Analysis of Heart Disease Prediction using Machine Learning Classification Algorithms

Anjani Kumar[1], Dr. Narendra Sharma[2]
Research Scholar, Department of CSE,SOE, SSSUTMS, Sehore Madhya Pradesh, India[1]
Research Scholar, Department of CSE,SOE, SSSUTMS, Sehore Madhya Pradesh, India[2]

Abstract: *In recent decades, heart disease, also known as cardiovascular disease, has emerged as the leading cause of death worldwide. It encompasses a range of conditions that affect the heart and is influenced by various risk factors. It has become increasingly imperative to develop accurate, dependable, and efficient methods for early diagnosis to facilitate timely disease management. To address this challenge, data mining has emerged as a valuable tool in the healthcare domain. Researchers have employed various data mining and machine learning techniques to analyze vast and complex medical datasets, aiding healthcare professionals in predicting the onset of heart disease. This research paper focuses on exploring different attributes associated with heart disease and building predictive models using supervised learning algorithms such as Naïve Bayes, decision trees, K-nearest neighbor, and the random forest algorithm. To conduct this analysis, an existing dataset from the Cleveland database of the UCI repository, comprising records of heart disease patients, is utilized. The dataset contains 303 instances and 76 attributes. However, for the purpose of this study, only 14 critical attributes are selected for testing to assess the performance of various algorithms. The primary objective of this research is to assess the likelihood of individuals developing heart disease. The findings reveal that the K-nearest neighbor algorithm achieves the highest accuracy score among the tested algorithms, demonstrating its effectiveness in predicting heart disease.*

Keywords: *Heart Disease, Cardiovascular Disease, Data Mining, Predictive Models, K-nearest Neighbor Algorithm*

1. Introduction

Over the past decade, cardiovascular disease, commonly known as heart disease, has continued to maintain its status as the leading cause of death worldwide. According to estimates from the World Health Organization, over 17.9 million deaths occur annually across the globe due to cardiovascular diseases. Of these fatalities, a staggering 80% can be attributed to conditions like coronary artery disease and cerebral stroke [1]. This disproportionately high number of deaths is particularly prevalent in low and middle-income countries [2]. Heart disease is influenced by a multitude of factors, including personal lifestyle choices, professional habits, and genetic predispositions. Among these factors, habitual risk elements like smoking, excessive alcohol and caffeine consumption, stress, and physical inactivity, combined with physiological factors such as obesity, hypertension, high blood cholesterol, and pre-existing heart conditions, contribute significantly to the prevalence of heart disease. The efficient, accurate, and early diagnosis of heart disease plays a pivotal role in implementing preventive measures to reduce mortality rates.

Data mining is the process of extracting valuable information from vast datasets across various fields, including the medical domain, business sector, and education sector. Machine learning, a rapidly advancing subfield of artificial intelligence, offers powerful algorithms capable of analyzing extensive datasets from diverse domains, with healthcare being a particularly important one. Machine learning serves as an alternative to conventional prediction modeling approaches by employing computers to discern complex, nonlinear interactions among various factors, thereby minimizing errors in predicting outcomes compared to real-world results [3]. Data mining involves the exploration of massive datasets to uncover hidden, critical decision-making information from a historical repository, which can be utilized for future analysis. The field of medicine, in particular, accumulates an immense volume of patient data, necessitating the application of various machine learning algorithms for data mining. Healthcare professionals harness these algorithms to analyze patient data effectively, enabling precise diagnostic decision-making.

Medical data mining, powered by classification algorithms, offers invaluable clinical support by facilitating in-depth analysis and aiding healthcare professionals in making informed decisions. Data mining is a crucial process involving the extraction of valuable data and information from vast databases. Within the realm of healthcare, specifically in predicting heart disease, various data mining techniques come into play. These techniques encompass regression, clustering, association rules, and classification methods like Naïve Bayes, decision trees, random forests, and K-nearest neighbors. This research undertakes a comparative analysis of these classification techniques to assess their efficacy in classifying various attributes related to heart disease. For this study, a dataset has been sourced from the UCI repository, serving as the foundation for the development of a classification model aimed at predicting heart disease. The research not only delves into the algorithms employed for heart disease prediction but also conducts a comprehensive comparison with existing systems in the field. This comparative analysis sheds light on the strengths and weaknesses of various approaches, aiding in the selection of the most suitable algorithm for accurate predictions. Furthermore, this research paves the way for future investigations and advancements in the domain of heart disease prediction. It outlines potential avenues for further research and development, highlighting the dynamic nature of this field and the ongoing quest for enhanced predictive accuracy and clinical utility.

2. Background

Heart disease continues to be a pervasive health concern, affecting millions of individuals globally, and it retains its status as the leading cause of mortality worldwide. In the realm of medical diagnosis, the imperative is clear: diagnoses must be proficient, dependable, and augmented by computer-based techniques to mitigate the effective costs associated with diagnostic tests. Data mining emerges as a software technology that empowers computers to construct and classify diverse attributes, making it a valuable tool in the context of heart disease prediction.

2.1 Machine Learning

Machine learning stands as an evolving subset of artificial intelligence, with its central objective being the creation of systems capable of learning and making predictions based on acquired experiences. This discipline entails the training of machine learning algorithms through the utilization of a training dataset, ultimately resulting in the construction of a predictive model. This model subsequently employs new input data to forecast the likelihood of heart disease. By harnessing machine learning, it becomes possible to uncover concealed patterns within the input dataset, facilitating the construction of predictive models that deliver accurate forecasts for new datasets.

The process entails the cleansing of the dataset and the handling of missing values to ensure data integrity. Subsequently, the model is employed to make predictions regarding heart disease, and its accuracy is assessed through testing. Machine learning techniques encompass a range of methods, including:

Supervised Learning

Supervised learning involves training a model on a labeled dataset, where each data point contains input information along with its corresponding outcomes. The dataset is typically categorized and then divided into two subsets: a training dataset used to educate the model and a test dataset that functions as new, unseen data to assess the model's accuracy. The training dataset imparts knowledge to our model, which is then tested using the separate dataset to gauge its performance. Supervised learning encompasses both classification and regression tasks.

Unsupervised Learning

In unsupervised learning, the dataset used for training lacks labels or classifications. The primary objective is to uncover hidden patterns within the data. The model is trained to identify and develop these patterns. Although it can predict concealed patterns in new input datasets, it primarily draws conclusions and insights from the dataset itself. In this approach, no predetermined responses or classifications are provided within the dataset. Clustering is a common example of an unsupervised learning technique.

Reinforcement Learning

Reinforcement learning operates without the use of labeled datasets or predetermined results associated with the data. Instead, the model learns through experience. It continually refines its performance based on interactions with its environment and aims to optimize its decision-making process. This involves assessing various possibilities and learning from the consequences of its actions to achieve the desired outcome.

3. Classification Machine Learning Techniques

Classification tasks are employed to predict future cases based on past data. Numerous data mining techniques, such as Naïve Bayes, neural networks, and decision trees, have been utilized by researchers to achieve precise diagnoses in the context of heart disease. The accuracy yielded by these various techniques can vary depending on the number of attributes considered. This research endeavors to provide diagnostic accuracy scores to enhance healthcare outcomes. The WEKA tool is employed for dataset preprocessing in this research, with the data formatted in ARFF (attribute-relation file format). Out of the 76 distinct attributes available, only 14 are selected for analysis to ensure precise results. By conducting comparisons and analyses utilizing different algorithms through the WEKA tool, it becomes possible to predict and address heart disease promptly, ultimately leading to improved health outcomes [5].

In pursuit of our research objectives, this study extensively explores the utilization of various machine learning algorithms on the dataset, and it includes a comprehensive analysis of the dataset itself. Furthermore, this paper provides insights into which attributes within the dataset exert a more significant influence on the prediction process, ultimately enhancing the precision of our predictive models. Such insights can potentially lead to cost savings in medical diagnosis, as not all attributes may carry equal weight in predicting outcomes [5].

This research draws upon a dataset sourced from the UCI Machine Learning Repository. The dataset consists of real-world data comprising 300 instances, each characterized by 14 distinct attributes. Among these attributes are 13 predictors and one class variable, encompassing factors such as blood pressure, type of chest pain, electrocardiogram results, and more (refer to Table 1). In our research, we employ four different algorithms to discern the determinants of heart disease and construct a model capable of achieving the highest attainable accuracy.

Real-world datasets often present challenges in the form of missing and noisy data. Therefore, data pre-processing steps are essential to address these issues and enable robust predictive modeling. Figure 1 illustrates the sequential flow of our proposed model.

Cleaning the Data: The data collected from real-life sources typically contain noise and missing values. To ensure accurate and effective results, these datasets undergo cleaning to eliminate noise and fill in missing values.

Transformation: Data transformation involves altering the format of the data from one representation to another, making it more comprehensible. This process encompasses tasks such as smoothing, normalization, and aggregation.

These pre-processing steps serve as crucial preparatory measures, ensuring that the data is of high quality and suitable for the subsequent application of machine learning algorithms.

Naïve Bayes Classifier

The Naïve Bayes classifier, a supervised algorithm, is a straightforward classification technique that leverages Bayes' theorem. It operates under the "naïve" assumption of strong independence among attributes. In essence, this means that the predictors are assumed to be unrelated to each other and possess no correlation. Instead, all attributes independently contribute to the probability to maximize predictive accuracy. The Naïve Bayes model does not employ Bayesian methods.

This algorithm is employed in various complex real-world situations, with Bayes' theorem serving as the mathematical foundation for obtaining probabilities. The key components include:

P(X/Y): Posterior probability
P(X): Class prior probability
P(Y): Predictor prior probability
P(Y/X): Likelihood probability of the predictor

Naïve Bayes is appreciated for its simplicity, ease of implementation, and efficiency in handling non-linear and complex data. However, its reliance on assumptions of class conditional independence may result in a loss of accuracy.

In previous studies, Naïve Bayes achieved an accuracy of 84.1584% with the selection of the 10 most important predictors using SVM-RFE, while an accuracy of 83.49% was attained using all 13 attributes from the Cleveland dataset [8].

4. Decision Tree

The Decision Tree algorithm is a classification method suitable for both categorical and numerical data. It constructs tree-like structures to facilitate data analysis. Decision trees are renowned for their simplicity and are commonly used in medical datasets. They provide a clear, tree-shaped visualization of data analysis, with three key nodes:
Root Node: The primary node governing the functioning of all other nodes.
Interior Node: These nodes handle various attributes.
Leaf Node: Represents the result of each test.
The algorithm partitions the data into two or more sets based on essential indicators. It calculates the entropy of each attribute and then divides the data, prioritizing predictors with maximum information gain or minimum entropy. Decision trees yield results that are easily interpretable.
This algorithm tends to exhibit higher accuracy compared to other methods due to its tree-based data analysis. However, it may lead to over-classification, and only one attribute is tested at a time for decision-making. In previous research, a decision tree achieved an accuracy of 71.43% [9], while another study reported a significantly lower accuracy of about 42.8954% [10].

K-Nearest Neighbor (K-NN)
The K-Nearest Neighbor (K-NN) algorithm, a supervised classification method, categorizes objects based on their proximity to neighboring data points. It is a form of instance-based learning, with the calculation of attribute distances typically measured using Euclidean distance. K-NN utilizes a group of labeled points to inform how to classify another point. The data are clustered based on similarity, and missing values can be filled using K-NN, followed by various prediction techniques applied to the dataset. Better accuracy can be achieved by exploring different algorithm combinations.
K-NN is known for its simplicity, as it does not require the creation of a model or the imposition of other assumptions. It is a versatile algorithm used for classification, regression, and search tasks. However, K-NN's accuracy can be affected by noisy and irrelevant features. In one study, an accuracy of 83.16% was achieved with a value of K=9 [8].
The Random Forest algorithm is a supervised classification technique that assembles multiple trees into a "forest." Each individual tree in the random forest provides a class prediction, and the class with the most votes becomes the model's prediction. Notably, a higher number of trees in the forest tends to result in increased accuracy.

5. Results and Analysis

The primary objective of this research is to predict the likelihood of a patient developing heart disease using supervised machine learning classification techniques, including Naïve Bayes, decision tree, random forest, and K-nearest neighbor, with a focus on the UCI repository dataset. The experiments were conducted using the WEKA tool on a system equipped with an 8th generation Intel Core i7 processor (8750H, up to 4.1 GHz CPU) and 16 GB of RAM. The dataset was divided into a training set and a test set, subjected to data preprocessing, and analyzed using supervised classification techniques: Naïve Bayes, decision tree, K-nearest neighbor, and random forest, to determine their accuracy.

6. Conclusion

The overarching goal of this study is to explore various data mining techniques for effective heart disease prediction, with an emphasis on achieving efficient and accurate predictions using a reduced number of attributes and tests. The study focuses on only 14 essential attributes from the dataset and applies four data mining classification techniques: K-nearest neighbor, Naïve Bayes, decision tree, and random forest. Among these techniques, K-nearest neighbor, Naïve Bayes, and random forest emerge as the algorithms demonstrating the most promising results in this model. Notably, the highest accuracy is achieved with K-nearest neighbors (k = 7). Looking ahead, there is room for expanding this research by incorporating additional data mining techniques such as time series analysis, clustering, association rules, support vector machines, and genetic algorithms. Acknowledging the limitations of this study, future research endeavors should explore more complex models and combinations to enhance the accuracy of early heart disease prediction. In summary, this study provides valuable insights into the potential of data mining and machine learning techniques for heart disease prediction, with practical implications for improving diagnostic accuracy and healthcare outcomes.

Reference

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
[2] Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart

disease in low-and middle-income countries. Curr Probl Cardiol. 2010;35(2):72–115.

[3] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[4] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[5] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.

[6] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16. https://doi.org/10.4236/jilsa.2017.91001.

[7] Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p. 500–504.

[8] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.

[9] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[10] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.

[11] Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In: 2017 IEEE 2nd international conference on big data analysis (ICBDA). IEEE. p. 228–32.

[12] Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M. Effective diagnosis and monitoring of heart disease. Int J Softw Eng Appl. 2015;9(1):143–56.

[13] Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.

[14] Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Comput Sci Inf Technol (IJACSIT). 2014;2:56–66.

[15] Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst (IJAIS). 2012;3(7):25–30.

[16] Deepika K, Seema S. Predictive analytics to prevent and control chronic diseases. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE. p. 381–86.

[17] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.

[18] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.