Positive (FP)	89	54	220	160
False Negative (FN)	55	55	248	223

Figure g: Shows the classification report and confusion matrix of the classifiers after the dataset was balanced.



	XGBoost	Random Forest	Logistic Regression	SVM
MAE	0.1260	0.0897	0.3209	0.2564
RMSE	0.3549	0.2995	0.5665	0.5064
R2 Score	0.4962	0.6413	-0.2836	-0.0256
Mean CV Score	0.8741	0.9119	0.6791	0.7463

Figure h: Shows the AUC values of the classifiers and other metrics used.

Hyperparameter tuning

There were different hyper-tunings done. The first was Randomized Search CV. For this, each classifier parameter was tuned, and CV scores were obtained at 5 folds. The whole classifiers improved in their performance. However, in all, the Random Forest gave the best performance, its metrics are shown below.

Grid Search CV was also used. Random Forest gave the best accuracy which was 93% again. The figures shown below give the metrics.

From the generated metrics, it performed best among the other classifiers.

Metrics \ Classifier		Random Forest
Accuracy (%)		92
Precision	Class 0	0.92
	Class 1	0.92
Recall	Class 0	0.92
	Class 1	0.91
F1 Score	Class 0	0.92
	Class 1	0.92
CV Score		0.9124
ROC Curve		0.98

Classifiers	Random Forest
True Positive (TP)	689
True Negative (TN)	636
False Positive (FP)	58
False Negative (FN)	56

Figure i: Shows the classification report of the best classifier (Random Forest) and the confusion matrix for Randomized Search CV.

```
Classifier: Random Forest
Best Hyperparameters: {'max_depth': 30, 'n_estimators': 200}
Best Cross-Validation Score: 0.9116991651871718
              precision    recall  f1-score   support

           0       0.92      0.94      0.93       745
           1       0.93      0.92      0.92       694

    accuracy                           0.93      1439
   macro avg       0.93      0.93      0.93      1439
weighted avg       0.93      0.93      0.93      1439
```

Figure 13: Shows classification report of the grid search model for the classifier (Random Forest) with best metrics.

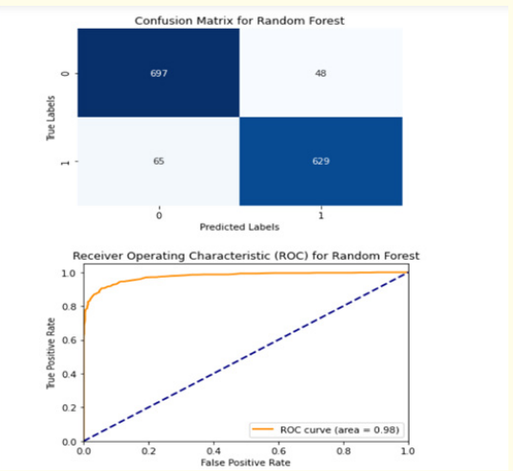


Figure 14: Shows confusion matrix and AUC value of the grid search model for Random Forest.

Ensemble techniques

Ensemble techniques (Bagging and AdaBoost) were employed. For Bagging, the classifier with best performance was XGBoost with an accuracy of 91%.

Metrics \ Classifier		XGBoost
Accuracy (%)		91
Precision	Class 0	0.89
	Class 1	0.93
Recall	Class 0	0.94
	Class 1	0.87
F1 Score	Class 0	0.91
	Class 1	0.90
CV_Score		0.8867
ROC Curve		0.96

Confusion matrix table for Bagging model

Classifiers	XGBoost
True Positive (TP)	697
True Negative (TN)	606
False Positive (FP)	88
False Negative (FN)	48

Figure j: Shows the XGBoost classification reports and confusion matirx for Bagging model.

After running the **AdaBoost** model, Random Forest has the best performance with accuracy of 93%.

Metrics \ Classifier		Random Forest
Accuracy (%)		93
Precision	Class 0	0.93
	Class 1	0.93
Recall	Class 0	0.93
	Class 1	0.92
F1 Score	Class 0	0.93
	Class 1	0.92
CV_Score		0.9080
MAE		0.0920
R2 Score		0.6322
ROC Curve		0.98
RSME		0.3032

Classifiers	Random Forest
True Positive (TP)	695
True Negative (TN)	639
False Positive (FP)	55
False Negative (FN)	50

Figure k: Shows the classification report and confusion matrix of Random Forest from Adaboost model.

Splitting the whole dataset into train and test

The whole dataset was split into train and test datasets (ratio of 80:20). There was data transformation of the two sets of datasets. The training dataset was used to get the charts of the actual values of each feature at their actual average points in the target variable (TenYearCHD). A model was built on the whole train dataset (the whole x_train and y_train datasets altogether concatenated) and the metrics were obtained by testing the model with the test dataset (the whole x_test and y_test datasets concatenatedalso). The test dataset was used to get the charts for actual and predicted values against their average values on the target variable (TenYearCHD). The model classifier with best metric was Random Forest. Accuracy - 93%.

Metrics \ Classifier		Random Forest
Accuracy (%)		93
Precision	Class 0	0.93
	Class 1	0.94
Recall	Class 0	0.94
	Class 1	0.92
F1 Score	Class 0	0.93
	Class 1	0.93
Aveage CV_Score		0.9110
MAE		0.0730
ROC Curve		0.98
RMSE		0.2701
R2_Score		0.7078

Classifiers	Random Forest
True Positive (TP)	692
True Negative (TN)	639
False Positive (FP)	55
False Negative (FN)	53

Figure l: Shows the classification report and confusion matrix of the split dataset model.

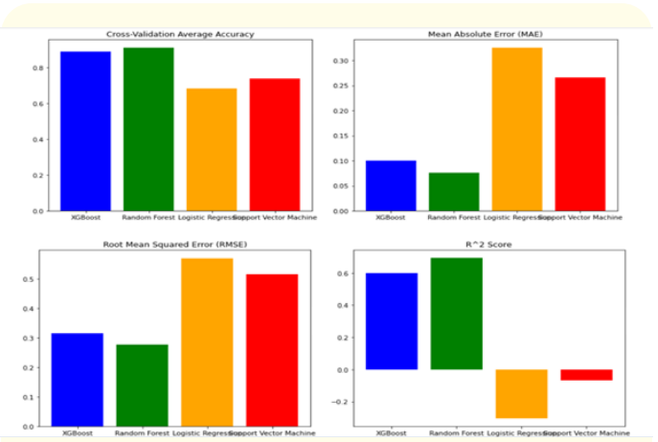


Figure 15: Shows the bar chart of the metrics used for all the Classifier used in the split dataset model.

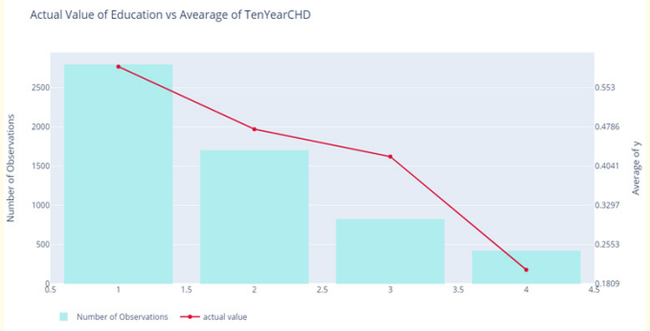


Figure 16: Shows Actual value for education vs. average of TenYearCHD using train dataset.

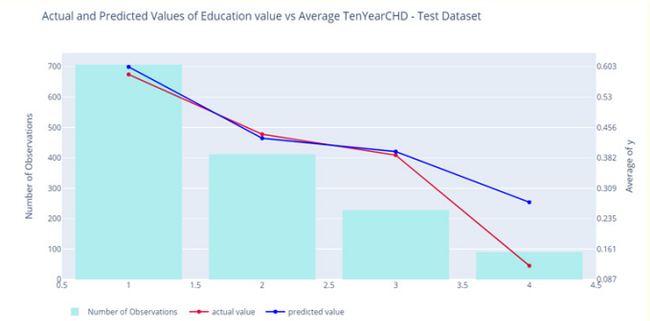


Figure 17: Shows Actual and predicted values for education vs average of TenYearCHD using the test dataset.

The above shows that the model did not predict well some values in some features like 4 in 'Education' and many other.

Comparison of the models' results with the best performance

Ensemble technique (Adaboost) and the model built when the dataset was split into test and train gave best metrics. They both gave an accuracy of 93%.

Metrics \ Model/Classifier		Adaboost/Random Forest	Split Dataset Model/Random Forest
Accuracy (%)		93	93
Precision	Class 0	0.93	0.93
	Class 1	0.93	0.94
Recall	Class 0	0.93	0.94
	Class 1	0.92	0.92
F1 Score	Class 0	0.93	0.93
	Class 1	0.92	0.93
CV_Score		0.9080	0.9110
MAE		0.0920	0.0730
R2 Score		0.6322	0.7078
ROC Curve		0.98	0.98
RSME		0.3032	0.2701

Confusion Matrix

Model/Classifiers	Adaboost/Random Forest	Split Dataset Model/Random Forest
True Positive (TP)	695	692
True Negative (TN)	639	639
False Positive (FP)	55	55
False Negative (FN)	50	53

Figure m: Shows the best results comparison between AdaBoost and Split dataset models

In all, Random Forest classifier performed best. Considering the precision, recall/sensitivity, and F1_score, the model with the split dataset gave the best performance. Also, with the CV-score, MAE, R2_Score, and RSME, it gave the best result. Although the classification of labels according to the confusion matrix shows that Adaboost classified better.

The model was checked if it overfit, but it did not.

```
Training Accuracy: 0.85
Test Accuracy: 0.86
The model does not appear to be overfitting.
```

Discussion

Comparison of results with past authors

Authors	Classifiers	Accuracy	Sensitivity	Specificity	AUC	CV Score	MAE	RSME	R2 Score
Karthiga., *et al.*	DT	98.28%	95.45	97.79	-	-	-	-	-
Muhammad., *et al.*	Extra Tree	94.41	94.93	94.89	0.942	-	-	-	-
Dangari and Apte	ANN	99.25%	-	-	-	-	-	-	-
Fahd S. A.	DT	93.19%	-	-	-	-	-	-	-
Alam and Rahman	RF	85.50%	85.60	-	0.915	-	-	-	-
Avinash Golande., *et al.*	DT	99.62	-	-	-	-	-	-	-
Hasan and Bao	XGBoost + wrapper method	73.74%	-	-	-	-	-	-	-
Chiradeep Gupta., *et al.*	LR	92.30%	96.05	87.50	-	-	-	-	-
Reddy K. V. V., *et al.* [19]	RF	97.91%	97.91	97.66	0.996	-	-	-	-
Current research	RF	93.0%	92.V	91.0	0.98	0.9110	0.0730	0.2701	0.7078

Table 1: Shows the results of this research and comparison with other authors'.

Limitations and potential areas for future research

From above, the model was not predicting 4 of education (this could be tertiary if assuming that 1 is nursery education, 2 is primary, 3 is secondary and 4 is tertiary, but 4 could also be primary if the case was reversed. This was not explicitly stated in the dataset dictionary). There are more values in other features with the same challenge. Also, there are monotonic relationships (implies that as the values of the features increase, the values of the target variable also consistently increase or decrease with the decrease in the feature) among 8 features against the target variable. In a way, it influences the performance of models. This could be dealt with, with any approach or "MonotonicConstraint". Lastly, for future purposes, new feature selection methods and any other hyperparameter optimization may be taken into consideration to enhance the machine learning models' performance.

Conclusion

The fast rising of death records as a result of heart disease has made it very importunate to establish medium that reliably predict this disease. The project's main goal is to, more accurately, predict heart disease. In all of the models built for the research, the Adaboost model and model_8 gave the best accuracy (93% each), but other metrics like MAE, Area Under Curve, CV_score, etc. showed that model_8 gave the best metrics. In all of the four classifiers, RF performed best. Model_8 did not show any signs of overfitting. Train and test datasets were used to visualize the charts of actual and predicted instances of every feature against the target feature, it was deduced that some unique instances of some features were not predicted by the model properly, and a monotonic relationship was seen among some (eight) of the features against the target features. This is aforementioned written to be dealt with using any "Monotonic constraint" approach – for future work.

Bibliography

1. Bhatt CM., *et al*. "Effective Heart Disease Prediction Using Machine Learning Techniques". *Algorithms* 16 (2023): 88.
2. Estes C., *et al*. "Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030". *Journal of Hepatology* 69 (2018): 896-904.
3. Purushottam Saxena K and Sharma R. "Efficient Heart Disease Prediction System". *Procedia Computer Science* 85 (2016): 962-969.
4. Vardhan Shorewala. "Early detection of coronary heart disease using ensemble techniques". *Informatics in Medicine Unlocked* 26 (2021): 100655.
5. Mozaffarian D., *et al*. "Heart disease and stroke statistics—2015 update: A report from the American Heart Association". *Circulation* 131 (2015): e29-e322.
6. Avinash Golande and Pavan Kumar T. "Heart Disease Prediction Using Effective Machine Learning Techniques". *International Journal of Recent Technology and Engineering* 8 (2019): 944-950.
7. H Schmidt. "Chronic disease prevention and health promotion" (2016).
8. Apte C S. "Improve study of Heart Disease prediction system using Data Mining Classification techniques".
9. Fahd Saleh Alotaibi. "Implementation of Machine Learning Model to Predict Heart Failure Disease". *(IJACSA) International Journal of Advanced Computer Science and Applications* 10.6 (2019).
10. Li J., *et al*. "Work stress and cardiovascular disease: A life course perspective". *Journal of Occupational Health* 58 (2016): 216-219.
11. Hasan N and Bao Y. "Comparing different feature selection algorithms for cardiovascular disease prediction". *Health Technology* 11 (2020): 49-62.
12. Karthiga A S., *et al*. "Early Prediction of Heart Disease Using Decision Tree Algorithm". *International Journal of Advanced Research in Basic Engineering Sciences and Technology* (2017).
13. Theresa Princy R and J Thomas. "Human heart Disease Prediction System using Data Mining Techniques". International Conference on Circuit Power and Computing Technologies, Bangalore (2016).
14. Amanda H Gonsalves., *et al*. "Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis". In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (ICDLT '19). Association for Computing Machinery, New York, NY, USA (2019): 51-56.
15. Nagaraj M Lutimath., *et al*. "Prediction Of Heart Disease using Machine Learning". *International Journal Of Recent Technology and Engineering* 8.2S10 (2019): 474-477.
16. Mohamed AAA., *et al*. "Parasitism—Predation algorithm (PPA): A novel approach for feature selection". *Ain Shams Engineering Journal* 11 (2020): 293-308.
17. Anjan Nikhil Repaka., *et al*. "Design And Implementation Heart Disease Prediction Using Naives Bayesian". International Conference on Trends in Electronics and Information (ICOEI 2019).
18. Muhammad Y., *et al*. "Early and accurate detection and diagnosis of heart disease using intelligent computational model". *Scientific Report* 10 (2020): 19747.
19. Reddy KVV., *et al*. "An Efficient Prediction System for Coronary Heart Disease Risk Using Selected Principal Components and Hyperparameter Optimization". *Applied Science* 13 (2023): 118.
20. Alam Z and Rahman MS. "A Random Forest based predictor for medical data classification using feature ranking". *Informatics in Medicine Unlocked* 15 (2019): 100180.