International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Heart Disease Prediction Using Machine Learning

Shourya Singh[1], Vishesh Singh[2], Youg Khanna[3]

[1,2,3]SCSE, Galgotia University, Greater Noida, India. Yougkhanna103@gmail.com

Abstract

Cardiovascular disease, essentially a broad term for heart disease, has topped the list of causes of death in the world for several decades. Since this condition is influenced by several risk factors, there is a considerable need for precise and reliable methods that will enable early diagnosis and timely management. Data mining is emerging as an important technique for managing large amounts of medical data, helping health professionals make more precise predictions regarding heart disease.The present work uses various techniques of machine learning and data mining to analyze complex medical datasets. This work specifically focuses on supervised learning algorithms, including Naïve Bayes, decision tree, KNN, and random forest. The employed data set comes from the Cleveland database in the UCI repository, which deals with data from patients affected by heart diseases. Although the dataset has 303 instances and 76 attributes, the analysis is performed only on the selected 14 key attributes to estimate the performance of a few models. In this research, the determination of heart disease in patients is the goal.

Keywords: Test case prioritization, Genetic algorithm, Machine learning.

Introduction

Machine learning is a powerful approach that allows us to analyze data, uncover hidden patterns, and extract meaningful insights. It is a rapidly growing field with expanding applications across various domains. Machine learning comprises different learning methods, including supervised, unsupervised, and ensemble learning, which help improve predictions and assess the accuracy of datasets. In our project on the prediction of heart diseases, HDPS, machine learning techniques are adopted to benefit humankind by picking out the victims of cardiovascular disease from their records. Cardiovascular diseases (CVDs) are now extremely common and impact millions of individuals worldwide. Approximately 17.9 million people die every year due to CVDs as reported by WHO, making this the biggest cause of death for adults. The system to be devised aims to help diagnose initial signs of heart diseases such as chest pains and high blood pressure early and therefore minimize the extensive number of medical tests to ensure prompt treatment. This project uses three major data mining techniques: Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. The accuracy of the model is as high as 87.5%, which is better than earlier systems that had used a single data mining technique. The usage of multiple algorithms increases the efficiency and accuracy of HDPS. Logistic regression is one of the most important supervised learning techniques. It is used to classify problems that have only discrete values. The dataset used for this project is from the UCI repository and is a collection of medical records for patients. By analyzing 14 key medical attributes, including age, gender, chest pain type, and fasting blood sugar level, the system predicts whether a patient is at risk of heart disease. These attributes are processed using the

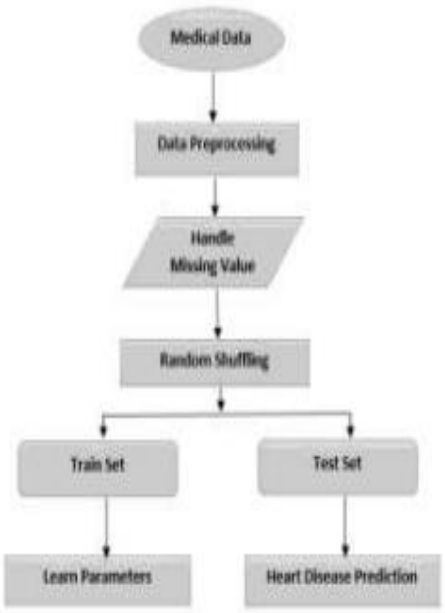
three machine learning models, with KNN yielding the highest accuracy of 88.52%. It, therefore classifies people into categories of likely to develop heart disease. The system is relatively cost-effective and would allow early diagnosis, hence possibly reducing the burden on health care systems while improving patient outcomes.

BACKGROUND AND RELATED WORK

The increasing usage of machine learning algorithms in the diagnosis of cardiovascular diseases has been a key motivation for this research. This paper provides an overview of existing literature and explores different predictive models for cardiovascular disease using algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. The results emphasize that each algorithm has its own unique advantages in achieving specific objectives. The prior models IHDPS stands for Integrated Heart Disease Prediction System, which applies existing and recently developed machine learning and deep learning models to formulate decision boundaries that might be developed from the analysis of the research. It implemented fundamental risk factors including family history regarding heart diseases. However, the accuracy associated with the performance of the model of IHDPS was more or less reduced in comparison with current approaches where advanced machine learning techniques and ANNs are incorporated to detect CHD. McPherson et al. researched key risk factors for the development of coronary heart disease and atherosclerosis. These improvements were achieved with the help of neural network-based techniques in developing the approach implemented to predict whether a patient is at risk or not to ensure positive changes towards improving the effectiveness of machine learning in the prediction of cardiovascular disease.

Methodology Predictions

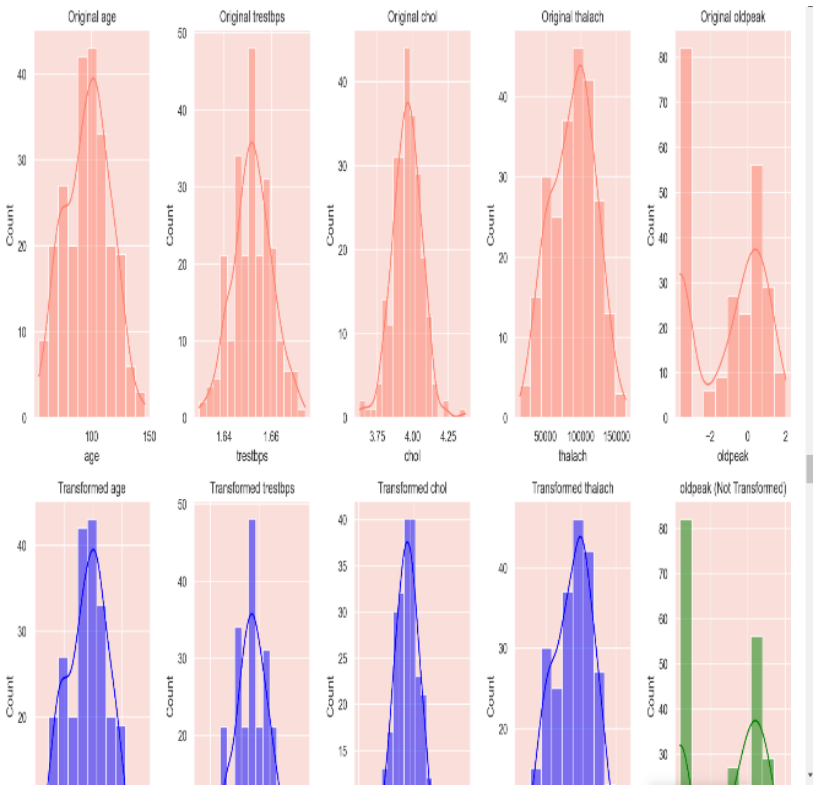
This This study employs a multi-algorithm approach to detect heart diseases using Machine Learning (ML). The methodology consists of the following stages:



Data Collection: A dataset of labeled emails (spam and non-spam) is collected from publicly available sources, such as Kaggle or UCI Machine Learning Repository, AWS dataset, Google etc.
Data Preprocessing: The collected data is preprocessed by removing stop words, punctuation, and special characters. The emails are then tokenized, and the frequency of each word is calculated using TF-IDF.
Dataset Used: We used the Cleveland Heart Disease dataset from the UCI Machine Learning Repository with features including age, sex, type of chest pain, resting blood pressure, cholesterol, and target (heart disease diagnosis).
Testing and Validation: The GUI-based system is tested and validated using a separate dataset to ensure its effectiveness in real-world scenarios.

Machine Learning Approaches

KNN algorithm:- was first introduced by Hodges et al. in the year 1951; it is a widely used nonparametric method for pattern classification, and it is considered to be one of the very simple yet highly effective classification techniques. Unlike other algorithms, KNN doesn't make prior assumptions about the data's distribution, which is useful for classification tasks where little to no information about the data is available.The algorithm KNN works on obtaining the 'k' numbers of closest data points in the training set to an unknown data point and determining its class based upon the neighbours. Classification is usually done by taking the most common class amongst the nearest neighbours or averaging them in regression tasks.Experiments performed have shown that the efficiency of KNN at different parameter settings. For example, if 'k' is 9 and the technique applied is a 10-fold cross-validation technique, then the model with accuracy of 83.16%. Other studies combining the Ant Colony Optimization algorithm reported that KNN could achieve a value for accuracy of 70.26% with an error rate of 0.526. Ridhi Saini et al. in their study had shown an efficiency level of 87.5% while working with KNN because KNN is most efficient in its classification methodologies.



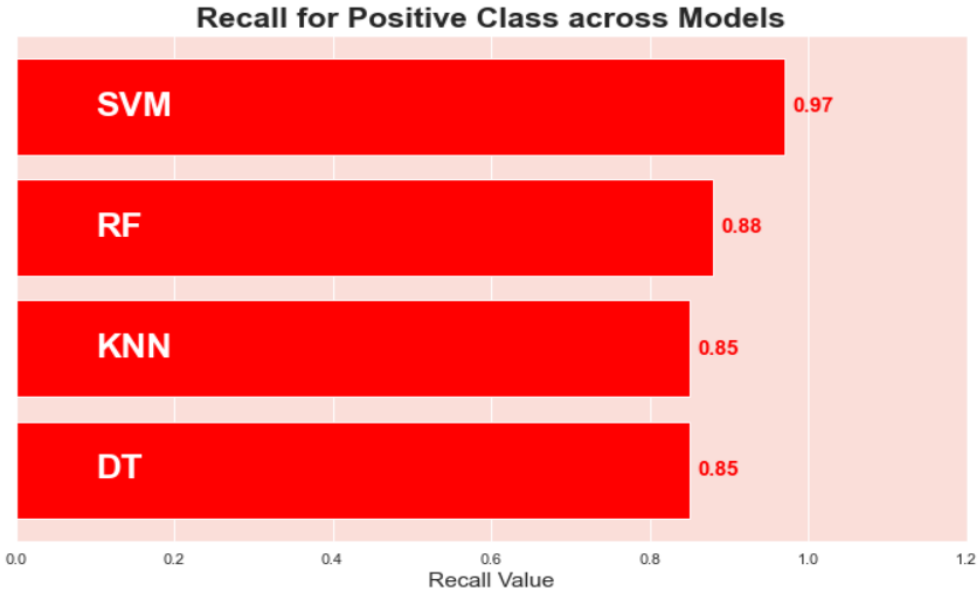
Random Forest is one of the most popular machine learning algorithms, often used in both classification and regression tasks, though it shines more clearly when it comes to classification.The algorithm works by creating multiple decision trees, which form an "ensemble" of trees. The idea is that using many decision trees together leads to more accurate and reliable predictions compared to relying on just one. This approach helps minimize overfitting and enhances the model's ability to generalize to new, unseen data. In classification tasks, Random Forest uses a **voting system**, where each tree casts a "vote" on the class of a given data point, and the class with the most votes becomes the final prediction. For regression, the algorithm averages the outputs of all the decision trees to arrive at the final prediction. Random Forest is particularly effective when working with large datasets and high-dimensional features, making it a versatile tool in many applications.

Predictions

This is a problem of binary classification (has-disease or no-disease cases). Scikitlearn has various classification algorithms and most traditional machine learning challenges begin from there, so we will start by looking into some of the classification alorithms from the sklearn libarary, which includes Logistic Regression, Nearest Neighbors, Support Vectors, Nu SVC, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis and Neural Net. We build simple models using the ML algorithms mentioned above, and later on, we will optimize them by appropriate tuning of the parameters.

Results

From the results, it's apparent that however numerous experimenters have used algorithms similar as SVC and Decision Trees for the opinion of cases suffering from heart complaint, styles similar as KNN, Random Forest Classifier, and Logistic Retrogression result in better issues. The algorithms used in this exploration yield lesser delicacy, are cheaper, and give quicker results than those in former studies. specially, in both KNN and Logistic Regression, the loftiest delicacy was set up to be 88.5, which came about by taking a larger set of medical attributes from the dataset used in this design. likewise, our findings indicate that, in this problem, Logistic Retrogression and KNN outperform the Random Forest Classifier.



Conclusion And Future Scope

A cardiovascular disease detection model has been developed using three ML This project predicts people with cardiovascular disease by extracting the patient medical history as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them having been diagnosed with a previous heart disease. Algorithms used in building the model are Logistic regression, Random Forest Classifier and KNN. The accuracy of our model is 87.5%. Using more training data ensures higher chances of the model to predict accurately whether the given person has a heart disease or not. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and KNN to get an accuracy of an average of 87.5% on our model which is better than the previous models having an accuracy of 85%. Also, it is concluded that accuracy of KNN is highest between the three algorithms that we have used i.e. 88.52%. 'Figure 6' shows 44% of people that are listed in the dataset are suffering from Heart Disease.

References

1. D Nucci, D., Panichella, A., Zaidmain, A. and De Lucia, A. 2020. “A Test Case Prioritization Genetic Algorithm Guided by the Hypervolume Indicator,” in IEEE Transactions on Software Engineering, vol. 46, no. 6, pp. 674-696, doi:10.1109/TSE.2018.2868082
2. Lachmann, R., Schulze, S., Nieke, M., Seidl, C. and Schaefer, I. 2016. “System-Level Test Case Prioritization Using Machine Learning,” 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016, pp. 361-368, doi: 10.1109/ICMLA.2016.0065.
3. Khatibsyarbini, M., Isa, M. A., Jawawi, D. N. A., Shafie, M. L. M., Wan-Kadir, W. M. N., Hamed, H. N. A., Suffain, M. D. M. 2021 “Trend Application of Machine Learning in Test Case Prioritization: A Review on Techniques, “in IEEE Access, vol. 9, pp. 166262-166282, doi:10.1109/ACCESS.2021.3135508.
4. Sabharwal, S., Sibal, R. & Sharma, C. 2014. Applying Genetic Algorithm for Prioritization of Test Case Scenarios Derived from UML Diagrams., Int J Comput Sc Issues. 8.
5. Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. 2002. “A fast and elitist multiobjective genetic algorithm: NSGA-II,” in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, doi: 10.1109/4235.996017
6. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
7. Devansh Shah, Samir Patel & Santosh Kumar Bharti ,Heart Disease Prediction using Machine Learning Techniques.345, (2020)
8. 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.
9. Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.
10. Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction", 4 th International Conference on Computing Communication And Automation(ICCCA) , 2018.

11. A. Hazra, S. Mandal, A. Gupta and Mukherjee, "A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, 2017.
12. J. Patel, P. Upadhyay and Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique", Journals of Computer Science Electronics, 2016.
13. V. Kirubha and S. M. Priya, Survey on Data Mining Algorithms in Disease Prediction, vol. 38, no. 3, pp. 124-128, 2016.
14. M. Nikhil Kumar, K. V. S. Koushik and K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools", International Journal of Scientific Research in Computer Science Engineering and Information Technology IJSRCSEIT, 2019.