Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach.

Ambily Merlin Kuruvilla¹ and Dr.N.VBalaji²
¹ Research Scholar, Department of CS,CA and IT, Karpagam Academy of Higher Education
² Dean, Faculty of Arts, Science and Humanities, Karpagam Academy of Higher Education
E-mail: E-mail:ambilykuruvilla@gmail.com

Abstract. Cardiac disease prediction helps physicians to make accurate recommendations on the treatment of the patients. The use of machine learning (ML) is one of the solutions for recognizing heart disease-related symptoms. The goal of this study is to suggest a methodology for identifying the most relevant features of cardiac disease characteristics by applying a feature selection technique. The data set used in this study was Framingham heart disease dataset (FHS). It was collected from KAGGLE Machine Learning repository. There are 16 attributes and a mark in the dataset that has been validated by four ML classifiers. There are two feature selection methods, Correlation Based Feature selection (CBFS) and Principle Component Analysis (PCA) was used for the comparison in the study. By using CBFS Method five highly correlated features are selected for the study, and by using PCA thirteen features are selected. The experimental result shows that Correlation Based Feature Selection with Multilayer perceptron (CBFS with MLP) obtained the highest accuracy for this dataset.

1. Introduction
The research concentrates on the two feature selection methods for data reduction before building the predictive models by classification algorithms. These reduced features are then passed into the classification algorithms to design the models for the heart disease prediction. These models are used for the comparison of accuracy of the classifier. Principle Component Analysis and Correlation Based feature selection methods are used for finding out the reduced features. The selected features are inputted to four different classifiers such as Naïve Bayes, ADABOOST, MLP and SMO. The accuracy of each model is compared with the other.

2. Background Study
Devansh Shah studied various attributes related to heart disease[1]. The study was conducted with Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithms[1]. The experimental result proves that K-nearest neighbor algorithm exhibits the highest accuracy.
Hamidreza Ashrafi Esfahani[2] formulated a model to predict cardiovascular disease. The model includes decision trees, Neural Networks, Rough set, Naïve Bayes and SVM for implementation. On comparing the results achieved, it was revealed that the hybrid model of Rough Set, Naïve Bayes and Neural Network obtained the highest accuracy. An ensemble strategy was implemented that allowed for the output to be combined that would result in
better accuracy. The performance of the classifiers was compared with the parameters such as its precision, sensitivity, accuracy and F-Measure.

3. Proposed Methodology
Framingham Heart Study (FHS) from Kaggle Machine Learning repository is used for the study. Two feature selection methods along with four classification algorithms are used for the study. CBFS and PCA are the methods used for the dimensionality reduction. MLP, Navie Bayes, Sequential Minimum Optimiser (SMO) and ADABOOST algorithms are used for classification. The reduced feature set from both the feature selection methods are inputted to different classifiers. Eight different Machine Learning Models were created for Heart Disease Prediction. Accuracy of these Models are compared with each other.

4. Description of the Data Set
The Framingham Heart Study (FHS) dataset was collected from Kaggle. The dataset consists of 4241 records. It contain sixteen features including AGE, PREVALENT HYP, SYSBP, DIABP, GLUCOSE, SEX, EDUCATION, CURRENT SMOKER, CIGSPERDAY, BPMEDS, PREVALENT STROKE, BMI, HEART RATE, DIABETES, TOTCHOL and PREDICTOR VARIABLE.

5. Classification Algorithms
In Machine Learning various forms of classification techniques are available. Classification techniques used for this study was described below.

6. Multilayer Perceptron (MLP)
MLP is a subset of Artificial Neural Network. MLP comprises one or more than one hidden layers aside from one input and one output plate. The Perceptron is made of an input layer and a totally linked output layer. MLPs have the same levels of input and output, but could have several levels concealed within them.

![Figure 1. Different layers in MLP.](image)

7. Adaboost
In machine learning, AdaBoost( Adaptive Boosting) is a supervised learning algorithm. It is used for combining several weak classifiers together to generate a strong classifier.

8. Naive Bayes
Naive Bayes is a Machine Learning algorithm based on Probability theory in statistics. The term naive suggests that the elements that go through the software are autonomous of each other, That is, the value of one characteristic, does not explicitly influence or alter the value of any of the other characteristics used in the algorithm. The Bayes theorem tells us how we can compute the conditional probability. The equation for conditional probability is,
P(Ai/Bi) = (P(Bi/Ai) * P(Ai)) / P(Bi)

P(Ai/Bi) defines the probability of an event Ai occurs corresponding to the event Bi has occurred.

P (Bi / Ai) is the conditional probability and it defines the probability of occurrence of event Bi corresponding to the occurrence of the event Ai.

P(Ai) and P(Bi) defines the probability of the events Ai and Bi occurs.

9. Sequential Minimal Optimization (SMO)
The sequential minimal optimization is more effective to solve the SVM problem compared to traditional Quadratic Programming algorithms such as the interior-point method. The SMO algorithm can be viewed as a method of decomposition by which a problem of optimization of multiple variables is decomposed into a set of sub problems, each optimizing an objective feature of a limited number of variables, usually only one, whereas all other variables are treated as constants which remain unchanged in the sub problem.

10. Feature Selection
During feature selection the most relevant features are extracted from the data set. Redundancy can be avoided using this method. Since irrelevant features are excluded from the input data, feature selection can increase the accuracy of prediction. In this study Correlation Based Feature Selection (CBFS) and Principle Component Analysis (PCA) is used for feature selection. After feature selection the reduced data set is applied to four different classification Algorithm.

11. Correlation Based Feature Selection (CBFS)
Correlation values are calculated by CBFS. The five highly correlated features are selected for the analysis. These features are given as the inputs for the classifiers.

| S/N | Selected Features | Correlated Values |
|-----|-------------------|-------------------|
| 1   | AGE               | 0.2254            |
| 2   | PREVALENTHYP     | 0.2164            |
| 3   | AGE               | 0.2254            |
| 4   | PREVALENTHYP     | 0.2164            |
| 5   | PREVALENTHYP     | 0.2164            |

12. Principal Component Analysis (PCA)
Thirteen features were selected by PCA during feature selection. The features selected by the PCA algorithm are AGE, PREVALENTHYP, SYSBP, DIABP, DIABETES, SEX, BPMEDS, TOTCHOL, PREVALENTSTROKE, CIGSPERDAY, EDUCATION, BMI, CURRENT SMOKER.

13. Result and Discussion
In the study two feature selection methods are used for comparison - Principle Component Analysis (PCA) and Correlation Based Feature Selection (CBFS). After dimensionality reduction the reduced dataset is applied to four different classification Algorithm such as
Multilayer Perceptron (MLP), AdaBoost, Naive Bayes and SMO. Five most correlated features were selected and applied to Classifiers in CBFS Method. Thirteen features were selected by Principal Component Analysis (PCA). The result is shown in Table 4. From the results it is found out that Correlation Based Feature Selection (CBFS) along with MLP algorithm shows maximum accuracy.

Table 2. Comparison of Predictive accuracy of Models.

| S/N | Algorithm     | CBFS       | PCA       | Before FS |
|-----|---------------|------------|-----------|-----------|
| 1   | MLP           | 84.9057    | 83.9151   | 84.1509   |
| 2   | ADABOOST      | 84.8113    | 84.8821   | 84.8821   |
| 3   | Naive Bayes   | 81.1792    | 80.0472   | 80.0472   |
| 4   | SMO           | 84.8113    | 84.8113   | 84.8113   |

Figure 2. Comparison of Accuracy of classifiers corresponding to CBFS and PCA Feature Selection Methods.

Eight Models are generated by combining two feature selection algorithms and four classifiers. They are CBFS-MLP, CBFS-ADABOOST, CBFS-NB, CBFS-SMO, PCA-MLP, PCA-ADABOOST, PCA-NB and PCA-SMO. The accuracy of various Models are shown in Table 1. The results from table proves that Correlation Based Feature Selection along with Multilayer Perceptron (CBFS-MLP) Model perform better than the other Models with the accuracy of 84.9057 Percentage.

The result from Table 2 proves that CBFS-MLP combination shows better performance than MLP classifier. Also when we are comparing CBFS-MLP and PCA-MLP models accuracy measures proves that CBFS-MLP combination shows more performance.
Table 3. Comparison of Predictive Accuracy of MLP.

| MLP         | Precision | Recall | F Measure | ROC  |
|-------------|-----------|--------|-----------|------|
| Before FS   | 0.790     | 0.842  | 0.798     | 0.671|
| After CBFS  | 0.805     | 0.849  | 0.791     | 0.649|
| After PCA   | 0.784     | 0.795  | 0.668     | 3.62 |

![Figure 3. Comparison of Accuracy of MLP Classifier.](image)

14. Conclusion and Findings
During the study the performance of two different feature selection methods CBFS and PCA are evaluated. Eight different classifier models are developed by combining the feature selection and classification algorithms. The performance of each model was evaluated. Performance measures such as Accuracy, Precision, Recall, F Measure and ROC are evaluated for finding out the best classifier. From the result it is proven that the model CBFS with MLP Classifier shows the maximum performance for FHS dataset.

References
[1] Devansh Shah, Samir Patel and Santosh Kumar Bharti 2020 Heart Disease Prediction using Machine Learning Techniques Springer.
[2] Hamidreza and Ashrafi Esfahani 2017 Cardiovascular disease detection using a new ensemble classifier IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI).
[3] V.V.Ramalingam 2018 Prediction of Heart Diseases Using Machine Learning International Journal of Engineering and Technology.
[4] Amin Ul Haq 2019 A Hybrid Intelligence System Frame Work for Prediction of Heart Disease Using ML, Mobile Information Systems.
[5] Shadman Nashif Heart Disease Detection by Using Machine Learning Algorithm and a Real time Cardiovascular Health Monitoring System World journal of Engineering and Technology,vol.06 no.04,2018,article id 88650.
[6] Poornima Singh 2018 Effective heart disease prediction system using data mining techniques International journal of Nano medicine.
[7] M. Gunay and T. Ensari 2019 Predictive churn analysis with machine learning methods IEEE 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.
[8] R. Suguna, M. Shyamala Devi and Rincy Merlin Mathew 2019 Customer Churn Predictive Analysis by Component Minimization using Machine Learning International Journal of Innovative Technology and Exploring Engineering.
[9] Shyamala Devi, Munisamy, Suguna Ramadass and Aparna Joshi 2020 Cultivar Prediction of Target Consumer
Class using Feature Selection with Machine Learning Classification Springer’s book series “Learning and Analytics in Intelligent Systems Springer, LAIS vol. 3, pp. 604-612, 2019.

[10] Suguna Ramadas and Shyamala Devi 2019 Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers Springer’s book series Learning and Analytics in Intelligent Systems, Springer.

[11] R. Suguna, M. Shyamala Devi, Rupali Amit Bagate and Aparna Shashikant Joshi 2019 Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification Journal of Statistics and Management Systems, Taylor Francis.

[12] M. Shyamala Devi, Rincy Merlin Mathew and R. Suguna 2019 Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis International Journal of Recent Technology and Engineering.

[13] Debjani Panda, Ratula Ray, Azian Azamimi Abdullah and Satya Ranjan Dash 2019 Role of Feature Selection in Prediction of Heart Disease International Conference on Biomedical Engineering (ICoBE) IOP Publishing

[14] Bandari Sai Santosh, Dharma Sahith Reddy, M Sai Vardhan and Dr. Shaik Subhani 2019 Heart Disease Prediction with PCA and SRP International Journal of Engineering and Advanced Technology.

[15] Mothe Sunil Goud 2019 Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning International Journal of Innovative Technology and Exploring Engineering.