Improving performance of Heart Disease prediction through feature dependency extraction

USMAN ABDULLAHI MUSA

Computer Science Department, Federal University Dutse

Usman.m@fud.edu.ng

ABSTRACT

Cardiovascular disease is a significant public health concern responsible for many deaths annually. It also causes a significant amount of morbidity and impairment to humans. An increase in health care data through the use of electronic health record (EHR) systems makes it possible to perform analysis on the data and forecasting diverse scenarios for numerous fields. The need to make accurate predictions of heart disease through the use of machine learning algorithms is as a result of many factors the human mind cannot process. Numerous machine learning algorithms such as Random Forest, Logistic Regression, ANN, K-Nearest Neighbor, SVM, etc. have been applied on Cleveland heart datasets however, not very much was done on modeling with a Bayesian Network (BN). The widely used 14 features of the Cleveland heart data collected from the UCI repository used and modeled using Bayesian Network modeling in this study. Experimental results prove that the use of feature reduction techniques does effectively improve the prediction performance of the classifier. The study aim is to assess how feature reduction could increase performance and extract those feature dependency that affects the performance of the classifier. The proposed method achieved an accuracy of 88.53%. Based on the results obtained, we observed feature reduction on Cleveland dataset could increase performance of Bayesian network

**Keywords**—*Machine Learning, Bayesian Network (BN), Naïve Bayes Logistic Regression, KNN, Heart Disease, Prediction*

INTRODUCTION

Cardiovascular diseases are conditions that influence the constructions or capacity of your heart, for example, Abnormal heart rhythms/arrhythmias, Aorta infection, and Marfan disorder, Congenital coronary illness, Heart assault, Heart disappointment, Heart muscle sickness (cardiomyopathy) and Stroke, and so on (Steinbaum 2019). These diseases share common risk factors namely; age, unhealthy diet sexual orientation, hypertension, diabetes mellitus, tobacco smoking, processed meat utilization, unnecessary liquor consumption, sugar consumption, family ancestry, weight, absence of exercise, psychosocial factors, and air contamination (WHO 2019). Over 36 million people die annually from Noncommunicable Diseases accounting for about 63% of all global deaths. Low- and middle-income countries are known to bear 86% of the number of these premature deaths giving rise to estimated cumulative economic losses of US$7 trillion over the next 15 years(WHO 2019). According to the Federal Ministry of Health (FMoH), “cardiovascular disease is a major public health concern, accounting for 11% of the over 2 million NCD deaths in Nigeria each year.” People with Cardiovascular Disease are often unaware of their condition until a disaster such as a stroke, heart attack, or death occurs.” (WHO 2019).

The increase in the amount of health data gathered through the electronic health record (EHR) systems makes the use of strong analysis tools necessary. Numerous machine learning algorithms such as Random Forest, Logistic Regression, ANN, K-Nearest Neighbor, SVM, etc. have been applied on Cleveland heart datasets however, not very much was done on modeling with a Bayesian Network (BN). The need of making accurate predictions of heart disease made the use of machine learning algorithms to point out predictions based on many factors. Many factors or input features even to machine learning often make a predictive modeling task more challenging to model. Feature engineering is important in increasing the efficiency and correctness of prediction on machine learning models. It entails translating raw data into characteristics that better reflect the underlying problem to prediction models, resulting in enhanced model accuracy on unseen data.

Bayesian networks (BNs) have received increasing research attention as it possesses potential significant benefit to the healthcare system. A Bayesian network represents the causal probabilistic relationship among a set of random variables, their conditional dependencies, and it provides a compact representation of a joint probability distribution. The proposed method in this study is the use of wrapper feature selection technique as a dimension reduction technique for extracting important features and the use of Naïve Bayes, Bayesian Network, KNN, and Logistic Regression to make predictions. The performance of all the models is also measured to make a comparison.

MATERIALS AND METHODS

The open-source program WEKA was used in the data mining procedure. WEKA is a tool for knowledge analysis that has multiple machine learning algorithms for data analysis(Srivastava 2014). The software is necessary to develop models in different sectors, such as bioinformatics, education, and medicine. To accurately, easily, and effectively predict cardiovascular disease, the flowchart of how the proposed system works in shown in Fig 1. The first step involves data collection followed by data preprocessing. After prepossessing the data, the full data set is put trained on the selected models and then evaluated. The preprocessed data then put into a feature reduction technique and feature selection is done. The reduced dataset is then trained on the models and evaluated. The steps taken in the flowchart are listed below:

1. The dataset to be used
2. Data Retrieval
3. Data preprocessing
4. Feature Selection
5. the model to be used for the study.
6. Performance Metrics

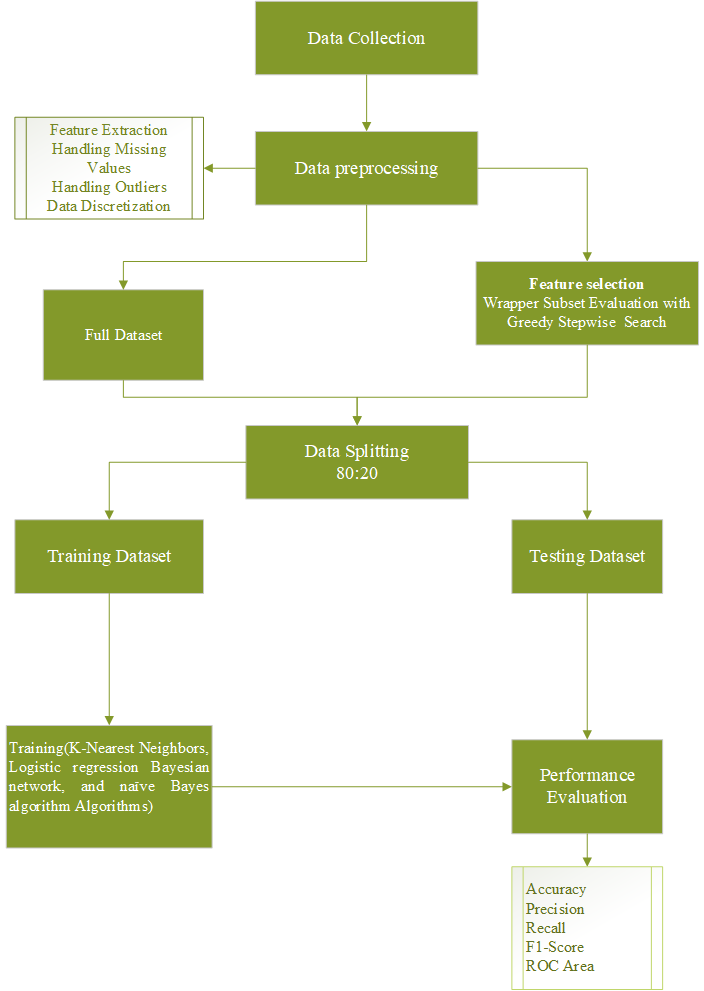


Figure 1: Flow Chart of the Proposed System

1. *Dataset*

This research is proposed to use the heart disease dataset from the UCI machine learning repository called Cleveland Heart Disease Data set. Cleveland Heart Disease Dataset is a publicly available supervised dataset provided by the Cleveland Clinic Foundation that was used for the ML model. This data set contains 14 total attributes of patient medical information for 303 patients. The figure below shows the chosen attributes and their information.

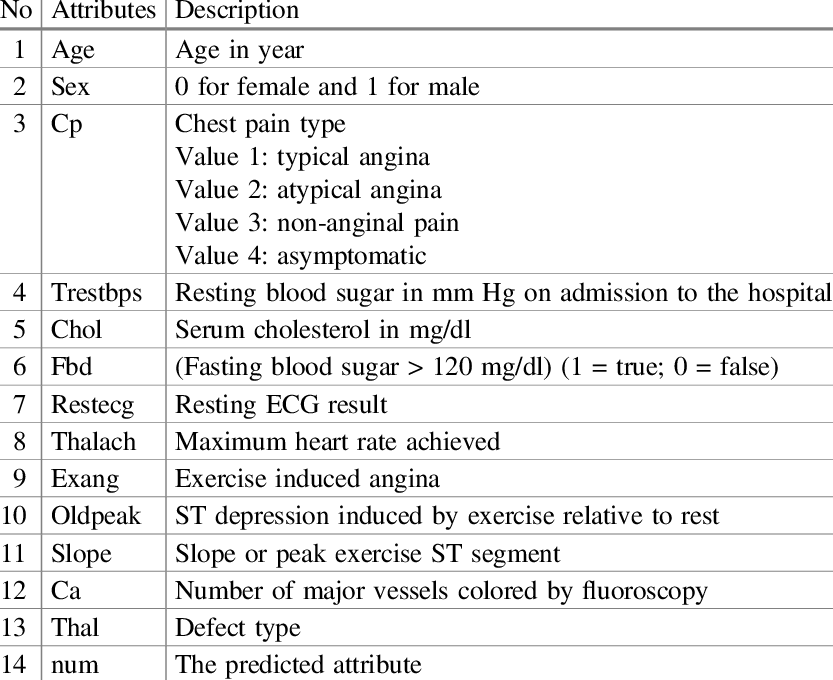


Figure 2: List of Full Cleveland Dataset

ii. *Data Retrieval*

The Dataset can be retrieved from University of California, Irvine dataset archive or from Kaggle dataset. Our dataset is to be downloaded from the UCI machine repository.

1. *Data Preprocessing*

Data preprocessing is also known as cleaning data. It is one of the most important steps to achieve the best from the dataset. This is a technique that removes data inconsistencies such as missing numbers, out-of-range values, unformatted data, and noise. Our preprocessing would involve data retrieval, handling missing values, and data discretization

Handling Missing Values

Missing data values is a common problem faced by analysts. This occurs due to different reasons such as incomplete extraction, corrupt data, failure to load the information, etc. This is a great challenge that must be fixed because good models are generated when you make the right decisions on how to fix them. Ways of handling missing data includes:

a. Deleting Rows

b. Replacing with mean/median/mode

c. Assigning a unique category

d. Predicting the missing values

e. Using algorithms that supports missing values

We would adopt the method of handling missing values that prove best.

Data Discretization

The technique of transforming continuous data attribute values into a finite set of intervals with minimal information loss is known as data discretization. Discretization can help improve significantly the classification performance of some as algorithms like Naïve Bayes that are sensitive to the dimensionality of the data(Lustgarten et al. 2008).

1. *Feature Selection*

To select the best features for our model. A wrapper method of attribute selection on WEKA was proposed. The wrapper method is a feature selection method that fits a model and selects the combination of attributes that will perform best.

1. *Proposed Model*

The proposed models comprise of Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. This is because the models proved to have high performance on predicting heart disease from previous studies.

*Naive Bayes*

Naïve Bayes or stupid Bayes is used to handle binary (two-class) and multiclass classification challenges, It has its name because the probabilities for each hypothesis are simplified to make its calculation tractable. (Jason 2014). In simple terms, a Naive Bayes classifier assumes that the presence of one attribute or feature in a class is unrelated to the presence of any other feature, i.e., predictor independence.

P(A|B) =

Where P(A) is the prior distribution of parameter A; P(A|B) is the posterior distribution, the probability of A given new data B; and P(B|A) the likelihood function, the probability of B given existing data.

*Bayesian Network*

A Bayesian network B = < N, A, θ > is a directed acyclic graph (DAG) <N, A> with a conditional probability distribution (CP) for each node, collectively represented by θ. Each arc a ∈ A between nodes represents a probabilistic dependency, and each node n ∈ N represents a domain variable. In general, a BN can be used to compute the conditional probability of one node, given values assigned to the other nodes; hence, a Bayesian Network can be used as a classifier to calculate the posterior probability distribution of a classification node given the values of other characteristics. (Cheng et al. 2002). A Bayesian network, for example, could reflect the probability correlations between diseases and symptoms. Given a set of symptoms, the network may be used to calculate the likelihood of the presence of certain diseases.

*Logistic regression*

The logistic function, also known as the sigmoid function, was created by statisticians to characterize the properties of population increase in ecology, such as how it rises swiftly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1, but never exactly between those two points.

Where e is the base of the natural logarithms (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform.

Logistic regression has two alternative outcomes of a goal variable. This means that the input and output have a linear relationship and calculates the likelihood of the goal variable of the data. LR still considers the dependent variable to be a bi-categorical variable. It is mostly used to forecast and calculate the likelihood of success. Molding the equation into the form of needed data entry is also part of LR. A basic equation is used here:

Y = β0 + β1X1 + β2X2 + …… βnXn

The regression coefficients are estimated using the Maximum Likelihood Ratio (MLR). It aids in the calculation of statistical significance for dependent variables using independent variables. MLR tests and assesses the part of the independent variables.

*K-Nearest Neighbour*

KNN makes predictions using the training dataset directly. For each new data point, predictions are formed by exploring the whole training set for the k most similar examples (neighbors) and summing the output variable for those k instances. The most similar of the k instances in the training dataset to a new input is determined using a distance metric. The most frequent distance measure for real-valued input variables is Euclidean distance. The square root of the total of the squared discrepancies between two points a and b across all input qualities *i* is used to calculate Euclidean distance.

KNN extracts data points from a dataset and calculates the closest output. Because there are various features in the heart disease dataset, this technique works well with pattern recognition. Along with the majority of KNN, it extracts logic and knowledge using the Euclidean distance Samples function d(Xi, Xj). Mathematically;

1. PERFORMANCE METRICS

Performance metrics are used to evaluate how different algorithms perform based on various criteria such as accuracy, precision, recall, etc. They are discussed below.

*Accuracy*

Accuracy is the ratio of the number of correctly classified instances to all the cases. It is the sum of TP and TN divided by the total number of instances. Accuracy = (TP + TN) / (TP+TN+FP+FN)

*Precision*

Precision is the proportion of true positive instances that are classified as positive. It shows how near the projected values are to each other. Precision = TP/(TP+FP)

*Recall/ Sensitivity*

The recall is the proportion of positive instances that are correctly classified as positive. Recall is known as sensitivity. Sensitivity = TP / (TP +FN)

*F1 Score*

F1 score combines both precision and recall and finds a balance between both. In other words, it computes the harmonic mean of precision and recall. F-measure = (2\*Precision + Recall) / (precision + Recall)

*MCC*

MCC known as Mathew correlation coefficient. Matthew’s correlation coefficient is a contingency matrix method of calculating the Pearson product-moment correlation coefficient between actual and predicted values. It ranges in the interval [ −1, +1], with extreme values –1 and +1 reached in case of perfect misclassification and perfect classification, respectively.

*AUROC curve*

It's a graph depicting the ratio of false positives to real positives. The area assesses discrimination, or the classifier's ability to accurately classify the test data.

*Kappa Statistics*

The kappa measure of agreement is the ratio K = P(A) - P(E) 1 - P(E)

Where P(A) denotes the percentage of times the k raters agree, i.e., the percentage agreement between the classifier and the ground truth, and P(E) is the proportion of times the k raters are expected to agree by chance alone i.e., the chance agreement. K=1 indicates perfect agreement and K=0 indicates chance agreement. The value greater than 0 means classifier is doing better. The classifier's result improves as the kappa statistic value rises.

RESULTS AND DISCUSSION

In total for machine learning classifiers, 303 records were used each with 14 total attributes. Three different attribute feature subset selection where made. Each feature selection technique performace is evaluated. This list of the feature selection with selected attributes are shown below:

1. **Wrapper Subset Evaluation:** Age, sex, cp, restecg, exang, oldpeak, slope, ca, thal
2. **Correlation-based Feature Subset Selection:** sex, cp, restecg, thalach, exang, oldpeak, slope, ca, thal
3. **Consistency Subset Evaluation:** age, sex, cp, fbs, restecg, thalach, exang, oldpeak, slope, ca,thal

The result of the three feature selection subset using Bayesian network is shown in the table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Precision** | **Recall** | **F-Measure** | **MCC** | **ROC Area** | **Accuracy** |
| **Dataset Before reduction** | 0.869 | 0.869 | 0.869 | 0.737 | 0.927 | 86.8852 |
| **Wrappers for feature subset selection** | 0.888 | 0.885 | 0.885 | 0.773 | 0.920 | 88.5246 |
| **Consistency Subset Evaluation** | 0.869 | 0.869 | 0.869 | 0.737 | 0.927 | 86.8852 |
| **Correlation-based Feature Subset Selection** | 0.838 | 0.836 | 0.836 | 0.674 | 0.919 | 83.6066 |

Out of the 3 different feature selection techniques, Wrapper feature subset proves to have better performance with the highest accuracy, Precision, Recall F-Measure and MCC score. The wrapper feature subset had selected 9 out of the 14 features as it shown to played a part in determining the best feature subset in heart disease diagnosis. The proportions of the feature set is also evaluated. The evaluation involved different classification methods, i.e., Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. In this study, WEKA was used to classify the Cleveland dataset. This is done using different proportions of the dataset for training and testing and the accuracy measure is taken as shown in Table 1. In Table 1, Splitting the training and testing data into 80:20 ration has the highest accuracy.

Table 1. Test proportion to determine the best ration for highest model accuracy (rounded to 4 digits)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ALGORITHMS** | **TRAINING/TESTING ACCURACY RATIO** | | | | | | | | |
| 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| Naïve Bayes | 82.4176 | 83.8843 | 85.3774 | 84.0659 | 86.0927 | 87.6033 | 86.8132 | 86.8852 | 86.6667 |
| Bayesian Network | 81.685 | 83.4711 | 86.3208 | 85.1648 | 87.4172 | 87.6033 | 86.8132 | 88.5246 | 86.6667 |
| KNN | 79.8535 | 79.3388 | 79.717 | 80.2198 | 82.1192 | 80.1653 | 82.4176 | 80.3279 | 86.6667 |
| Logistic Regression | 68.8645 | 65.2893 | 75.000 | 82.4176 | 86.0927 | 83.4711 | 81.3187 | 85.2459 | 83.3333 |

In determining which ML algorithm to use, accuracy scores were produced for each of the common machine learning algorithms. Figure 1 shows the accuracy scores of the K-Nearest Neighbors, Logistic regression Bayesian network, and Naïve Bayes algorithm Algorithms, and proves how the Bayesian Network had the highest accuracy score on 80:20 training/testing split.

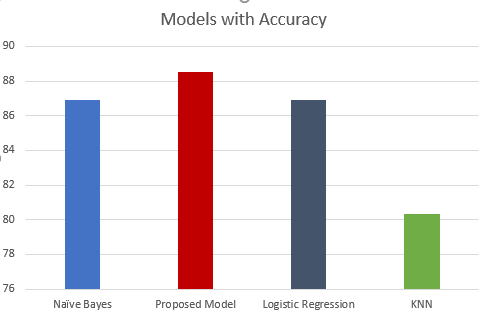


Figure 3: Accuracy Comparison of the models

ROC graphs were also established for the Bayesian Network (Fig 4.), Naïve Bayes (Fig 5), Logistic Regression (Fig 6), and KNN (Fig 7). Bayesian network and Naïve Bayes had the highest a score of approximately 92%. Logistic regression had a score of approximately 91% with KNN having a Score of 87.3%

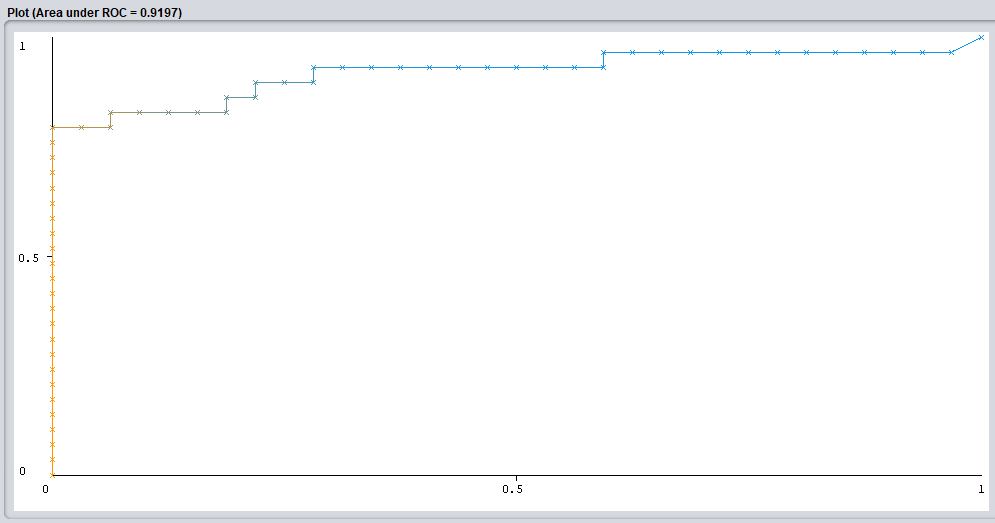
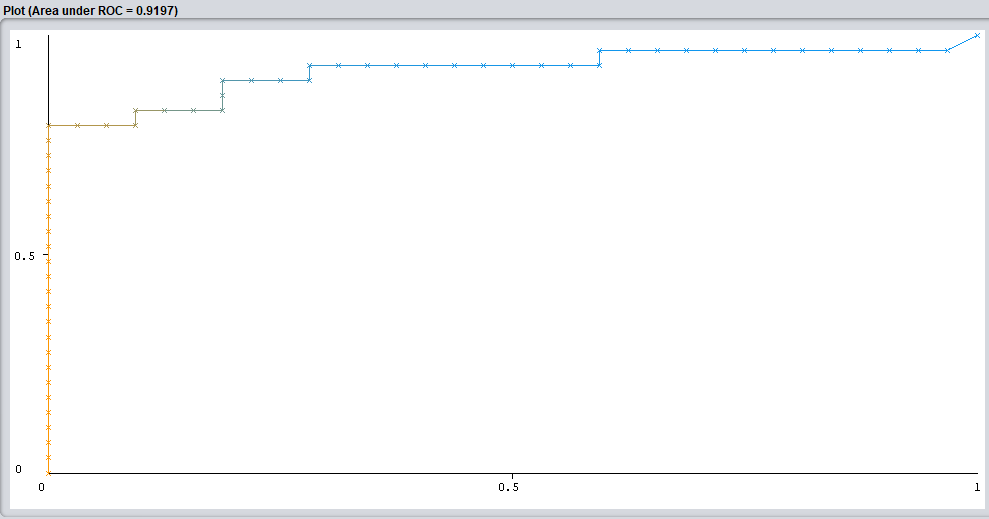
 

Figure 4: Proposed Model Figure 5: Naive Bayes

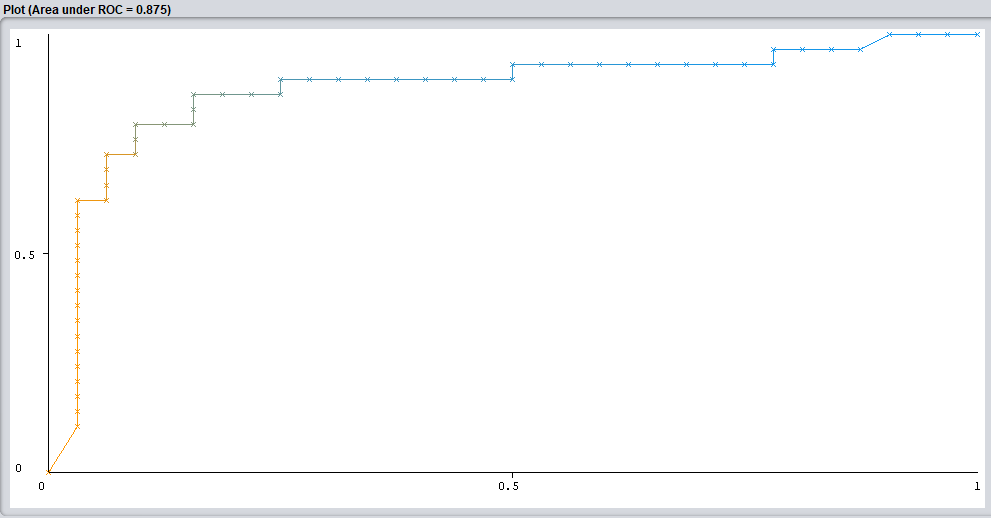
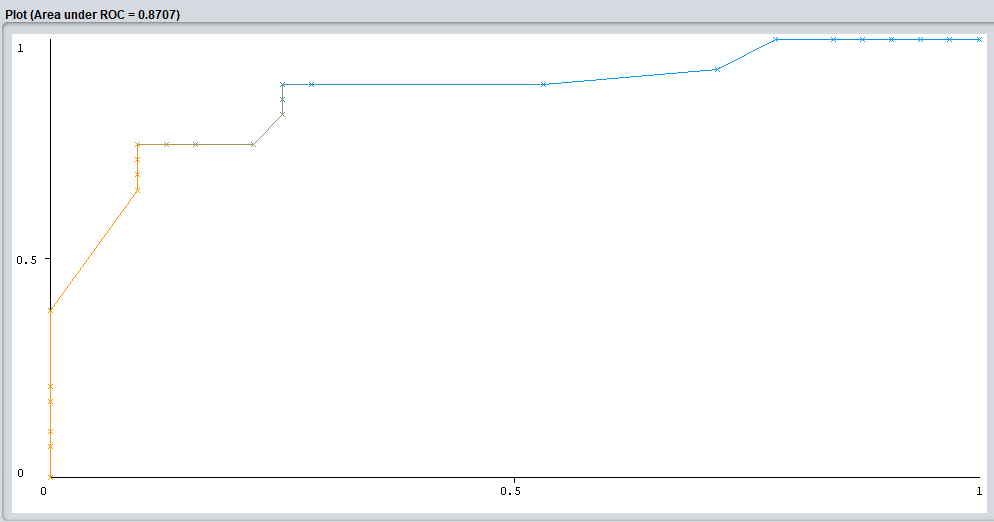
 

Figure 6: Logistic Regression Figure 7: KNN

Comparison of the models in relation to Matthews Correlation Coefficient (MCC) chart is shown in Figure 8 below. The Chart shows Bayesian network having the highest score of 77% followed by Naïve Bayes and Logistic regression with similar score of about 74%.KNN has the least MCC score of about 61%.

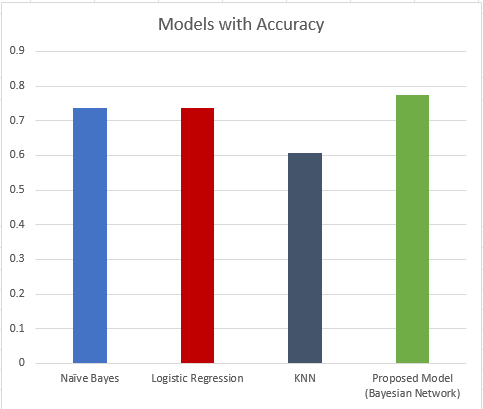


Figure 8: MCC chart of the Models

Table 2: Metrices Comparison between Naive Bayes, Bayesian Network, Logistic Regression and KNN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F-Measure** | **MCC** | **ROC Area** | **Kappa statistic** |
| **Naïve Bayes** | 86.8852 | 0.870 | 0.869 | 0.868 | 0.738 | 0.920 | 0.7362 |
| **Logistic Regression** | 86.8852 | 0.870 | 0.869 | 0.868 | 0.738 | 0.908 | 0.7362 |
| **KNN** | 80.3279 | 0.804 | 0.803 | 0.803 | 0.606 | 0.873 | 0.6043 |
| **Proposed Model (Bayesian Network)** | 88.5246 | 0.888 | 0.885 | 0.885 | 0.773 | 0.920 | 0.7688 |

Looking at the performance of the various models, each model had proved to have a good predictive power to heart disease However, looking at the MCC score and accuracy score Bayesian Network, despite having a similar AUC score with Naive Bayes was the overall winner. Bayesian Network proved to be the better overall model as seen by its much higher performance when looking at accuracy score and Matthews Correlation Coefficient. Thus, it can be deemed that it was the best algorithm out of the 4 tested.

Table 3: Comparison Various Approaches with our Proposed Approach

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S/N | Author and Year | Method | Dataset | Metrics |
|  | Mistura Muibideen & Rajesh Prasad, 2020 | Bayesian Network | Cleveland dataset: 14 Attributes | Accuracy: 85%  Precision: 86%  Recall: 85%  F1- Score: 85% |
|  | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton **-**2020 | 2-layer CNN | NHANES dataset:  7 attributes | Accuracy: 81.78%  Recall: 77.3%  Specificity: 81.8 %  AUC: 76.78 % |
|  | Sahithi Ankireddy -2020 | Deep Neural  Network (DNN) | Cleveland dataset: 14 Attributes | Accuracy: 85.60% |
|  | Ekta Maini, and Bondu Venkateswarlu **-**2021(Maini and Venkateswarlu 2021) | Ensembling techniques (Naïve Bayes, SVM, Logistic Regression and and Multilayer Perceptron) | Cleveland dataset: 14 Attributes | Accuracy: 87.5% |
|  | **Our proposed approach** | Bayesian Network with Wrapper subset evaluation (For feature selection) | Cleveland dataset: 8 Attributes. Namely: age, sex, cp, exang, oldpeak, slope, ca, thal | Accuracy: **88.53%**  Precision: **88.8** %  Recall: **88.5**%  F1- Score: **88.5** %  ROC Area**: 92.0%** |

Conclusion

It is worth researching much of what is required to forecast and diagnose any disease using machine learning effectively. Heart disease is considered one of the major threats to life and now it is a critical challenge to predict heart disease at an early stage in the area of clinical data analysis in order to minimize the death rate. Our method reduces the dimensionality of the dataset using WEKA wrapper method of data selection to select the best subset on Cleveland dataset features for better accuracy and efficiency predicting heart disease. The selected features are 8 in numbers and they include: age, sex, cp, exang, oldpeak, slope, ca, thal . The proposed method in the study has been evaluated with various metrics, and its performance results are compared with explores different machine learning algorithms.

A very detailed, useful, and highly preferable Machine Learning based model in this paper that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification window. The paper used Naïve Bayes, Bayesian Network, KNN, and Logistic Regression on the reduced features. The same features are used to both train and test the dataset. The outcome reveals that these data mining techniques can predict heart disease early with an accuracy of approximately 89%.

**Acknowledgements**:

The author(s) wish to acknowledge the kind support of Dr. Muhammad Sirajo Aliyu for providing the useful guidance in the research work.

References

Cheng, Jie, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. 2002. “Learning Bayesian Networks from Data: An Information-Theory Based Approach.” *Artificial Intelligence* 137 (1–2): 43–90. https://doi.org/10.1016/S0004-3702(02)00191-1.

Jason, Brownlee. 2014. “Naive Bayes for Machine Learning.” *Machine Learning Algorithms*. https://machinelearningmastery.com/naive-bayes-for-machine-learning/.

Lustgarten, Jonathan L, Vanathi Gopalakrishnan, Himanshu Grover, and Shyam Visweswaran. 2008. “Improving Classification Performance with Discretization on Biomedical Datasets.”

Maini, Ekta, and Bondu Venkateswarlu. 2021. “Improving the Performance of Heart Disease Prediction System Using Ensembling Techniques.” *AIP Conference Proceedings* 2316 (February). https://doi.org/10.1063/5.0036478.

Muibideen, Mistura, and Rajesh Prasad. 2020. “A Fast Algorithm for Heart Disease Prediction Using Bayesian Network Model,” 1–11. http://arxiv.org/abs/2012.09429.

Sai Shekhar, M., Y. Mani Chand, and L. Mary Gladence. 2021. “Heart Disease Prediction Using Machine Learning.” *Lecture Notes in Electrical Engineering* 708 (Inccst 20): 603–9. https://doi.org/10.1007/978-981-15-8685-9\_63.

Srivastava, Shweta. 2014. “Weka: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule Mining.” *International Journal of Computer Applications* 88 (10): 26–29. https://doi.org/10.5120/15389-3809.

Steinbaum, Suzanne R. 2019. “Cardiovascular (Heart) Diseases: Types and Treatments.” 2019. https://www.webmd.com/heart-disease/guide/diseases-cardiovascular.

Webb, Geoffrey I., Eamonn Keogh, Risto Miikkulainen, Risto Miikkulainen, and Michele Sebag. 2011. “Naïve Bayes.” *Encyclopedia of Machine Learning*, 713–14. https://doi.org/10.1007/978-0-387-30164-8\_576.

WHO. 2019. “WHO and Nigerian Government Move to Curb Cardiovascular Diseases | WHO | Regional Office for Africa.” 2019. https://www.afro.who.int/news/who-and-nigerian-government-move-curb-cardiovascular-diseases.