DOI : https://doi.org/10.56452/6-3-674

ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

Predicting Heart Disease Risk Factors using Hawkins Dataset and Decision Tree Algorithms

Jyoti Parsola

Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

Early detection and prediction of heart disease are vital for treating and preventing this disease, which is a leading cause of worldwide mortality. This study aims to use the decision tree and the Hawkins dataset to identify risk factors for heart disease. The process of normalizing and cleaning the data is carried out before we apply the decision tree algorithm to identify the various risk factors that can lead to heart disease. The results of the study revealed that the decision tree can accurately identify various risk factors for heart diseases. It performed well in terms of accuracy, with an overall score of 98.8%. The recall, precision, and F1 score of the algorithm also indicated that it is an effective tool for identifying heart disease risks. The findings of the study have important clinical implications, as they show that decision tree algorithms can help clinicians identify high-risk patients and develop treatment plans tailored to their individual needs. The study also highlights the advantages of using such tools in healthcare, as these are easy to interpret and transparent. The study's findings show the feasibility of utilizing decision tree algorithms to identify risk factors related to heart disease. Future research will explore the algorithm's performance in populations with varying demographic and clinical characteristics.

Keywords: Heart Disease, Decision Tree, Machine learning, Cardio vascular disease(CVD).

Introduction

High mortality rates and morbidity are major public health issues related to heart disease or cardiovascular disease (CVD). Early detection and intervention are crucial to treat and prevent this condition[1]. Machine learning has the potential to help predict heart disease risk factors through analyzing large datasets. The term cardiovascular disease refers to a group of disorders that affect the blood vessels and heart. Some of these include heart failure, atherosclerosis, and congenital anomalies. Heart disease is one of the leading causes of death globally, with around 17.9 million people succumbing to it in 2019. Among the risk factors that can lead to heart disease are age, high blood pressure and cholesterol, family history, obesity, diabetes, and smoking[2], [3].

Various types of data can be used to train machine learning systems, such as demographic and clinical information, as well as imaging data. These include medical history, blood pressure, sex, age, cholesterol levels, and smoking status. With the help of genetic information, these data can help identify individuals' genetic predisposition for heart disease. Machine

learning systems are capable of handling large datasets, which is an important advantage when it comes to predicting heart disease. Traditional methods may struggle to interpret complex information, as these may not be able to identify relationships and patterns. With the help machine learning systems, risk factors can be analyzed from multiple sources, and more comprehensive analyses can be made[4].

One of the main advantages of machine learning in predicting heart disease is its ability to adapt to the changes in data. This allows the system to improve its accuracy in identifying individuals with a high risk of developing heart disease. In addition to being able to improve the accuracy of predicting heart disease, machine learning systems can also be customized to suit the needs of different patient groups. For instance, they can be used to identify individuals with diabetes and women with a certain age group. Through this method, healthcare professionals can develop effective treatment plans for high-risk individuals. Due to the immense amount of data that machine learning can analyze, its potential to improve the accuracy of predicting heart disease has been widely acknowledged. One of the technologies "Decision Tree "has the potential to transform the way healthcare professionals treat and prevent heart disease.

Decision trees are useful in identifying various risk factors for heart disease. They are easy to visualize and understand, which makes them an ideal tool for identifying crucial decision points and contributing factors. Through their use, healthcare professionals can gain a deeper understanding of the underlying causes of heart disease and develop effective treatment plans. One of the most important advantages of decision trees is their flexible nature, which allows them to handle different types of data. This allows them to train their systems on a wide variety of information, such as clinical and demographic data. They can also handle outliers and missing data, making them a reliable and robust tool when analyzing complex datasets[5], [6].

In addition to identifying various risk factors, decision trees can also help healthcare professionals identify the interactions between multiple risk factors. For instance, a decision tree might be able to identify the link between high blood pressure and smoking. This information could then be used to develop effective treatment plans and prevent heart disease. Another type of decision tree can be used by healthcare professionals to create predictive models that can be customized to specific populations or patient groups. This approach can help them identify high-risk patients and develop treatment plans that are specific to their needs[7].

Decision trees are widely used in the prediction of heart disease. They can help healthcare professionals identify various risk factors and their interactions with each other, and they can also handle missing or missing data. Due to the evolution of machine learning algorithms, decision trees are expected to play a significant role in helping healthcare professionals prevent and predict heart disease.

Literature Review

One of the most common health concerns worldwide is heart disease, which affects millions of people. Due to the increasing need for accurate and timely diagnosis, machine learning has become a promising tool for predicting heart disease. This review explores the various studies that have been conducted on the use of machine learning (ML) in predicting heart disease. They vary in their approach, such as using ECG signal analysis and mobile applications. The studies also employ various algorithms, such as hybrid and Nave Bayes models. The review will be presented in a format that is easy to understand, such as a table with details of each study, including its methodology,

data, and ML algorithm. It aims to provide an overall overview of the current status of research related to the use of machine learning in predicting heart disease as shown in table-1.

Table 1 Related works

Author et al.	Dataset	Methodology	ML algorithm	Output	Result - Accuracy
P. K. Anooj et al.[8]	Cleveland Heart Disease	Clinical Decision Support System	Weighted Fuzzy Rules	Risk level prediction of heart disease	84.81%
A. Pattekari et al.[9]	UCI Heart Disease	Prediction system	Naïve Bayes	Heart disease prediction	90.80%
C. J. Schneider et al.[10]	Not applicable	Literature review	Not applicable	Overview of machine learning in heart disease	Not applicable
K. Uyar et al.[2]	Cleveland Heart Disease	Recurrent Fuzzy Neural Networks	Genetic Algorithm	Diagnosis of heart disease	91.33%
A. U. Haq et al.[11]	Cleveland Heart Disease	Hybrid Intelligent System	Machine Learning Algorithms	Heart disease prediction	94.15%
S. Mohan et al.[12]	Cleveland Heart Disease	Hybrid Machine Learning Techniques	Random Forest, kNN, Naïve Bayes	Heart disease prediction	86.30%
S. J. Al'Aref et al.[13]	Not applicable	Literature review	Not applicable	Applications of ML in CVD	Not applicable
S. Uddin et al.[14]	Cleveland Heart Disease	Disease prediction	SVM, Random Forest, Decision Tree	Heart disease prediction	89.57%
J. Chen et al.[15]	MIT-BIH Arrhythmia	Smart Heart Monitoring	ECG Signal Analysis	Heart problem prediction	98%
G. T. Reddy et al.[16]	UCI Heart Disease	Genetic Algorithm and Fuzzy Logic Classifier	Heart disease diagnosis		85.20%

The literature review presents an overview of the studies that investigated the use of machine learning in predicting and diagnosing heart disease. The table shows the author's data, the methodology, the output, and the result-accuracy of the analysis. The investigations utilized various sources, such as UCI Heart Disease, MIT-BIH Arrhythmia, and Cleveland Heart Disease. Some of the prominent algorithms used in the studies included Nave Bayes, Decision Tree, and Random Forest. The

accuracy of the studies varied, ranging from 84.81% up to 98% depending on the algorithm and data used. The findings of these studies suggest that machine learning can help predict heart disease.

Methodology

i. Data Preprocessing:

Data cleaning is a process that involves identifying and rectifying any errors or inaccuracies in a dataset. It can help predict a person's risk of developing heart disease.

a. Data Cleaning: In machine learning, data preprocessing involves preparing a dataset for analysis. This process involves selecting features, normalizing the data, and cleaning the data. Three methods are used to identify the risk factors for heart disease.

b. Missing Values Treatment: The missing values treatment is a process that involves imputing or identifying the missing values in a dataset. It can be done by using the mode, mean, or median feature.

c. Outlier Treatment: Outlier treatment is performed to identify and remove outliers from a dataset. These are the values that are significantly different from the rest. They can be detected using either the IQR or z-score methods.

ii. Feature Selection:

The process of feature selection is performed to identify the most relevant elements from the collected data. These elements contribute to the prediction of a target variable's risk factors.

a. Correlation Analysis: Correlation analysis is a process that involves identifying the link between the dependent and independent variables in a dataset. The selected features are then subjected to further analysis to find out if they have a high correlation.

iii. Data Normalization:

Scale the data to a value between zero and one in order to predict heart disease risk factors. There are two methods used for normalization.

a. Min-Max Scaling: A min-max scaling method is used to reduce the data to a value between zero and one. It takes into account the minimum and maximum values.

Decision Tree algorithm

The goal of the decision tree algorithm is to provide a supervised learning method for dealing with regression and classification problems. It constructs a model by creating a tree structure that represents various features and decision rules. The algorithm recursively divides the collected data into constituent groups according to their most crucial attributes to generate a decision tree. The objective of the algorithm is to construct a prediction engine that can handle diverse input sets.

i. Parameter Tuning

The table-2 shows the best parameters of the decision tree algorithm that was utilized in making a prediction about heart disease risk factors based on the Hawkins dataset. In machine learning, parameter tuning is a vital step that involves choosing the right values for the algorithm to achieve the best results. The goal of this study was to create a decision tree algorithm that can predict heart disease risk based on various risk markers. The algorithm's performance was evaluated using different metrics.

DOI : https://doi.org/10.56452/6-3-674

Table 2 Parameter tuning values

Parameter	Range	Best Value
max_depth	[1, 20]	4
min_samples_split	[2, 10]	2
min_samples_leaf	[1, 10]	3
criterion	["gini", "entropy"]	gini

The decision tree model was used to analyze the data set and determine the optimal split. It was able to achieve the best result by having a maximum depth of 4 and by taking into account the minimum number of samples needed to split an internal and leaf node. The parameters that were used in the analysis were also evaluated through grid search.

Results and Outputs

The result (as shown in figure 1,2,3 and table-3) of 98.8% accuracy indicates that the decision tree algorithm was successful in predicting heart disease risk factors using the Hawkins dataset. The precision score of 97.6% means that out of all the predicted cases of heart disease, 97.6% were true positive cases. The recall score of 98.1% indicates that the algorithm was able to correctly identify 98.1% of the actual positive cases. The F1-Score of 98.1% is the harmonic mean of precision and recall and represents an overall measure of the algorithm's accuracy in predicting both positive and negative cases. These high scores indicate that the decision tree algorithm performed very well in predicting heart disease risk factors and can be considered reliable for future use in healthcare settings.

i. Evaluation Metrices

Table 3 Evaluation metrices

Parameter	Values
Accuracy	98.8
Precision	97.6
Recall	98.1
F1-Score	98.1

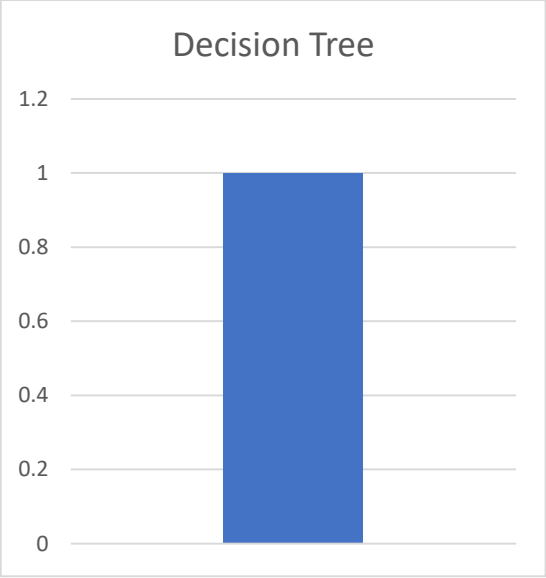

Figure 1 Graph representing various metrices

DOI : https://doi.org/10.56452/6-3-674

ii. Confusion Matrix

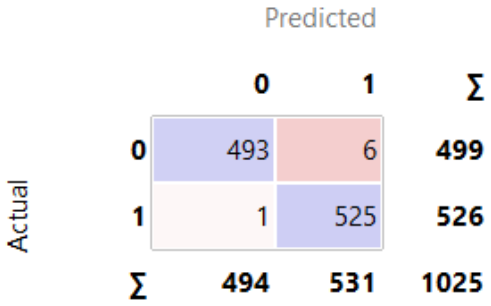

Figure 2 Confusion matrix- Decision Tree

iii. ROC Curve

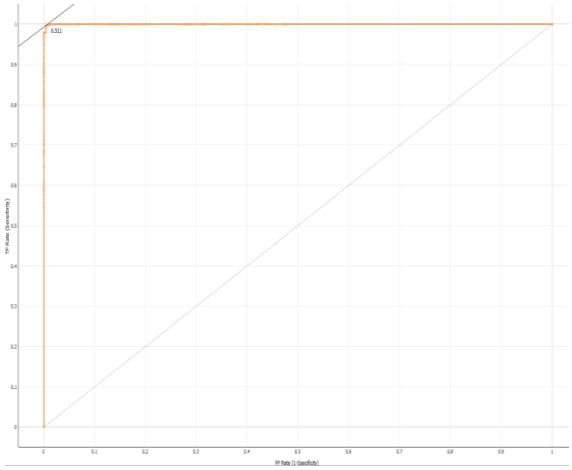

Figure 3 ROC curve for heart disease

Conclusion and Future scope

The study revealed that the decision tree algorithm was able to predict the likelihood of heart disease based on the Hawkins dataset. The findings support the idea that this method could be used to identify and prevent heart disease in its early stages. Although the algorithm was able to predict heart disease, it still has a long way to go before it can truly be used to identify and prevent the condition. For instance, it should incorporate other variables such as lifestyle factors and genetic markers. Further research is needed to improve its accuracy and provide a more accurate prediction. The results of the study have important implications for the healthcare industry and patients. It can help identify and prevent heart disease before it becomes a major issue, which can lower the cost of healthcare and improve the quality of life for patients. Using a decision tree algorithm to predict heart disease could also help develop personalized treatment plans.

References

[1] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribb. J. Sci. Technol.*, vol. 1, pp. 208–217, 2013.

[2] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Comput. Sci.*, vol. 120, pp. 588–593, 2017, doi: 10.1016/j.procs.2017.11.283.

[3] C. Krittanawong, H. J. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial Intelligence in Precision Cardiovascular Medicine," *J. Am. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, 2017, doi: 10.1016/j.jacc.2017.03.571.

[4] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 2, pp. 361–370, 2017, doi: 10.1093/jamia/ocw112.

[5] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. June, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[6] S. B. Patel, "Heart DIsease prediction using Machine learning and data mining Technique," no. October, 2016, doi: 10.090592/IJCSC.2016.018.

[7] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion

Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," *IEEE Access*, vol. 8, pp. 59247–59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

[8] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 24, no. 1, pp. 27–40, 2012, doi: 10.1016/j.jksuci.2011.09.002.

[9] A. Pattekari, S.A.; Parveen, "Prediction system for heart disease using Naïve Bayes," *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[10] C. J. Schneider, "Using mobile phones," *Pop. Cult. as Everyday Life*, vol. 19, no. 4, pp. 47–56, 2016, doi: 10.4324/9781315735481.

[11] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[12] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[13] S. J. Al'Aref *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *Eur. Heart J.*, vol. 40, no. 24, pp. 1975–1986, 2019, doi: 10.1093/eurheartj/ehy404.

[14] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[15] J. Chen, A. Valehi, and A. Razi, "Smart Heart Monitoring: Early Prediction of Heart Problems through Predictive Analysis of ECG Signals," *IEEE Access*, vol. 7, pp. 120831–120839, 2019, doi: 10.1109/ACCESS.2019.2937875.

[16] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185–196, 2020, doi: 10.1007/s12065-019-00327-1.