International Journal of Computer
Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 7, Issue 2, Feb–Mar 2018; 19–26

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

HEART DISEASES PREDICTION USING CLASSIFICATION

Poornima M

Research Scholar, Department of Electronics and Communication Engineering, Government Polytechnique, Mirle, India

ABSTRACT

Over the past ten years, heart disease has been the leading cause of death worldwide. In the United States alone, heart disease claims the lives of nearly one person every minute. Researchers have been assisting medical personnel in the identification of heart disease by employing a variety of data mining approaches. On the other hand, fewer tests are needed when data mining techniques are used. To lower the number of heart disease-related deaths, a rapid and effective detection method is required. Among the useful techniques for data mining is the decision tree. This study examines several categorization algorithms to improve heart disease diagnostic accuracy. Neural networks, Support Vector Machines(SVM), Logistic regression and Navie Bayes algorithms are employed. The algorithms' performance is tested and validated using pre-existing datasets of patients with heart disease from the UCI repository's Cleveland database. There are 13 attributes and 303 instances used for the study. Ten-fold cross validation method is used, and among all Naïve bayes classifier and Neural Network gave highest accuracy of 84% compared to other classifiers.

KEYWORDS: *Navie Bayes; Support Vector Machine; Neural Network, Logistic Regression; Classification*

Article History

Received: *19 Mar 2018* | ***Revised:*** *25 Mar 2018* | ***Accepted:*** *31 Mar 2018*

INTRODUCTION

According to the World Health Organization (2007), heart disease has been the world's top cause of mortality for the last ten years. According to the European Public Health Alliance, 41% of deaths are related to circulatory illnesses, heart attacks, and strokes (European Public Health Alliance 2010). Heart illness can have a variety of symptoms, which makes it challenging to identify it more accurately and quickly. Working with databases of individuals with heart disease is analogous to real-world application. The ability of doctors to weigh each attribute. The characteristic with the most influence on illness prediction is given more weight. To aid in the diagnosis process, it seems acceptable to attempt employing the expertise and experience of multiple professionals gathered in databases. Additionally, it gives medical professionals another source of information to draw from when making decisions.

Large volumes of health-care data are gathered by the healthcare sector, and these data must be mined to find hidden information necessary for wise decision-making. Researchers have been employing data mining techniques to assist medical professionals in the diagnosis of heart disease because of the annual global increase in heart disease patient mortality and the abundance of patient data available for the extraction of valuable knowledge [1]. The process of examining huge datasets to find hidden and undiscovered links, patterns, and knowledge that are challenging to find using conventional statistical approaches is known as data mining[2].Data mining, then, is the process of mining or obtaining

knowledge from vast amounts of data. Applications for data mining will be utilized to improve health policy-making, avoid hospital errors, detect illnesses early, and stop avoidable hospital mortality [3]. Using patient clinical data, a heart disease prediction system can help doctors make predictions about heart disease[4]. Therefore, it is possible to forecast more accurately the likelihood that a patient would receive a heart disease diagnosis by establishing a heart disease prediction system employing data mining techniques and performing some type of data mining on various heart disease features. In this paper we have used different classification algorithms for prediction.

LITERATURE REVIEW

Even while data mining has been around for longer than 20 years, its full potential is just now becoming understood. To uncover hidden patterns and relationships in huge databases, data mining integrates statistical analysis, machine learning, and database technology [5]. "A process of nontrivial extraction of implicit, previously unknown, and potentially useful information from the data stored in a database" is how Fayyad defines data mining [6]. Giudici describes it as "a process of selecting, exploring, and modeling large quantities of data to find previously unknown regularities or relations with the goal of obtaining clear and useful results for the database owner" [7].Two learning methodologies are used in data mining: supervised and unsupervised learning. While unsupervised learning (such as kmeans clustering) uses no training set, supervised learning uses a training set to learn model parameters [8].

The goal of each data mining approach varies based on the modeling target. The two most popular modeling goals are prediction and categorization. While prediction models forecast continuous-valued functions, classification models forecast categorical labels (discrete, unordered) [9]. While Regression, Association Rules, and Clustering utilize prediction algorithms, Decision Trees and Neural Networks use classification techniques [10].

CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5 are examples of decision tree algorithms. The way these algorithms choose which splits to make, when to halt a node from splitting, and which class to give to a non-split node are different. A partition's or a set of training tuples' impurity is measured by CART using the Gini index [9]. High dimensional category data can be handled by it. Regression and other continuous data types can also be handled by decision trees, but they need to be transformed into categorical data first.

The foundation of many machine learning and data mining techniques is Naive Bayes, sometimes known as Bayes' Rule [11]. Models having predictive power are produced using the rule (algorithm). It offers fresh approaches to data exploration and comprehension. By determining the correlation between the target, or dependent variable, and other, or independent, variables, it gains knowledge from the "evidence."

Three layers make up neural networks: input, hidden, and output units (variables). The assigned value (weight) of each individual input unit determines the relationship between it and the hidden and output units. It is more significant the heavier it is. Both linear and sigmoid transfer functions are used by neural network techniques. When training with minimal inputs and a vast quantity of data, neural networks operate well. It's employed when more methods don't work well enough.

MATERIALS AND METHODOLOGY

The comparison of classification techniques involves the utilization of the Cleveland dataset sourced from the UCI repository, a renowned hub for machine learning datasets. This dataset comprises 303 records, each associated with 76 attributes. However, for the specific study, only 13 attributes are considered pertinent, likely due to their relevance to the classification task at hand. The selection process for these attributes is typically informed by domain knowledge, feature importance analysis, or prior research. By focusing on a subset of attributes, the study aims to streamline the analysis process and ensure that only the most informative features are used for classification purposes.

The overarching goal of the study is to evaluate and contrast various classification techniques using the Cleveland dataset. This involves applying different machine learning algorithms to the dataset and assessing their performance in accurately classifying instances. Through this comparative analysis, researchers seek to identify the strengths and weaknesses of each technique, thereby gaining insights into their suitability for real-world applications. Ultimately, the findings of this study can inform decision-making processes in fields such as healthcare, where accurate classification of medical conditions, such as heart disease in this case, holds significant importance. The details of the attributes are tabulated in table 1. The correlation between the attributes were calculated using Pearson correlation factor and figure 1 shows the correlation between the attributes. The attributes related to exercise and depression have negative correlation with the target attribute whereas slope and cp have positive correlation.

Table 1: Attributes of the Data Set

Name	Type	Description
Age	Continuous	Age in years
Sex	Boolean	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Boolean	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang	Boolean	Continuous Maximum Heart Rate Achieved Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Oldpeak	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopythat ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Target	Boolean	1 = Yes 0 = No

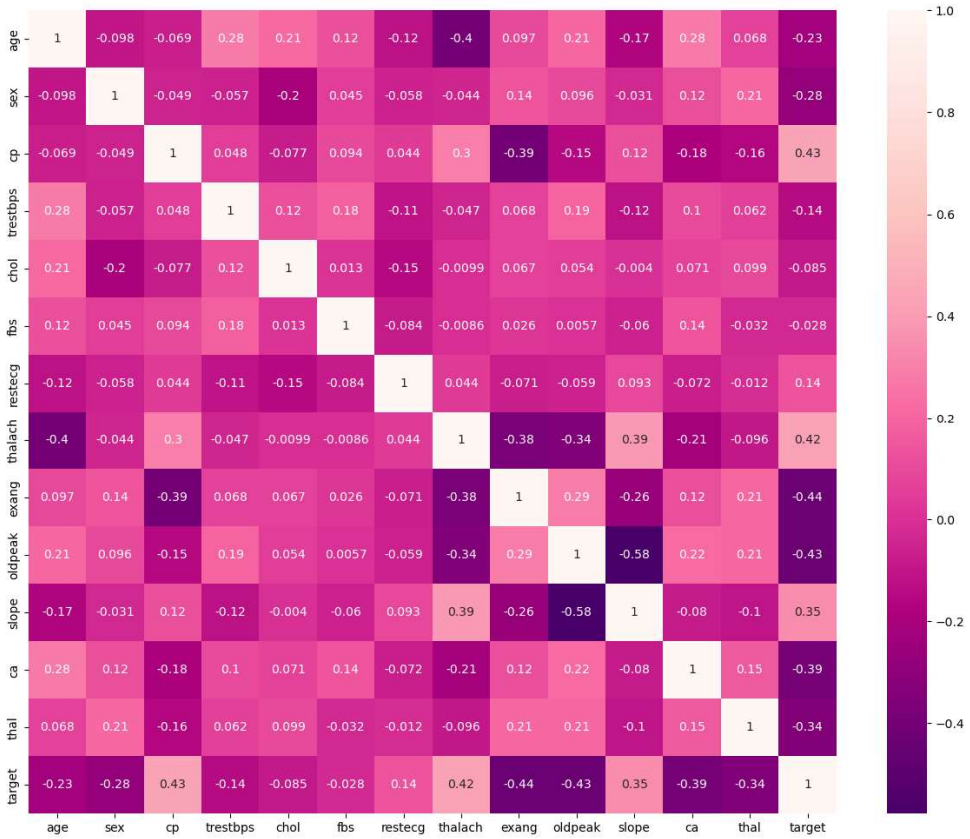

Figure 1: Correlation Factors between the Attributes.

The following classification techniques are used for the study.

Neural Networks

Neural networks, inspired by the structure of the human brain, are powerful computational models consisting of interconnected nodes arranged in layers. Each node, or neuron, processes input data using an activation function and passes the result to subsequent layers, allowing for complex pattern recognition and decision-making. Through a process called training, neural networks adjust the weights of connections between neurons to minimize the difference between predicted and actual outcomes. With their ability to learn from large datasets and handle nonlinear relationships, neural networks excel in tasks such as image and speech recognition, natural language processing, and predictive analytics, making them fundamental tools in modern artificial intelligence and machine learning applications.

Support Vector Machines

Support Vector Machines (SVMs) are a powerful class of supervised learning algorithms used for classification and regression tasks. SVMs aim to find the optimal hyperplane that separates data points into different classes while maximizing the margin between the classes. They achieve this by identifying support vectors, which are the data points closest to the decision boundary. SVMs are effective in high-dimensional spaces and can handle complex datasets by using kernel functions to map the data into higher-dimensional spaces. They are known for their ability to generalize well, even in cases where the number of features exceeds the number of samples. SVMs have found applications in various fields, including image recognition, text classification, and bioinformatics.

Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability of an instance belonging to a particular class. Despite its name, logistic regression is a linear model that applies the logistic function (also known as the sigmoid function) to map input features to a probability value between 0 and 1. This probability is then used to make binary decisions. Logistic regression estimates parameters through maximum likelihood estimation, optimizing them to best fit the observed data. It is widely used due to its simplicity, interpretability, and effectiveness in various domains such as healthcare (e.g., predicting disease outcomes), marketing (e.g., customer churn prediction), and finance (e.g., credit risk assessment). Additionally, logistic regression can be extended to handle multiclass classification tasks through techniques like one-vs-rest or multinomial logistic regression.

Naive Bayes

It is a simple, yet powerful probabilistic classification algorithm based on Bayes' theorem with the "naive" assumption of feature independence. Despite its simplicity, Naive Bayes performs well in various real-world applications, particularly in text classification and spam filtering. It calculates the probability of a given instance belonging to each class based on the probability distributions of its features. Despite its "naive" assumption, Naive Bayes often produces surprisingly accurate results and is computationally efficient, making it suitable for large-scale datasets. Additionally, it requires relatively few parameters to be estimated, making it robust to overfitting and suitable for situations with limited training data. Despite its simplicity, Naive Bayes can serve as a strong baseline model for more complex classification tasks and is widely used in combination with other algorithms in ensemble methods.

RESULTS AND DISCUSSION

The classifiers are applied for the model with 10 ten-fold cross validation. In 10-fold cross-validation, a dataset is divided into ten equal-sized subsets or folds. In each iteration, nine folds are used for training the model, while the remaining fold is used for testing. This process is repeated ten times, with each fold serving as the test set exactly once. By averaging the performance across all iterations, 10-fold cross-validation provides a robust estimate of model performance, helping to mitigate issues such as overfitting and bias. This technique is widely used to assess model generalization and select the best-performing algorithm or parameter configuration.

The Neural Network classifier is configured with 100 neurons, which are the computational units within the network, and it employs the hyperbolic tangent (tanh) activation function. The tanh activation function is commonly used in neural networks to introduce non-linearity into the model, allowing it to capture complex patterns in the data. Additionally, the Neural Network is trained with a maximum of 200 iterations, indicating the number of times the entire dataset is passed forward and backward through the network during training to adjust the model parameters.

The SVM classifier utilizes a linear kernel, implying that it seeks to find the optimal linear decision boundary to separate the classes in the dataset. The iteration limit is set to 100, indicating the maximum number of iterations the SVM algorithm will perform during training to optimize the decision boundary.

For the Logistic Regression model, Ridge regularization is employed. Ridge regularization is a technique used to prevent overfitting by adding a penalty term to the loss function, thereby constraining the magnitudes of the coefficients. This helps to generalize the model and improve its performance on unseen data.

After training and evaluating these classifiers, the performance metrics, such as accuracy, precision, recall, and F1-score, are tabulated in Table 2. These metrics provide insights into how well each classifier performs in correctly classifying instances within the dataset. The comparison in Table 2 allows for an informed decision regarding the most suitable classifier for the specific classification task at hand, based on its performance characteristics.

Table 2: The Performances of the Classifiers

Model	AUC	Accuracy	F1	Precision	Recall
SVM	0.81	0.76	0.76	0.76	0.76
Neural Network	0.90	0.84	0.83	0.84	0.84
Naive Bayes	0.91	0.84	0.84	0.84	0.84
Logistic Regression	0.90	0.82	0.82	0.82	0.82

The results indicate that Naive Bayes and the Neural Network achieved the highest AUC scores of 0.91 and 0.90 respectively, showcasing their strong discriminatory ability in distinguishing between classes. Moreover, both models demonstrated comparable performance in terms of accuracy, precision, recall, and F1-score, with Naive Bayes slightly edging out the Neural Network by achieving identical scores of 0.84 across these metrics. SVM and Logistic Regression also exhibited respectable performance, albeit slightly lower compared to Naive Bayes and the Neural Network, with AUC scores of 0.81 and 0.90 respectively, and maintaining consistency across other performance measures. These results collectively suggest that Naive Bayes and the Neural Network may be preferred choices due to their superior discriminatory power and well-rounded performance in this classification task.

The ROC (Receiver Operating Characteristic) curve as shown in figure 2, provides a comprehensive visualization of the performance of classification models across different discrimination thresholds. In the context of the provided results, the ROC curves for each model—SVM, Neural Network, Naive Bayes, and Logistic Regression would illustrate their ability to trade off true positive rate (sensitivity) against the false positive rate (1-specificity) across various threshold values. AUC (Area Under the ROC Curve) values accompany these curves, summarizing the overall discriminatory power of each model. With Naive Bayes and the Neural Network boasting the highest AUC scores of 0.91 and 0.90 respectively, their corresponding ROC curves would likely exhibit a steeper ascent, indicating superior discriminative ability. SVM and Logistic Regression, with AUC scores of 0.81 and 0.90 respectively, would also present ROC curves, albeit with slightly less pronounced performance. Overall, the ROC curves provide a visual aid to assess the trade-offs between true positive and false positive rates, aiding in the selection of the most suitable model for the classification task at hand.

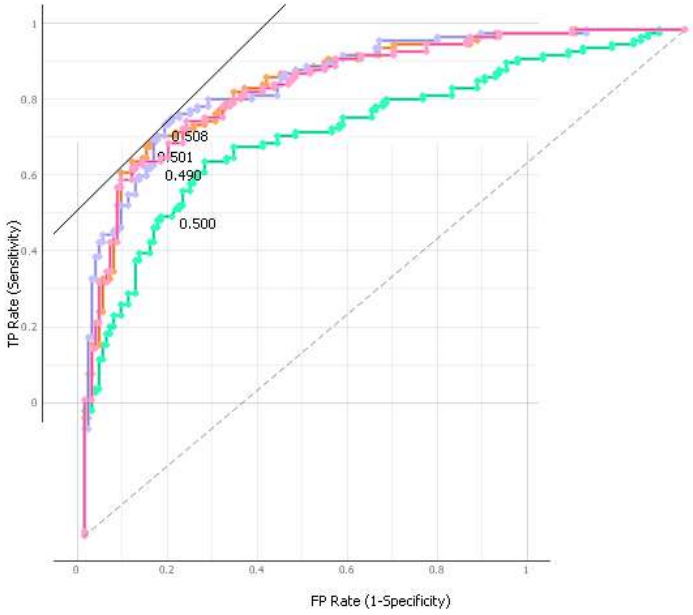

Figure 2: The ROC Analysis of the Classifiers.

CONCLUSION

In conclusion, the performance evaluation of classification algorithms on the pre-existing dataset of patients with heart disease from the UCI repository's Cleveland database reveals noteworthy insights. Naive Bayes and the Neural Network emerged as top performers, exhibiting high AUC scores of 0.91 and 0.90 respectively, along with consistent accuracy, precision, recall, and F1-scores of 0.84. Both models demonstrated robust discriminatory power, making them promising candidates for heart disease classification tasks. While SVM and Logistic Regression displayed slightly lower AUC scores of 0.81 and 0.90 respectively, they still showcased respectable performance across various metrics. These findings underscore the efficacy of machine learning techniques in healthcare decision-making, particularly in diagnosing heart disease using existing patient data.

In future work, it would be beneficial to explore ensemble methods that combine the strengths of multiple classifiers to further enhance predictive performance for heart disease classification. Additionally, investigating feature engineering techniques to optimize model inputs and leveraging deep learning architectures may yield improvements in accuracy and generalization. Furthermore, extending the analysis to include more diverse and larger datasets from different sources could enhance the robustness and applicability of the developed models in real-world clinical settings. Lastly, conducting prospective studies to validate the effectiveness of the developed classifiers in clinical practice would be essential for translating these findings into actionable insights for healthcare professionals.

REFERENCES

1. *C. Helma, E. Gottmann, and S. Kramer, "Knowledge discovery and data mining in toxicology," Statistical methods in medical research, vol. 9, pp. 329-358, 2000.*
2. *I.-N. Lee, S.-C. Liao, and M. Embrechts, "Data mining techniques applied to medical information," Medical informatics and the Internet in medicine, vol. 25, pp. 81-102, 2000.*

3. *R. Canlas, "Data mining in healthcare: Current applications and issues," School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.*
4. *S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS international conference on computer systems and applications, 2008, pp. 108-115.*
5. *B. Thuraisingham, "A primer for understanding and applying data mining," It Professional, vol. 2, pp. 28-31, 2000.*
6. *U. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific databases," in Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No. 97TB100150), 1997, pp. 2-11.*
7. *P. Giudici, Applied data mining: statistical methods for business and industry: John Wiley & Sons, 2005.*
8. *M. K. Obenshain, "Application of data mining techniques to healthcare data," Infection Control & Hospital Epidemiology, vol. 25, pp. 690-695, 2004.*
9. *J. Han, M. Kamber, and D. Mining, "Concepts and techniques," Morgan Kaufmann, vol. 340, pp. 94104-3205, 2006.*
10. *C. Kleissner, "Data mining for the enterprise," in Proceedings of the Thirty-First Hawaii International Conference on System Sciences, 1998, pp. 295-304.*
11. *Z. Tang and J. Maclennan, Data mining with SQL Server 2005: John Wiley & Sons, 2005.*