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
Department of Computational Science and Engineering
North Carolina Agricultural and Technical State University

ABSTRACT
Heart disease, an example of cardiovascular diseases is the number one notable reason for the death of many people in the world. Of recent, studies have concentrated on using alternative efficient techniques such as data mining and machine learning in the diagnosis of diseases based on certain features of an individual. This study will use data exploratory and mining techniques to extract hidden patterns using python. By this, machine learning algorithms (logistic linear regression, decision tree classifier, Gaussian Naïve Bayes models) will be developed to predict the presence of heart diseases in patients. This will try to seek better performance in predicting heart diseases to reduce the number of tests require for the diagnosis of heart diseases. The k-fold cross validation approach will be used in assessing the resulting models for receiver operating characteristic (ROC) curves (sensitivity against specificity). The dataset was collected from UCI machine learning repository which contains information on patients with heart disease. The dataset has 14 attributes and measured on 303 individuals.

General Terms
Algorithms, pattern recognition, supervised learning, machine learning, heart disease.

Keywords
Classification, regression, k-fold cross validation, Receiver Operator Characteristics

1. INTRODUCTION
Heart diseases such as heart failure, myocardial infarction have been ranked as the highest cause of death in the United States [1]. According to the center of disease control and prevention, about 610,000 people die every year from heart diseases. Health professionals have put measures in place for early detection of heart diseases, as this is key in preventing and curbing them. However, early-stage adopted strategies in identifying heart diseases have not been successful, due to the associated complexities [2]. According to [3] , unlike traditional statistical methods, data mining techniques can detect and extract hidden inconspicuous patterns, relationships in large dataset. Support vector machine, Naïve Bayesian, artificial neural networks, logistics regression, etc., models have been developed and used in healthcare research [4][5]. They have shown immense potential in accurate prediction of heart diseases based on clinical data of patients [2].

In the diagnosis of heart diseases, series of laboratory tests are required. However, the numerous tests impede the rapidity and efficiency in diagnosing heart diseases in patients. Data mining techniques provide alternative approach for quick and efficient detection of heart diseases at the early stages. The primary objective of this project is to develop three different classification models, Gaussian Naïve Bayes models (GNB), Logistic Regression (LR) and decision tree classifier (DCT) to predict heart diseases in patients based on clinical data sample trained and tested. Also, the models' performance efficiencies were evaluated and compared using accuracy scores, 10-fold cross validation, and area under the curve receiver operating curves (AUCROC).

Age, cholesterol, family history among other factors are considered risk factors for heart diseases. Early identification of heart diseases among patients can reduce the fatality that is associated with them. Many research studies have involved the use of machine learning in predicting heart diseases among patients [6][7] . However, findings have differed in the metrics used in evaluating models, culminating in differences in accuracies. Some research work had involved the development of ML algorithms using the Cleveland dataset which is been used in this projected. [3] developed J48, Logistic model tree and Random Forest algorithms. The highest accuracy score was found in the J48 algorithm (56.76%),and the least in Logistic model tree algorithm (55.77%). [8] also developed Classification and Regression Tree and Iterative Dichotomized 3 (ID3), and Decision Table based on this dataset with the models scoring accuracies of 83.49%, 72.93%, and 82.50%, respectively. However, they adopted feature selection, leaving out important features such as number of major vessels (0-3) colored by fluoroscopy (ca), ST depression induced by exercise relative to rest (oldpeak). [9] observed one common problem, that is, many authors have different parameters for testing the accuracies of their models. This has made it difficult to be conclusive on the best model.

2. METHOD
Three classification models namely; Linear Regression (LG), Decision Tree Classifier (CART) and Gaussian Naïve Bayes (GNB) were developed. The data was analyzed and implemented in python. Data preprocessing techniques such as, feature transformation, and training, testing with the individual models, and finally comparison of the performance of the models were the steps followed through to achieve the aim this research.

2.1 Preprocessing
The dataset called the Cleveland Heart Diseases was collected from UCI machine learning repository which contains information on patients with heart disease. The dataset has 14 attributes including patients age, sex, cholesterol level, etc.

which was measured on 303 individuals. (i) The first step of the preprocessing was identifying and removing duplicated rows. (ii) Subsequently, rows containing missing values were also identified and removed. (iii) Also, the box and whisker plots were used to detect outliers, and the (iv) rows with outliers (i.e. values that are outside the range of -3δ and +3δ) were subsequently removed. (v) One duplicated row, and 15 rows containing outliers were removed.

2.1.1 Data Transformation

The following features were normalized age, thalach, and oldpeak to range between 0 and 1. This was done before the feature reduction process. The other features were not normalized because either they are categorical, or they are already gaussian.

2.1.2 Feature Reduction

The aim of feature reduction has been searching for a projection of the data on features which preserve the information, pattern and trend as much as possible (Hira & Gillies, 2015). In this study, single value decomposition (SVD) method was used to construct enriched features in the data. The features were simplified from thirteen (13) to four (4) features using SVD.

2.2 Machine Learning Model Development

Three classification models, Decision tree classifier, Logistic regression, Gaussian Naïve Bayes models. The three classifications models were trained to find the best fit for the models by splitting the data into trained and test dataset. The training dataset is used to fit the model. In splitting data into training and testing sets, it is important to avoid bias. The most efficient training method is using the k-fold cross validation. The 10-fold cross validation resampling method was adopted in training and testing of the model. The dataset was split into 10 folds. The first iteration uses the data in the 1st fold to test the model while the remaining 9 folds are used to train the model. The second iteration uses the 2nd fold as the test set, and the remaining 9 as training set. This procedure is repeated until all the folds have been used as the testing data. In each iteration, a model is fit on the training set, and evaluated on the test set.

2.2.1 Decision Tree Classifier Model (CART)

A decision tree operates by concluding the value of a dependent attribute given the values of the independent features [10]. The classification and regression tree, a type of decision was adopted implemented in this projected. It works by splitting the dataset into several segments through posing series of questions about the features. The was employed in this case because of its successful use in medical research as a powerful statistical tool for classification, interpretation, and data manipulation (Song & Lu, 2015).

2.2.2 Logistic regression (LR)

LR has been one of the widely used machine learning models for analyzing multivariate regression problems in the health fields [11]. It is used for predicting the outcome of a dependent variable with a continuous independent variable. It models binary dependent variables and it fits an equation to the data.

2.2.3 Gaussian Naïve Bayes Model (GNB)

The Naïve Bayes model is a classification model which is based on the Bayes theorem. It is a simple probabilistic model which is premised on the assumption that all the features are linearly independent of each other, for a given categorical variable [12]. The Gaussian Naïve Bayes (GNB) classifier is known for the prediction or for recognizing pattern in a data. This model operates by taking each data point, and subsequently assign it the nearest class to it. The GNB instead of using the Euclidean distance from the class means, it considers also the compared class variance [13].

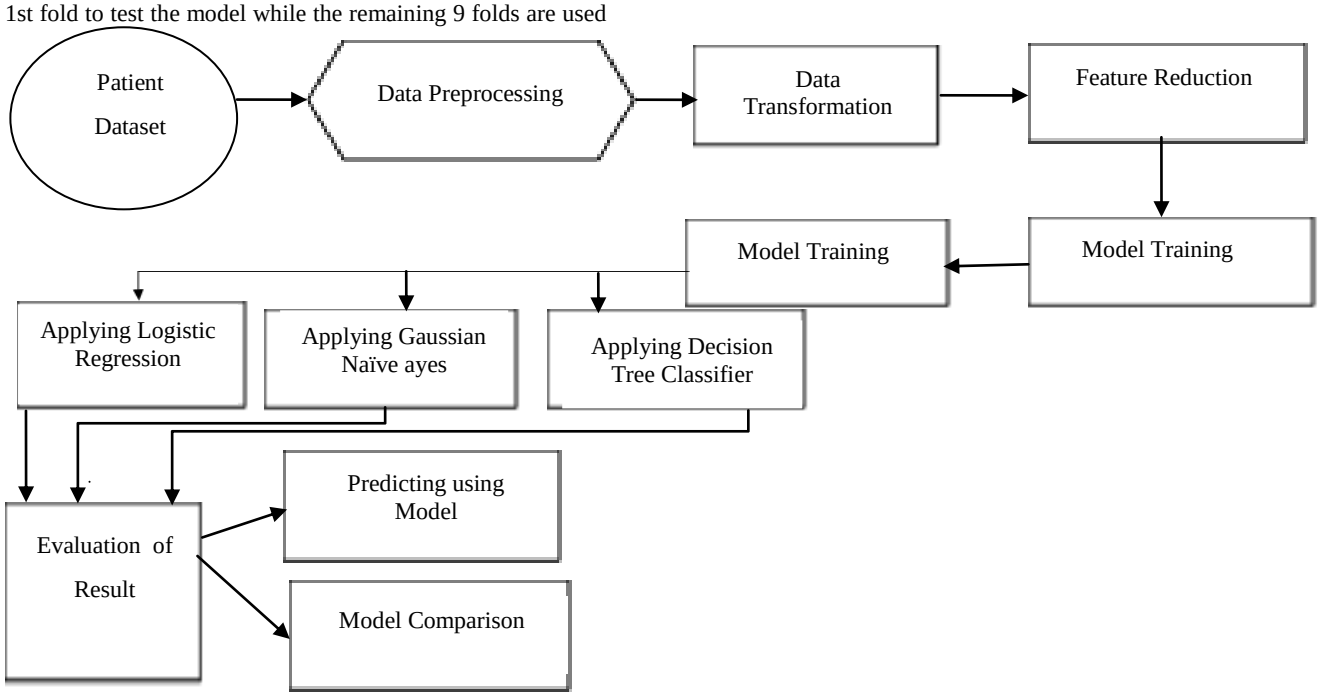

Fig.1 A high level block diagram summarizing the methodology adopted in this paper

3. RESULTS

3.1 Evaluation of Classification Algorithms

The three classification models were analyzed by evaluating precision, recall, f1, and accuracy scores. The performance of each classification model on the test data was visualized using confusion matrix. Area under the curve (AUC) receiver operating characteristic (ROC) curves were used to visualize the performance output of each of the models. It plots the true positive rates (sensitivity) against false positive rate (specificity). AUC is between 0 and 1. The greater the AUC value, the better the classification model. The reliability of the proposed models was tested by dividing data with the 10-fold cross validation method. **Figure 11** provides a comparison of the accuracy scores of the three algorithms (LR, GNB, CART).

$$Accuracy = \frac{True\ Positives + False\ Negatives}{Total\ number\ of\ samples}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

3.1.1 Decision Tree Classifier model

The confusion matrix for the decision tree classifier is presented in the figure below in table 1. The accuracy score obtained for this model was 79.31%. Table 2 presents evaluation metrics with precision, (ability of the model not to label patients as having heart diseases that do not have heart disease, recall, and f1-score. The AUC value obtained was 0.81,with the corresponding ROC graph displayed below in figure 8.

Table 1. Confusion matrix of decision Tree Classifier model

	Predicted (absence)	Predicted (presence)
Actual (absence)	20	7
Actual (presence)	5	26

Table 2. Classification report of Decision Tree Classifier model

	precision	recall	f1-score	support
0.0	0.80	0.74	0.77	27
1.0	0.79	0.84	0.81	31

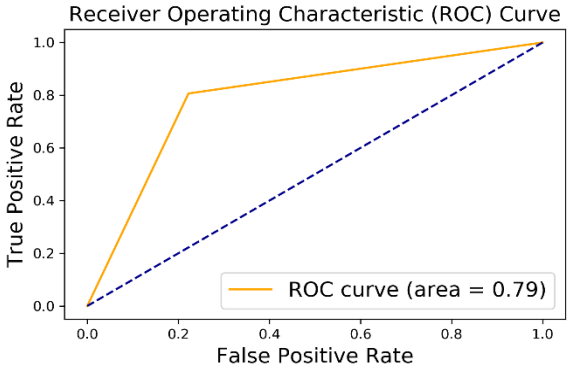

Fig 3: ROC curve showing the performance output of Decision Tree Classifier model

3.1.2 Gaussian Naïve Bayes Model

Displayed below (table 3) is the result of the confusion matrix for describing the performance of the GNB classifier, indicating how the testing dataset was predicted. An accuracy score, which indicates how correctly the model was able to predict the presence or absence of heart disease. The model was able to correctly predict 76% of the test dataset as either the presence or absence of heart disease in patients. Other performance metric such as precision, recall and f1-score are displayed in table 4. The recall indicates how many of either absence or presence of the disease the model was able to capture through classifying it as the absence of presence of the heart disease, respectively. Moreover, the AUC for this model was 0.87, and the ROC curve is represented in fig. 3.

Table 3. Confusion matrix of Gaussian Naive Bayes model

	Predicted (absence)	Predicted (presence)
Actual (absence)	21	6
Actual (presence)	4	27

Table 4. Classification report of Gaussian Naive Bayes model

	precision	recall	f1-score	support
0.0	0.81	0.73	0.77	30
1.0	0.76	0.84	0.80	31

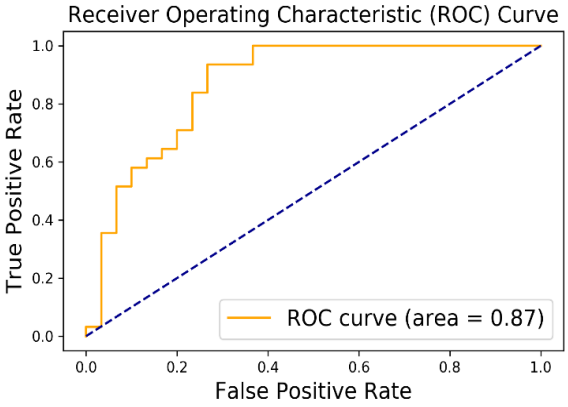

Fig 4: ROC curve showing the performance output of Gaussian Naive Bayes model

3.1.3 Logistic Regression

The confusion matrix associated with LR model is represented below indicating how many of the test samples that were predicted accurately as the presence or absence of heart disease (table 5). The summary of the precision recall and f1 scores are represented in table 6. The accuracy score for this model was 82.75%. The AUC for this model was 0.86, and the output is presented graphically in figure 4.

Table 5. Confusion matrix of Logistic Regression model

	Predicted (absence)	Predicted (presence)
Actual (absence	21	6
Actual (presence)	4	27

Table 6. Classification report of Logistic Regression model

	precision	recall	f1-score	support
0.0	0.84	0.78	0.81	27
1.0	0.82	0.87	0.84	31

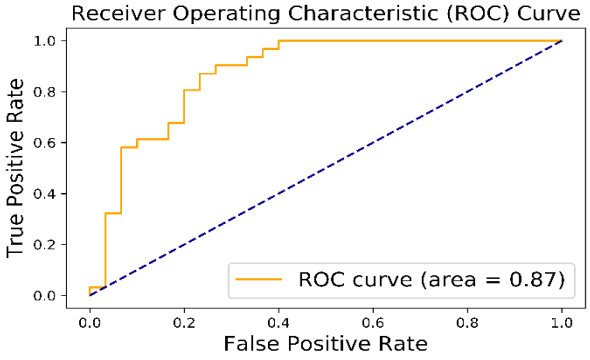

Fig 4. ROC curve showing the performance output of Logistic Regression model

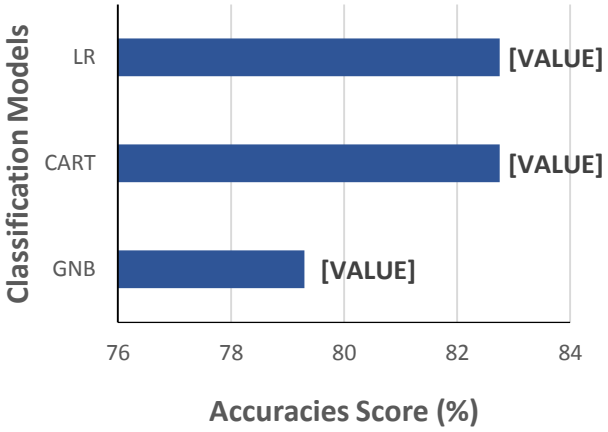

Fig 5. A diagrammatic comparison of the predicting accuracies (%) of the three models.

4. DISCUSSION

The purpose of this research was to compare the accuracies of the three algorithms (CART, LR and GNB) in predicting heart diseases in patients. Surprisingly, the highest predicting accuracy score was obtained with the LR and GNB. They both had predicting score of 82.75%, precision, recall,f1 scores. However, the AUCROC value for the GNB was higher than LR model. . Naïve Bayes algorithms are documented to be effective in practical medical diagnosis [14]. The competitive performance of GNB in classification could be attributed to the dependence distribution [15]. The CART model scored the lowest predicting accuracy, among the three models. The least accuracy score obtained with CART algorithm could be due to the relatively smaller sample size of the dataset [16].

5. CONCLUSION

Heart disease detection at the early stages with few clinical tests to diagnosing it is crucial in preventing the many deaths associated with it. The burgeoning influence of data mining techniques and machine learning in the medical field in detecting subtle patterns in large dataset make their applicability in heart disease diagnostics relevant. The performance metric used in evaluating the three models puts the GNB model ahead of the three classifiers. The greater AUCROC value (0.87) from GNB model makes it a better choice than LR (0.86). Future research could focus on including different models such as random forest, K Neighbors Classifier, support vector machine, etc.

6. REFERENCES

[1] H. K. Weir et al., "Heart Disease and Cancer Deaths - Trends and Projections in the United States, 1969-2020," Prev. Chronic Dis., vol. 13, pp. E157–E157, Nov. 2016.

[2] C. S. Dangare and M. E. Cse, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," Int. ournal Comput. Appl., vol. 47, no. 10, pp. 44–48, 2012.

[3] J. Patel, S. Tejalupadhyay, and S. Patel, Heart Disease prediction using Machine learning and Data Mining Technique. 2016.

[4] S. Sakr et al., "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project,"

BMC Med. Inform. Decis. Mak., vol. 17, no. 1, p. 174, Dec. 2017.

[5] S. Sakr et al., "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project.," PLoS One, vol. 13, no. 4, p. e0195344, 2018.

[6] M. Shouman, T. Turner, and R. Stocker, Using decision tree for diagnosing heart disease patients, vol. 121. 2011.

[7] A. H. Babar, "Comparative Analysis of Classification Models for Healthcare Data Analysis," vol. 07, no. 04, pp. 170–175, 2018.

[8] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining Techniques," Caribb. J. Sci. Technol., vol. Vol.1, pp. 208–217, Dec. 2013.

[9] H. Sharma, "Prediction of Heart Disease using Machine Learning Algorithms : A Survey," Int. J. Recent Innov. Trends Comput. Commun., vol. 5, no. 8, pp. 99–104, 2017.

[10] N. Bhargava and G. Sharma, "Decision Tree Analysis on J48 Algorithm for Data Mining," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 6, pp. 1114–1119, 2013.

[11] E. W. Steyerberg, M. J. Eijkemans, F. E. J. Harrell, and J. D. Habbema, "Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets.," Med. Decis. Making, vol. 21, no. 1, pp. 45–56, 2001.

[12] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," J. Inf. Sci., vol. 44, no. 1, pp. 48–59, Nov. 2016.

[13] R. D. S. Raizada and Y.-S. Lee, "Smoothness without Smoothing: Why Gaussian Naive Bayes Is Not Naive for Multi-Subject Searchlight Studies," PLoS One, vol. 8, no. 7, p. e69566, Jul. 2013.

[14] I. Rish, "An Empirical Study of the Naïve Bayes Classifier," IJCAI 2001 Work Empir Methods Artif Intell, vol. 3, Jan. 2001.

[15] H. Zhang, The Optimality of Naive Bayes, vol. 2. 2004.

[16] R. Nichenametla, T. Maneesha, S. Hafeez, and H. Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," Int. J. Eng. Technol., vol. 7, pp. 363–366, May 2018.