A Cardiovascular Disease Prediction Using Machine Learning Algorithms

Sara Omar, Nada Mohamed
Arab Academy for Science Technology & Maritime Transport, CCIT, Egypt
saraomar429@gmail.com,
nadaelmonateh2020@gmail.com
Supervisor: Nashwa El bendary
nashwa.elbendary@gmail.com

Abstract

Heart disease commonly occurring disease and is the major cause of sudden death nowadays. This disease attacks the persons instantly. Most of the people do not aware of the symptoms of heart disease. Timely attention and proper diagnosis of heart disease will reduce the mortality rate. Medical data mining is to explore hidden pattern from the data sets. Supervised algorithms are used for the early prediction of heart disease. Nearest Neighbor (KNN) is the widely used lazy classification algorithm. KNN is the most popular, effective and efficient algorithm used for pattern recognition. Medical data sets contain 14 features is obtained from UCI Machine Learning Repository. Feature subset selection is proposed to solve this problem. Feature selection will improve accuracy and reduces the running time. This paper investigates to apply KNN for prediction of heart disease. Experimental results show that the algorithm performs very well with 86% accuracy. This system also provides the relation between diabetes and how much it influences heart disease

Keywords: Medical data mining, Heart disease, KNN, Feature selection.

Introduction

The leading cause of death worldwide is the heart and blood vessels. It is a combination of various cardiovascular diseases such as heart disease, heart attack, stroke, heart failure, arrhythmia, heart valve problems, etc. High blood pressure, high cholesterol, diabetes and lack of physical activity are some of the main reasons behind the increased risk of developing this disease. By reducing behavioral risk factors such as smoking, an unhealthy diet, alcohol use and physical inactivity, this disease can be prevented.

If people can have prior knowledge of this disease before it moves to a more serious level, we can reduce the number of deaths and high-risk patients by a reasonable amount. With the help of development in machine learning and high computing power, they have been able to make doubly advances in artificial intelligence in the field of medicine, where people can use these techniques and come up with the basic model in the early stages.

In this document, a machine learning model is proposed and implemented to determine whether or not a person has this disease by focusing on factors such as factual information, medical test results, and patient-related information. The nearest neighbor classification algorithm which is well known and has good performance was used to implement this model.

Methodology

The cardiovascular disease prediction methodology was implemented using KNN ALGORITHM and the results were compared. Figure 1 describes the engineering scheme for predicting cardiovascular disease.

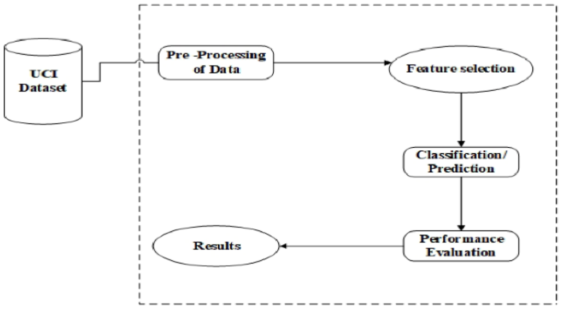


FIGURE 1. Experiment workflow with UCI dataset.

Algorithm Selection

K Nearest Neighbors is a simple yet incredibly functional algorithm that stores all available states and classifies the new data or state based on a similar scale. It is suggested that if the new point added to the sample is similar to the adjacent points, then that point will belong to a certain class of adjacent points. Generally, KNN ALGORITHM uses search applications where people search for similar items. K in KNN

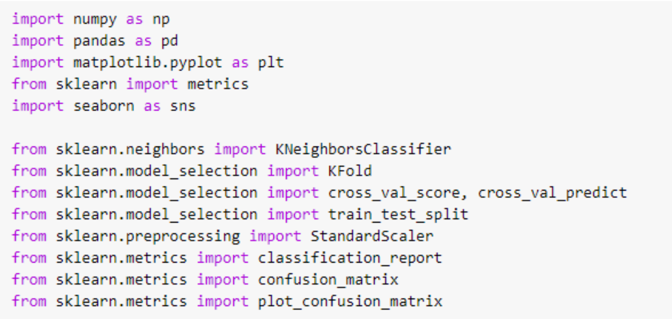
ALGORITHM indicates the number of nearest neighbors from the new point to be predicted.

Data collection

In order to predict whether a person has cardiovascular disease, a data set from KAGGALE.COM was selected. This dataset includes 14 features obtained from the UCI machine learning repository.

Model Implementation

As the first step all the required libraries for high-performance calculation, data visualization, and data model analyzation were imported as follows.



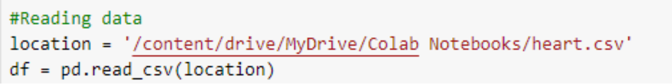
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import metrics
import seaborn as sns

from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score, cross_val_predict
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
```

Figure 2: Model implementation

Importing the dataset

An online dataset for cardiovascular disease was imported as a CSV file in order to do the analysis as follows.



```
#Reading data
location = '/content/drive/MyDrive/Colab Notebooks/heart.csv'
df = pd.read_csv(location)
```

Figure 3: import dataset

Data Visualization

A graphical representation has been conducted the number of patients who have heart disease:

So out of all the patients 165 patients actually have heart disease.

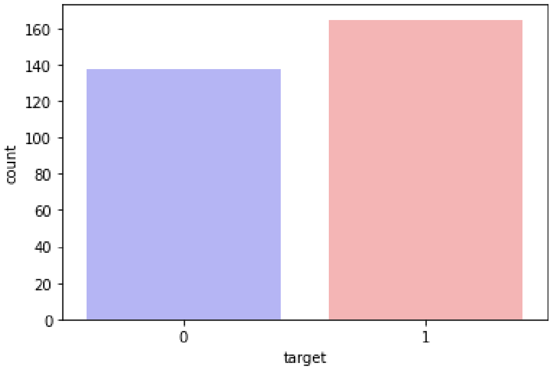


Figure 4: target show us if the person is suffering from heart disease or not.

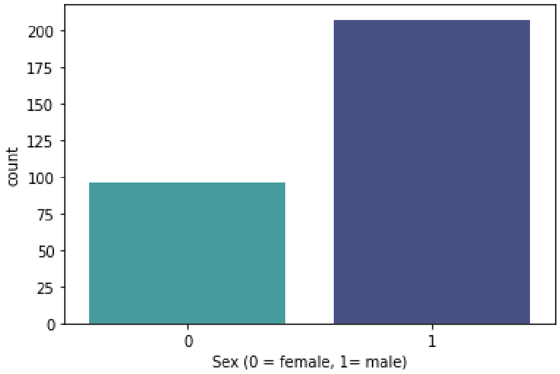


Figure 5: Classify target variable between male and female and visualize the result.

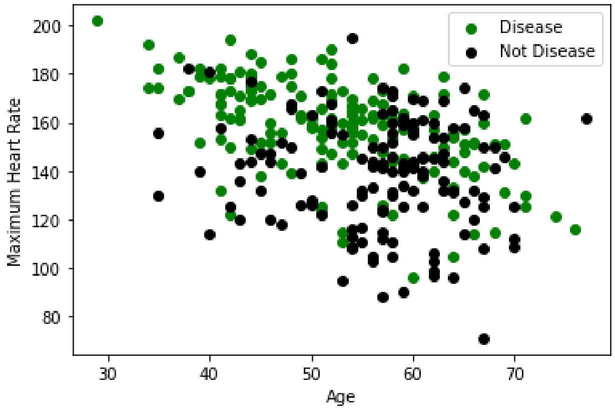


Figure 6: The relation between "Maximum Heart Rate" and "Age".

The maximum heart rate occurs in between age 50–60 years.

Algorithm

1. Load the heart disease dataset.

2. After Preprocess, Split the heart disease dataset into train and test data with the proportion of 80:20

3. K-Fold Cross Validation is wherever a given knowledge set is split into a K range of sections/folds wherever every fold is employed as a testing set at some purpose.

4. Train the model using train set.

5. Make predictions on the test fold.
6. Calculate the accuracy.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Where, TP- True Positive (prediction is yes, and they do have the disease.

TN-True Negative (prediction is no, and they don't have the disease.)

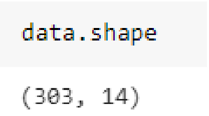
FP-False Positive (We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

FN-False Negative (We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

The accuracy obtained by using KNN algorithm is 82%

Experimental Results

To predict heart disease the dataset containing 303 instances is collected from UCI repository. Information about heart disease data set is shown in fig.



```
data.shape

(303, 14)
```

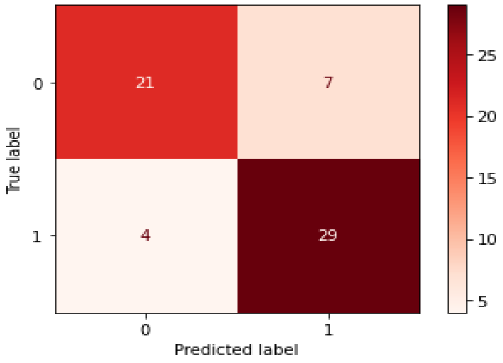
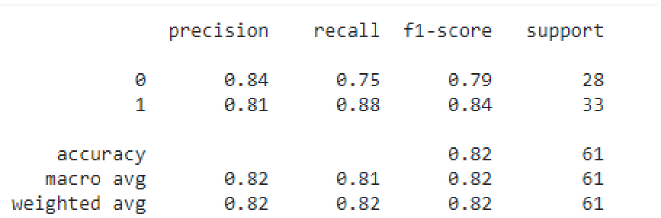


Figure 7: Confusion matrix for K=7

From the confusion matrix, it will give the classification report represented as in fig with an accuracy of 82. But this prediction has been done for K=7.

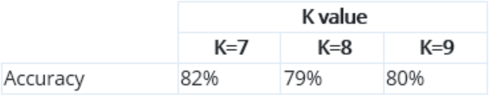


```
              precision    recall  f1-score   support

           0       0.84      0.75      0.79        28
           1       0.81      0.88      0.84        33

    accuracy                           0.82        61
   macro avg       0.82      0.81      0.82        61
weighted avg       0.82      0.82      0.82        61
```

Figure 8: Experimental Results



	K value		
	K=7	K=8	K=9
Accuracy	82%	79%	80%

Accuracy is maximum that is 82% when K=7

.

Conclusion

Manually determining the odds of cardiovascular disease based on risk factors can be hard. Using Machine learning techniques we can predict the outcome with the help of existing data. But still, we can’t trust the machine always. As you can see from this prediction, we got some percentage of “False positives and false negatives”. The only way to prevent cardiovascular disease is to stay healthy.

Reference

[1]https://www.semanticscholar.org/paper/Effective-Heart-Disease-Prediction-Using-Hybrid-Mohan-Thirumalai/2bc3644ce4de7fce5812c1455e056649a47c1bbf

[2]https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html

[3] Cardiovascular Disease Prediction Using KNN Algorithm | by Cibhi Baskar | Analytics Vidhya | Medium

[4] Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm - ScienceDirect

[5] Heart Disease Classification using KNN Algorithm | Kaggle