Improving the Accuracy of Prediction of Heart Disease Risk Coronary Using Feature Processing Techniques

Author:

Osama Anmar Sami Sami

Supervisor:

Dr.Yousef Elsheikh

Co-Supervisor:

Dr. Fadi Almasalha

This Thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

Applied Science Private University
Deanship of Scientific Research and Graduate Studies
Amman - Jordan 2020

Abstract

The heart is a small organ that is responsible for pumping oxygen-rich blood to the rest of the human body. The oxygen is also very important for the heart to continue his function, and any change in oxygen range affects the heart function, any change in heart function may affect the function of other organs. In coronary heart disease, the coronary artery that supplies the heart with oxygen becomes blockage or narrowing this means the oxygen reaching to the heart decreases or in sometimes cutting off. Coronary heart disease may be led to damage in a part of the heart muscle that the oxygen didn’t reach.

These days, people are working hard to live in a comfortable life. Because people care about their work as they eat any food, they forget to live in a healthy lifestyle, and they do not do any physical activity, and all of that led to an increased risk of coronary heart disease at the age of early. Predicting the possibility of coronary heart disease can help people change their lifestyle and eating habits to avoid coronary heart disease.

The prediction of coronary heart disease is a complicated process depending on many factors, such as in diabetic patients the risk of diagnose with coronary heart disease increase even if they are young more than non-diabetic patients are. Sometimes many factors together increase the risk very much, and each factor alone increases the risk too, therefore the effects of each factor need to understand, and the effect of each factor with other factors also needs to understand. Machine learning solves the complicated problem of predict coronary heart disease.

Rapidminer is an open-source tool for data science developed in 2006 by a rapidminer company. Rapidminer includes many libraries for machine learning, deep learning, and features processing.

Framingham heart study dataset is the first study for cardiovascular disease such as (coronary heart disease, hypertension, and many other cardiovascular disease) was done by the National Heart, Lung, and Blood Institute. The dataset contains features for many types of possible risk factors for cardiovascular disease such as (demographic features, behavioral features, medical history features, medical examination features, and Laboratory Features).

In this thesis, we are going to improve the accuracy of prediction of coronary heart disease using machine learning techniques (Decision Tree, Naïve Bayes, Random Forest, KNN, and Neural Network) and feature processing (Impute Missing Values, Feature Normalization, Feature Standardization, Encoding Categorical Features, and Discretization)

The results obtained from this thesis were very encouraging compared to other studies that used the same data set. For example, the accuracy of prediction obtained for decision tree was 91.39%,for random forest was 92.80%, multilayer perceptron was 92.64%, for k-nearest neighbors was 92.74%, and for naïve bayes was 92.74