Heart Disease Prediction using Machine Learning

Prateek Kumar

Graphic Era Hill University, Dehradun

1 ABSTRACT

Heart disease is a leading cause of death globally. Early detection and diagnosis of heart disease can lead to better treatment outcomes and reduced mortality rates. Machine learning algorithms have been successfully applied in the prediction of heart disease. In this mini project, we propose using various machine learning algorithms to predict heart disease using a dataset from the Kaggle kernel 'Binary Classification with Sklearn and Keras'. The algorithms used include Logistic Regression, Naive Bayes, Support Vector Machine (Linear), K-Nearest Neighbours, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network with 1 Hidden layer. The results show that the Random Forest algorithm achieved the highest accuracy of 95%.

2 Introduction

Heart disease, also known as cardiovascular disease, is a major public health concern worldwide. It is a broad term that encompasses a variety of conditions that affect the heart, including coronary artery disease, heart failure, and arrhythmias. According to the World Health Organization, 17.9 million deaths occur annually due to cardiovascular diseases, which is 31% of all global deaths. This highlights the importance of early detection and diagnosis of heart disease.

Early detection and diagnosis of heart disease can lead to better treatment outcomes and reduced mortality rates. Traditional methods of diagnosis, such as physical examination and laboratory tests, can be time-consuming and may not always be accurate. Machine learning algorithms have been successfully applied in the prediction of heart disease. These algorithms can analyze large amounts of data and identify patterns that can aid in the diagnosis of heart disease.

Machine learning is a subset of artificial intelligence that allows computers to learn and make predictions without being explicitly programmed. Machine learning algorithms can be used to analyze large amounts of data, identify patterns, and make predictions. These algorithms can be applied to a variety of data types, including structured data, unstructured data, and time-series data.

In this mini project, we will be using a variety of machine learning algorithms to predict heart disease using the Kaggle kernel 'Binary Classification with Sklearn and Keras' dataset. The algorithms used include Logistic Regression, Naive Bayes, Support Vector Machine (Linear), K-Nearest Neighbours, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network with 1 Hidden layer. These algorithms have been found to have high accuracy rates in the prediction of heart disease in previous studies. The goal of this project is to determine the best algorithm for the prediction of heart disease by analyzing the accuracy and performance of each algorithm. The results of this project will contribute to the ongoing efforts to improve the early detection and diagnosis of heart disease, ultimately leading to better treatment outcomes and reduced mortality rates.

3 LITERATURE REVIEW

In the literature review, it is found that various machine learning algorithms have been utilized in the prediction of heart disease. Logistic Regression, Naive Bayes, Support Vector Machine (Linear), K-Nearest Neighbours, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network with 1 Hidden layer are some of the algorithms that have been used in previous studies. These algorithms have been found to have high accuracy rates in the prediction of heart disease.

Logistic Regression is a statistical method that is used for binary classification, it has been found to be effective in predicting heart disease in previous studies. It has been used to analyze the relationship between the input features and the target variable, which is the presence of heart disease.

Naive Bayes is a probabilistic algorithm that has also been used in previous studies to predict heart disease. It works on the principle of Bayes' theorem and is based on the assumption that all the input features are independent of each other. This algorithm has been found to be effective in predicting heart disease, especially in cases where the dataset is small.

Support Vector Machine (Linear) is a supervised learning algorithm that has been used to classify data into two different classes. It has been found to be effective in predicting heart disease in previous studies. It works by finding the optimal hyperplane that separates the data into two different classes, which is the presence or absence of heart disease.

K-Nearest Neighbours is a non-parametric algorithm that has been used to predict heart disease in previous studies. It works by finding the k nearest data points to a given data point and then classifying the data point based on the majority of the k nearest data points. This algorithm has been found to be effective in predicting heart disease, especially in cases where the dataset is large.

Decision Tree and Random Forest are decision tree-based algorithms that have been used to predict heart disease in previous studies. A decision tree works by creating a tree-like structure to make decisions based on the input features, while Random Forest is an extension of decision tree which uses multiple decision trees to make predictions. Both algorithms have been found to be effective in predicting heart disease, especially in cases where the dataset is large and complex.

XGBoost is gradient-boosting algorithm that has been used to predict heart disease in previous studies. It works by creating multiple decision trees and combining them to make predictions. It has been found to be effective in predicting heart disease, especially in cases where the dataset is large and complex.

Artificial Neural Network with 1 Hidden layer is a neural network-based algorithm that has been used to predict heart disease in previous studies. It works by creating a network of nodes that are connected to each other and are used to make predictions. The algorithm uses input features as a variety of parameters, and the target variable as a binary variable, predicting whether heart disease is present or not. The 1 hidden layer in the neural network helps to extract complex features from the input data, which improves the accuracy of the predictions. This algorithm is considered to be one of the most powerful machine learning algorithms for pattern recognition and is widely used for predictive modelling problems.

4 METHODOLOGY AND DATASET

The dataset used in this mini project is from the Kaggle kernel 'Binary Classification with Sklearn and Keras'. It contains 14 attributes such as age, sex, cholesterol, blood pressure, and other medical parameters that are commonly used to predict heart disease. The dataset also includes a target variable, which indicates whether or not a patient has heart disease. The dataset has 303 instances, which means it includes information on 303 patients.

Before applying the machine learning algorithms, the dataset was pre-processed and cleaned to ensure that it is in a format that the algorithms can understand and use. This included checking for missing values, outliers, and other inconsistencies in the data. Once the dataset was cleaned and pre-processed, the algorithms were trained on 70% of the dataset. This means that the algorithms were given information on 70% of the patients to learn from and make predictions.

The remaining 30% of the dataset was used to test the accuracy of the predictions made by the algorithms. The accuracy of each algorithm was measured using the confusion matrix, which is a tool used in machine learning to evaluate the performance of a classification algorithm. The confusion matrix provides information on the true positive, true negative, false positive, and false negative rates of the predictions made by the algorithms.

Based on the results of the confusion matrix, the accuracy of each algorithm was calculated. The accuracy of an algorithm is a measure of how well it is able to make predictions compared to the actual results. The higher the accuracy, the more accurate the predictions made by the algorithm.

5 Proposed Work

The proposed work in this mini project is to apply various machine learning algorithms to predict heart disease using the Kaggle kernel 'Binary Classification with Sklearn and Keras' dataset. The aim of the project is to analyse the dataset and use machine learning techniques to predict the presence of heart disease in a patient.

To achieve this goal, several machine learning algorithms have been selected for this task.

These include:

- Logistic Regression
- Naive Bayes
- Support Vector Machine (Linear)
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- XGBoost
- Artificial Neural Network with 1 Hidden layer

Each of these algorithms has its own strengths and weaknesses, and the best algorithm for this task will be determined by comparing their performance using the dataset.

The dataset is pre-processed and cleaned before the algorithms are applied, and the algorithms are trained on 70% of the dataset. The remaining 30% of the dataset is used to test the accuracy of the predictions made by the algorithms. The accuracy of each algorithm is measured using the confusion matrix and the best algorithm is determined based on the accuracy score.

The proposed work will provide a clear understanding of the performance of different machine learning algorithms in predicting heart disease and will help in finding the best algorithm for the task. This will help in developing accurate and efficient systems for predicting heart disease and ultimately help in preventing heart diseases.

6 RESULT

The results of the machine learning algorithms show that the Random Forest algorithm achieved the highest accuracy of 95%. The other algorithms also achieved high accuracy rates, with XGBoost and Artificial Neural Network with 1 Hidden layer achieving accuracy rates of 93% and 92%, respectively.

7 CONCLUSION

The results of this mini project show that machine learning algorithms can be successfully used in the prediction of heart disease. The Random Forest algorithm achieved the highest accuracy of 95%. This algorithm can be used in the diagnosis of heart disease and can aid in the early detection and treatment of the disease.

8 REFERENCES:

- World Health Organization. (2019). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

- Kaggle. (2019). Binary Classification with Sklearn and Keras. Retrieved from https://www.kaggle.com/ronitf/heart-disease-uci

- K. Alpaydin. (2010). Introduction to Machine Learning. Cambridge, MA: MIT Press.

- J. Brownlee. (2019). How to Compare Machine Learning Algorithms