Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230082

Application and comparison of decision tree algorithm and K-Nearest Neighbors algorithm in heart disease prediction

Yiran Wang

School of Management, Hefei University of Technology, Hefei, Anhui, 230000, China

2020211627@mail.hfut.edu.cn

Abstract. In the past two decades, rapid industrialization and urbanization have led to tremendous economic growth and an improvement in people's living standards. However, the impact of people's irregular lifestyles and habits on their health has gradually emerged. Among them, cardiovascular diseases have become particularly prominent, with increasing incidence and mortality rates, especially in developing countries. Heart disease is a major cause of the rising death rates. Early-stage prediction of heart disease poses a major challenge in clinical analysis. Today, the adoption of appropriate decision support systems to achieve cost reduction in clinical trials has become a future development trend for many hospitals. This study compares decision tree classification and K-nearest neighbors (KNN) classification algorithms to seek better diagnostic performance for heart disease. The existing dataset of heart disease patients from the Cleveland database is used to te3st and demonstrate the performance of all algorithms, providing support for the establishment of a heart disease prediction system. This, in turn, can assist doctors in making more accurate diagnoses and timely interventions before the onset of heart disease, thereby reducing the mortality rate of heart disease from the source.

Keywords: Machine Learning, Heart Disease, Classification, Prediction, Decision Tree, K-Nearest Neighbors

1. Introduction

Predictive research on heart disease is crucial for improving diagnosis and preventing potential life-threatening events, as stated by Professor Jeremy Pearson from the British Heart Foundation. Thousands of families are devastated by heart disease each year, making it imperative for doctors to be able to make quick judgments through simple and efficient means. The heart is the engine of the body's functions and one of the most vital organs, providing the driving force for blood circulation and governing human life [1]. Once the heart encounters problems, the body cannot function properly. According to a report on "Top Ten Causes of Death" released by the World Health Organization in December 2020, heart disease is the leading cause of global mortality.

As a non-communicable disease, heart disease is difficult to identify in its early stages. For example, a significant portion of heart disease patients do not exhibit typical symptoms such as chest tightness, shortness of breath, general fatigue, and angina, a condition known as "silent heart disease" [2]. Many individuals fail to realize that their hearts are experiencing problems in a timely manner, resulting in missed opportunities for optimal treatment. Therefore, analyzing the causes of heart disease holds significant importance.

DOI: 10.54254/2755-2721/30/20230082

With the advancement of scientific technology, utilizing machine learning, artificial intelligence, big data, and other techniques to support decision-making in the medical field has become an inevitable trend [1]. Six popular machine learning classification techniques, including k-nearest neighbors, naive Bayes, decision trees, random forests, support vector machines and logistic regression, are available. Based on the aforementioned analysis, this study utilizes real data from heart disease patients in the Cleveland dataset, processes the data accordingly, and employs both k-nearest neighbors and decision tree classification methods. Analyzing 14 attributes, the study effectively combines machine learning and disease prediction to establish a heart disease prediction model, providing valuable support for medical decision-making.

2. Data And Features

2.1. Data description

The Cleveland database's 76 properties make up the dataset that was used. All published trials, however, only indicated the usage of the 14 traits. The original dataset has 303 rows and 14 columns, with each column representing a variable. Among the 14 variables, 13 are used as feature variables, and the last variable is used as the target variable, indicating whether a patient has heart disease [3]. Table 1 provides an explanation of the dataset's properties.

Table 1. Attributes meaning.

Serial Number	Variable	Data Type	Interval
1	age	int	[16,19]
2	sex	boolean	0 or 1
3	cp	int	1,2,3,4
4	trestbps	int	≥0
5	chol	float	≥0
6	fbs	boolean	0 or 1
7	restecg	int	0,1,2
8	thalach	int	≥0
9	exang	boolean	0 or 1
10	oldpeak	float	≥0
11	slope	int	1,2,3
12	ca	int	0,1,2,3
13	thal	int	3,6,7
14	target	int	0,1,2,3,4

In the dataset, the values sex=1 and sex=0 denote male and female, respectively. The values cp=1 and cp=2 correspond to usual angina chest pain, cp=3 to atypical angina chest pain, cp=4 to no symptoms of chest pain, respectively. Restecg=0 represents normal electrocardiogram, restecg=1 represents abnormal electrocardiogram, and restecg=2 represents left ventricular hypertrophy. Exang=1 denotes angina brought on by exertion, whereas exang=2 denotes its absence. When the target variable is equal to 0, heart disease is not present; when it is equal to 1, 2, 3, or 4, heart disease is present.

2.2. Data preprocessing

The original dataset comprises 303 rows and 14 columns, and the 'ca' column and 'thal' column both have 4 and 2 missing values, respectively. Common methods for handling missing values include algorithmic imputation, no treatment, and direct deletion. Since the original dataset has a small number of missing values, the direct deletion method is used here. After removing the missing values, the new dataset has 297 rows and 14 columns. The 'thal' indicator is replaced with 0, 1, 2, and the 'target' indicator is changed to exist (value 1) and not exist (value 0) for easier binary classification.

DOI: 10.54254/2755-2721/30/20230082

3. Model Building

3.1. Decision tree model creation

Decision tree is a basic classification and regression method. Its basic principle is to deduce a series of questions using if/else statements to ultimately make related decisions. Three components make up the majority of the decision tree algorithm: feature selection, tree formation, and tree pruning.

3.1.1. Feature selection. The objective of feature selection is to select features that can classify the training set. The key criteria for feature selection include Gini coefficient, information gain, and information gain ratio.

By calculating the Gini index for each potential split and choosing the one that minimizes the impurity the most, the Gini coefficient is used to identify the optimum split at each node in a decision tree. A lower Gini coefficient indicates a more homogeneous node and is preferred in decision tree algorithms [4]. Consequently, a decision tree model may be created to accomplish an appropriate data categorization impact, thereby lowering the system's level of disorder.

The following formula is used to compute the Gini coefficient [5]:

$$Gini(T) = 1 - \sum(p_i * p_i) \tag{1}$$

Information entropy is another classical measure of the disorder level of a system, in addition to the Gini coefficient. It can also be used to help reasonably divide nodes.

The information entropy is evaluated using the following formula:

$$H(X) = -\sum_{n=1}^{\infty} p_i \, log_2(p_i) \tag{2}$$

After classifying the data, the reduction in information entropy is called information gain. The larger the information gain, the lower the disorder level of the system after classification, indicating a better classification effect.

The information gain is calculated using the following formula:

$$Gain(A) = H(x) - H_A(x) \tag{3}$$

3.1.2. Tree generation. The process of generating a decision tree is shown in Figure 1, which involves recursively selecting the best features and partitioning the dataset [6].

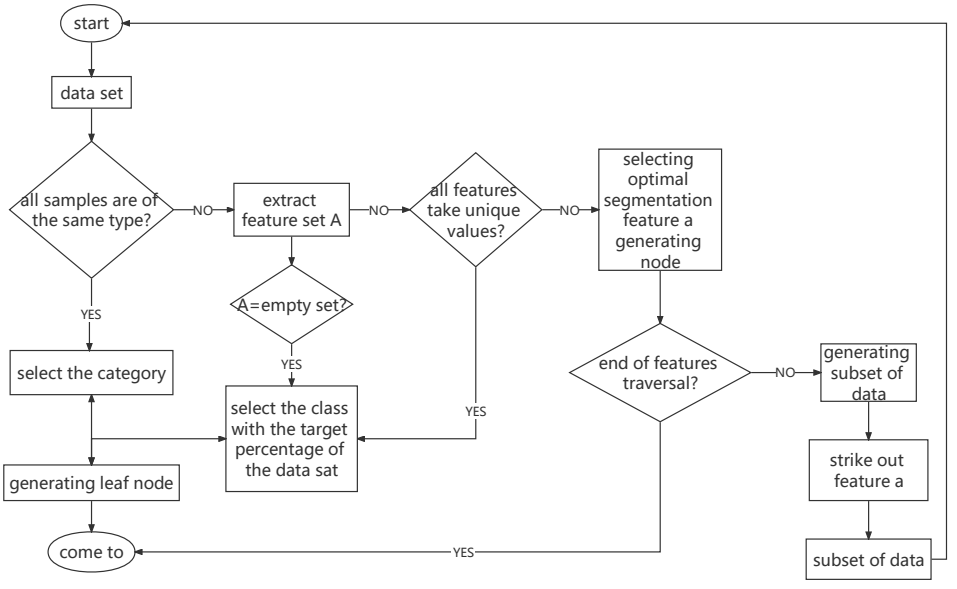

Figure 1. Flowchart for generating Decision Tree Model.

DOI: 10.54254/2755-2721/30/20230082

3.1.3. Tree pruning. The process of trimming the already generated tree is called tree pruning. Tree pruning aims to minimize the overall loss function or cost function of the decision tree [5].

3.2. K-Nearest Neighbors (KNN) model creation

K-Nearest Neighbors algorithm, abbreviated as KNN algorithm, is a classification algorithm suitable for binary classification problems.

3.2.1. Algorithm principle. KNN algorithm works by calculating the distances between an unknown sample and the labeled training samples to determine the nearest neighbors. Based on the class labels of the nearest neighbors, it either takes a majority vote or computes the average to predict the class or regression value of the unknown sample. The key aspects of the KNN algorithm include selecting an appropriate distance metric, determining the number of neighbors (K), and performing data preprocessing and parameter tuning to enhance accuracy and performance[7].

3.2.2. KNN algorithm generation. The three basic elements of the algorithm are the choice of K value, distance measurement and classification decision rule, all of which are depicted in Figure 2 [8].

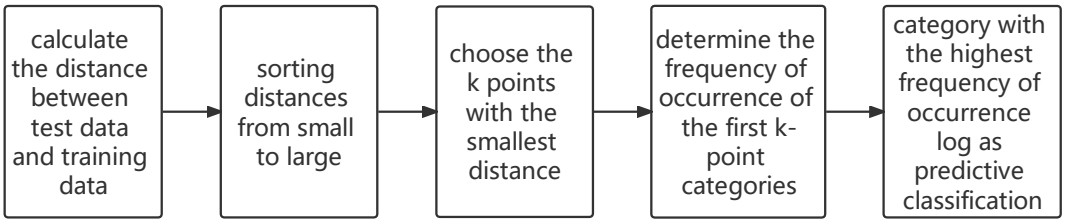

Figure 2. Flowchart for generating K-Nearest Neighbors Model.

4. Model Evaluation

4.1. Evaluation indicators

Confusion matrix, also known as error matrix, is one of the criteria to judge the classification degree [9]. When the classification results are binary, the confusion matrix is provided in Table 2 as shown.

Table 2. The confusion matrix for two classes.

	Positive Prediction	Negative Prediction
Positively labeled	True Positive(TP)	False Negative(FN)
Negative labeled	False Positive(FP)	True Negative(TN)

True Positive Rate (TPR) gauges the percentage of positive cases that the model properly identifies, sometimes referred to as sensitivity; The False Negative Rate (FNR) measures the percentage of genuine positive cases that the model wrongly classifies as negative; The False optimistic Rate (FPR) shows how frequently the model generates erroneous warnings or incorrectly optimistic predictions; The percentage of genuine negative occurrences (negative events that were accurately predicted) among all negative instances is known as the true negative rate (TNR), often referred to as specificity.

The proportion of accurate positive predictions to all of the model's positive predictions is known as precision; The percentage of positive samples that are accurately categorized as positive is used to determine recall; The fraction of accurate predictions—both true positives and true negatives—out of all possible instances is known as accuracy; And average Accuracy is calculated from the arithmetic average of accuracy for each class; The error rate is the percentage of cases that are wrongly categorised relative to the total number of occurrences.

DOI: 10.54254/2755-2721/30/20230082

A binary classification model's performance is assessed using the F1 score. It gives a fair evaluation of the model's accuracy by combining precision and recall into a single statistic. The harmonic mean of recall and accuracy is the definition of the F1 score. The scale goes from 0 to 1, with 1 being the highest attainable result and signifying flawless recall and precision. A higher F1 score suggests that the classification model is doing better overall[10].

4.2. Experimental analysis

4.2.1. Result comparsion. Train the established decision tree and KNN models. Since the 'target' field is the label column, the target column should be removed from the feature subset, and use cross-validation to train the model and evaluate its predictive classification performance. The dataset is divided into two parts using the sklearn.model_selection function train_test_split, including 30% for the test set and 70% for the training set. Use the training set to create 5 models, and five-fold cross-validation to assess each model's performance [8]. This involves calculating the cost function values for the cross-validation set using the five models and selecting the model with the smallest cost function value. Finally, use this selected model to calculate the cost function value for the test set.

The best values for max_depth and min_samples_leaf should be obtained by scanning the values of max_depth and min_samples_leaf throughout the construction of the decision tree model method and evaluating the parameters. Utilize the model's parameters after choosing the best feature to serve as the splitting feature.

During the creation of the KNN model algorithm, determine the range of K values and iterate through them. The K value with the highest classification accuracy should be chosen after comparing the classification accuracy of the various K values. The KNN algorithm model should then be constructed using the parameter values, and its performance should then be assessed using the dataset. The comparison of evaluation metrics after testing two models using the test dataset is presented in Table 3 and Figure 3.

Table 3. Comparison of model evaluation indicators.

Algorithmic model	Accuracy	Precision	Recall	F1_score
Decision tree	0.793	0.247	0.251	0.238
KNN	0.855	0.328	0.109	0.169

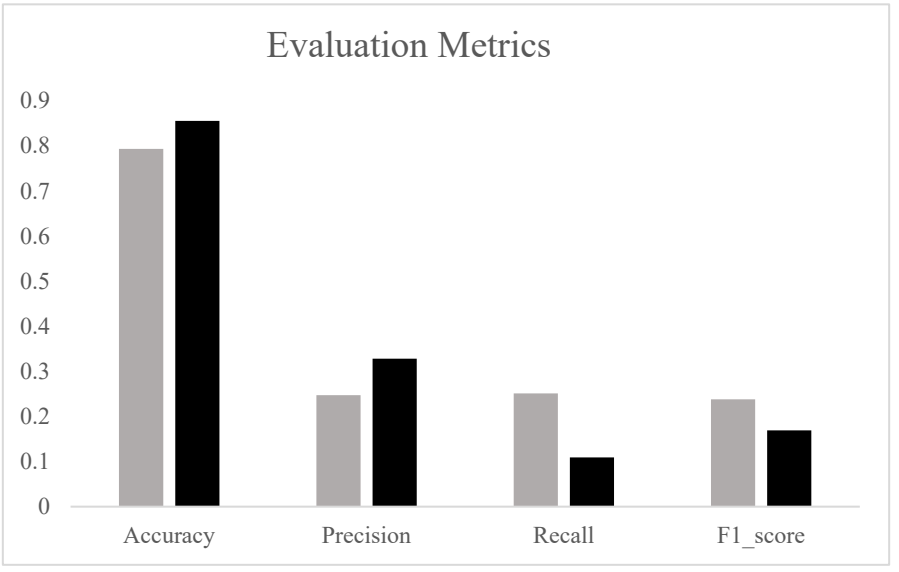

Figure 3. Visualization of comparison results.

DOI: 10.54254/2755-2721/30/20230082

4.2.2. Evaluation of the two model. The Decision Tree model algorithm has fast speed, low computational complexity, easily interpretable output results, and is suitable for high-dimensional data. It is not sensitive to missing intermediate values and can handle unrelated feature data. However, it may suffer from overfitting issues. Since the training data may contain noisy data, some nodes in the decision tree may have noisy data as splitting criteria, which can lead to the decision tree not accurately representing the true data.

The most straightforward and efficient approach for categorizing data is the K-Nearest Neighbors (KNN) model algorithm. It has high accuracy, is not sensitive to outliers, and has no assumptions about the input data [11]. However, it has the drawbacks of high computational complexity and high space complexity, resulting in longer calculation times.

In summary, the Decision Tree model algorithm has the advantages of fast computation and easy interpretation, but it can be prone to overfitting. On the other hand, the KNN model algorithm has high accuracy and is robust to outliers, but it has higher computational and space complexity.

5. Conclusion

Utilizing machine learning techniques that can be utilized to forecast cardiac disease is the major goal of this work. A prediction model for heart disease was constructed using the k-nearest neighbors and decision tree methods, and it was successfully applied to the dataset. The experimental results indicate that the k-nearest neighbors algorithm outperforms the decision tree algorithm in terms of classification prediction. After training on the Cleveland dataset, the k-nearest neighbors algorithm achieved an accuracy of 0.855 in predicting heart disease. Therefore, the k-nearest neighbors algorithm is more suitable for constructing a heart disease prediction diagnostic model. The results obtained in this study are quite satisfactory.

Moreover, the successful research on heart disease prediction can also be applied to other medical prediction domains, which is of significant importance for the development of intelligent healthcare driven by machine learning. As one of the major factors contributing to global mortality, cardiovascular disease has a high incidence and mortality rate. However, the risk factors and conditions associated with heart disease can vary among individuals. Heart disease prediction algorithms offer personalized risk assessments and treatment recommendations based on individual characteristics and data. This helps doctors develop more accurate treatment plans, including medication choices, dosages, and surgical approaches, thereby improving treatment outcomes and patients' quality of life. Moreover, heart disease prediction algorithms provide the capability for large-scale data analysis and risk assessments, offering crucial insights for national and societal health policy development. By accurately predicting the risk of heart disease, countries can implement targeted health screening and intervention measures, prioritizing healthcare services and resources for high-risk individuals. This maximizes resource utilization efficiency, reduces medical costs and treatment expenses, minimizes unnecessary waste, and ensures timely and appropriate care and attention for those who need it most.

Although satisfactory results have been obtained from model training, there are still many areas that need improvement to enhance the reliability of heart disease forecasting. Additionally, this study only constructed two heart disease prediction models using the k-nearest neighbors and decision tree algorithms. Further exploration is required to investigate the application of other classification prediction model algorithms in the field of heart disease prediction.

References

[1] Qin C C 2022 Research on heart disease prediction based on catboost model *Qufu Normal University* p 53 doi:10.27267/d.cnki.gqfsu.2022.001453.

[2] Wang X 2022 Study on machine learning based heart disease prediction model *Xinan Daxue* p 57 doi:10.27684/d.cnki.gxndx.2022.001659.

[3] Xin R H, Dong Z Y, Miao F B, Wang T T, Li Y R and Feng X 2022 Research on heart disease prediction model based on machine learning *Jilin Huagong Xueyuan Xuebao* **39(09)** p 27-32 doi:10.16039/j.cnki.cn22-1249.2022.09.006.

[4] Zheng L J and Song B 2023 Pre- Pruning and Optimization of Decision Tree Classification Algorithm *Zidonghua Yibiao* **44(05)** p 56-62 doi: 10.16086/j.cnki.issn1000-0380.202302 0066.
[5] Kotsiantis S B 2013 Decision trees:a recent overview *Artificial Intelligence Review* **39(4)** p 261-283 doi:10.1007/s10462-011-9272-4.
[6] Zhao X M, Wei X J, Wang N and Lei X J 2020 Feature Aggregation Decision Tree Prediction Model for Rainfall Landslide Disaster *J. Catastrophology* **35(01)** p 27-31
[7] Baidya A, Pasha A, Pavani B R, Paul A and Wali A 2020 Comparative Analysis of Multiple classifiers for Heart Disease Classification *International J. Advanced Research Comp. Sci.* **11(3)** p 6-11 doi:10.26483/ijarcs.v11i3.6523.
[8] Liang J H and Xu Y J 2022 Research on Predictive Diagnosis Model of Heart Disease Based on Machine Learning Algorithm *Modern Inf. Tech.* **6(19)** p 67-70 doi: 10.19850/j.cnki.2096-4706.2022.19.017.
[9] Tang Y F, Ke Y B, Zhuang L Y, Ji R D, Chen J Z and Yu K H 2023 Pipelines Ultrasonic Guided Wave Classification Based on Confusion Matrix Neural Network *Chinese J. Election Devices* **46(02)** p 469-477
[10] Chicco D and Jurman G 2020 The advantages of the Matthews correlation coefficient (MCC)over F1 score and accuracy in binary classification evaluation *BMC Genomics* **21(6)** p 4-5 doi:10.1186/s12864-019-6413-7.
[11] Xing W C and Bei Y L 2020 Medical Health Big Data Classification Based on KNN Classification Algorithm *IEEE Access* **8(86)** p 28808-28819 doi: 10.1109/access.2019.2955754.