Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
ISSUE (29), February (2020)
ISSN: 2616-9185
www.mecsj.com

Analysis Heart Disease Using Machine Learning

Mashael S. Maashi (PhD.)
Software Engineering Department
Collage of Computer and Information Sciences
King Saud University, Saudi Arabia
MMaashi@ksu.edu.sa

Abstract:

The heart diseases have great importance in the body, as it performs many basic functions, and any defect in it harms the entire body and leads to complications and great risks that may reach death if it is not discovered and treated early, and the kidneys may be exposed to many health problems and diseases, which must be paid attention for it, therefore, it is necessary to constantly and continuously detect to predict a possible disease.

There are many machine learning tools that help in predicting the presence of heart diseases or not, but in this study we will rely on Orange to predict correctly. This study aimed to predict heart disease using data extraction techniques, an analytical approach was used to reach the results. The Cleveland database and Stat log Heart Data Sets from UCI Heart Disease Dataset were used to search and download the dataset related to heart diseases, then the author performed the preprocessing on data set, after that the data mining techniques were applied using two algorithms which are Support K Nearest Neighbor (KNN), and Naïve Bayes (NB).

These techniques were applied using Orange. We found the best result of accuracy was provided by KNN algorithm which gave 88% accuracy while the other algorithm Naïve Bayes classifier gave 85% accuracy.

Keywords: Heart Disease, Machine Learning, Orang, K Nearest Neighbor (KNN), Naïve Bayes (NB).

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
ISSUE (29), February (2020)
ISSN: 2616-9185
www.mecsj.com

الملخص

لأمراض القلب أهمية كبيرة في الجسم، حيث يؤدي العديد من الوظائف الأساسية، وأي عيب فيه يضر الجسم كله ويؤدي إلى مضاعفات ومخاطر كبيرة قد تصل إلى الموت إذا لم يتم اكتشافه وعلاجه في وقت مبكر، وقد تصاب الكليتان بالعديد من المشكلات الصحية والأمراض التي يجب الانتباه إليها، لذلك من الضروري أن نكشف باستمرار وبشكل مستمر للتنبؤ بمرض محتمل.

هناك العديد من أدوات التعلم الآلية التي تساعد في التنبؤ بوجود أمراض القلب أو لا، ولكن في هذه الدراسة سوف نعتمد على Orange للتنبؤ بشكل صحيح. تهدف هذه الدراسة إلى التنبؤ بأمراض القلب باستخدام تقنيات استخراج البيانات، وقد استخدم المنهج التحليلي للوصول إلى النتائج.

تم استخدام قاعدة بيانات Cleveland و Stat log Heart Data Sets من مجموعة بيانات أمراض القلب UCI للبحث وتنزيل مجموعة البيانات المتعلقة بأمراض القلب، ثم قام المؤلف بإجراء المعالجة المسبقة على مجموعة البيانات ، بعد أن تم تطبيق تقنيات استخراج البيانات باستخدام خوارزمية هما K Nearest Neighbor (KNN)، Support و Naïve Bayes (NB).

تم تطبيق هذه التقنيات باستخدام Orange. وجدنا أن أفضل نتيجة للدقة كانت بواسطة خوارزمية KNN والتي أعطت دقة 88٪ بينما أعطت المصنفات الأخرى Naïve Bayes دقة 85٪.

الكلمات المفتاحية: أمراض القلب ، التعلم الآلي ، Orange ، K Nearest Neighbor (KNN), Naïve Bayes (NB).

1- Introduction:

Nowadays, with the increased data volume in the world, we are overwhelmed with data. Machine learning is one of the important fields that used to extract the knowledge and valuable information from the data sets using many of techniques such as clustering, association, and classification. It's used in a different field such as financial data analysis, medical data analysis, market analysis, and social network analysis, etc.

One of the most important domains using machine learning is the healthcare industry. Daily, healthcare industry generates a large amount of data about patients, disease, hospital, medical equipment, and treatment cost, etc. Data mining helps physicians in making the appropriate decision for treatment and prediction of disease in early stages which help in preventing disease or reduce its effects, such as Heart disease, Cancer disease, and Chronic Kidney Disease, etc.

The heart is one of the most important organs of the body and it is responsible for pumping blood to the rest of the body, there are a number of diseases that affect the heart muscle and affect it, including heart artery diseases, blood vessels, heart rhythm disorders, cardiomyopathy, heart infection, congenital heart defects, and heart valve disease. Due to the close association between heart health and vascular health in the body, these diseases are also referred to as "cardiovascular diseases." There are many reasons that lead to heart disease, including birth defects in the heart, obesity and fat accumulation, smoking, which is one of the most common habits, diabetes, excessive consumption of alcohol and caffeine, exposure to stress, excessive intake of foods that contain a proportion great fats and oils.

Moreover, heart disease is a common disease that is not limited to a specific age group, which requires more attention due to the high death rate in previous years. Moreover, heart disease is a common disease that is not limited to a specific age group, which requires more attention due to the high death rate in previous years. Globally, about 80% of deaths occurred in decline Middle-income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million person with heart disease are expected. (Devika, Avilala & Subramaniyaswamy, 2019).

This study came with the aim of using machine learning in order to analyze heart diseases, where orange learning machine tool was used with different algorithms.

2- Problem Statement:

Heart diseases are used to express any of the various diseases and health problems that may affect the heart, and are represented by vascular diseases such as coronary artery disease, heart rhythm disorders, and congenital heart defects. Heart diseases affect blood vessels, such as narrowing of blood vessels, which may result in heart attack, angina, or stroke, and other health problems that affect the heart as well.

However, the symptoms differ from one person to another based on several variables, including sex, age, weight, and many variables, as these variables require a great effort and time for doctors to analyze and study them, so it is necessary to use a machine learning tool that works to analyze these variables with high accuracy, to reach accurate results and close to the results of the doctors.

3- Research Significant:

Heart diseases are among the most dangerous and common causes of death in our modern world. Therefore, it is necessary to study the years that affect heart disease periodically and repeatedly in order to early detection and treatment of disease states, prevent complications of disease states, discover physical disabilities, work on their rehabilitation, work on correcting physical defects, and work to raise the health level of the individual and society.

The importance of this research lies in studying the importance of factors that affect the incidence of heart disease, and access to a useful tool that helps doctors to analyze heart disease to reduce the time and effort spent on examining patients, as the tool is a reference for them in their work.

4- Research Question:

1- What are the factors that affect heart disease?
2- What is the appropriate machine learning tool for the analysis of heart disease?
3- What algorithms are used in the machine learning tool for analyzing heart disease?
4- What are the best algorithms in learning a machine that helps analyze heart disease?

5- Research objectives:

1- Identify the factors that affect heart disease.
2- Identify the suitable machine learning tool for heart disease analysis.
3- Identify the algorithms used in the machine learning tool to analyze heart disease.
4- Identify the best algorithms in machine learning that help analyze heart disease.

6- Background:

In this research, one of the machine learning tools, Orange, will be used.

Orange: It is a visual software programming component for explorative data analysis, visualization, data mining and machine learning, its open source and written in python programming language which is developed by "Bioinformatics Laboratory of the Faculty of Computer and Information Science at University of Ljubljana"(Demšar & Zupan, 2013).

K-Nearest Neighbor (KNN): is among the simplest of all machine learning algorithms. It is the non- parametric method used for classification and prediction. It can use to give weight to the contributions of the neighbors, so the nearer neighbors contribute more to the average than more distant ones (Alex & Shaji, 2019).

Naïve Bayes: a classifier calculated the probability of a given dataset to perform classification. Each attribute in data is independent of others. The highest probability of class is the output class (Bashir et al, 2019).

7- Literature reviews:

7.1 Machine learning in medical field:

Khaleel, Pradham, & Dash, (2013) noted about there is an increase in demand for data mining techniques in the medical field, whether it by cluster, classification, or registration, as it is characterized by high accuracy in prediction. This study came to analyze the data mining technique to find Locally Frequent Diseases such as cancer disease and heart disease, the technical evaluation according to four specifications, namely accuracy, speed, cost, and performance, as the evaluation was according to the conventional method. The study also concluded that the Data Mining Technique is a distinguished higher utility. Moreover, the Data Mining Technique helps in taking the appropriate decision in the medical field.

According to (Kaur& Wasan, 2006) and (Kumar, N., & Kumar, S, 2018) the Diabetes diseases are one of the most prevalent diseases of this age, and its symptoms are related to each other in a hidden and unclear relation, and many tools cannot be discovered, as this problem was solved through the use of data mining technique.

The study of Kaur& Wasan (2006) aimed to use Data Mining techniques in the medical field, including ANN, KDD, Rule Based, and Decision Tree, it was applied to diabetes-specific data set and then analyzing the entire attribute alone and studying its relationship with diabetes based on the data mining technique. They found that Data Mining is a highly predictable technique in the early stages.

While the study of Kumar, N., & Kumar, S. (2018) aimed to diagnose diabetes by relying on cluster and classification. The study was conducted on 650 patients, the enamel was used by applying it to 3 Cluster, and every cluster was divided depending on the level of risk (mid, moderate, and severe), to find performance, and to calculate the performance based on accuracy and error rate. The study result showed that the C4.5 tree was the highest accuracy obtained which equal 100%.

We noted that in both studies the data mining techniques were used to predict for Diabetes diseases, but this study relied on data mining techniques to predict kidney disease.

Alonso, S. G., et al, (2018) noted about the field of mental health needs of a deep understanding as the factors affecting it are neutralized. This study aims to analyzing and reviewing previous studies to determine what technologies are used and what are the most prevalent diseases, they were limited to studies from 2008-2018, as mental health has been linked with diseases such as, Alzheimer's, dementia, depression, and schizophrenia. The studies have shown that the most accurate tools for determining the factors influencing are Data Mining techniques where they give high-precision results with a high ability to predict a possible disease.

7.2 Machine learning in heart disease:

By 2030, it is expected that there will be 23 million people with heart disease, so we need techniques to help us to study and predict this disease, they use the data mining technique in heart diseases. Soni et al. (2011) pointed for prediction of the diagnosis of heart diseases, where they relied on five techniques for Data Mining, which are decision tree, Bayesian, KNN, Neural Networks, and classification based on clustering.

Hence, this technique was applied to the same data set, consisting of 10 instances, they concluded by the technique that got the highest accuracy in results is decision tree, which was 99.2%, then Bayesian was 96.5, and then KNN, Neural Networks, and classification respectively. Also, when applying the genetic algorithm to reduce the size of the dataset, it was found that the results for the decision tree and Bayesian got the highest percentage of accuracy with the least time to obtain the results.

Ayatollahi (2019) mentioned about the forced on finding positive predictive value (PPV) through the comparison between Support Vector Machine and Artificial neural network, this study was done by collecting data at AJA University during the years 2016 and 2017 and this data set was formed from 1324 record and 25 variables, after collecting data, the use of the statistical program for analyzing it, after that applying the techniques, it was found that the SVM had higher accuracy, Precision, power, and performance than Artificial neural network for positive predictive value (PPV).

The study of Tarawneh & Embarak (2019) derived new approach relying on data mining techniques, able to classify disease, as this proposal works to combine more than one techniques, it called "Hybridization", is based on combined between ANN, Naïve Bayes, SVM, KNN, and Tree J48, with each other, the proposed approaches was applied on Cleveland heart disease data set and it consisted of 14 features and 303 records when applying each Algorithm alone, the Naïve Bayes and SVM were more accurate, but when applying the proposal got more accuracy in results, thus merging is more than a better technique than relying on a technique alone. We noted in previous studies, the data mining techniques were used to predict heart disease, but our studies relied on it to predict chronic kidney disease.

Chen et al. (2017) built a prediction model for Coronary Heart Disease (CHD) instead of Coronary Angiography (CAG), as well as established a more simple and noninvasive method for the diagnosis and treatment of a large number of patients with CHD. First, their dataset contained 599 cases of CHD patients and 398 healthy people over 60 years old, then in dimensionality reduction they used K-means cluster analysis and main component analysis, after that in prediction phase they used Back Propagation (BP) neural network method and entered for it 25 kinds of independent variables were as input, and the output was a type of CHD.

The result of using K-means cluster analysis and main component analysis was selected total 25 variables such as basic information (e.g. sex, BMI, heart rate), symptoms (e.g. angina, lifestyle) and comorbidities (e.g. stroke, diabetes). The ratio of the control group and the experimental group was 1: 1.5, where unstable angina accounted for 32.2%, myocardial infarction accounted for 14.9%, stable angina accounted for 29.8%, ischemic cardiomyopathy accounted for 7.5% and asymptomatic myocardial ischemia accounted for 15.6%.

The BP neural networks achieved 93.4% accuracy for predict 5 types of CHD and propose 8 main types of health intervention solutions.

Deekshatulu et al. (2013) proposed a new algorithm which combines KNN with genetic algorithm for effective classification for heart disease. They used genetic algorithms to prune redundant and irrelevant attributes as well as to rank more attribute which give more best classification, First , they built classifier and measured the accuracy , also they built classifier combined KNN with genetic algorithm. They used seven datasets from UCI repository include various types of heart diseases such as coronary heart disease, heart failure and inflammatory heart disease also common risk factors of heart disease like gender, diabetes…etc. They entered the parameter of genetic algorithm crossover probability=0.6, mutation probability=0.033, maximum generations=20, population size=20 report frequency=20, seed =1 also for KNN algorithm they entered K=1,2,.N, cross validate=true, debug=True, distance, weighting=weight by 1 distance, mean squared=True, no normalization=false. From the results they observed that integrating GA with KNN outperforms the other methods with greater accuracy. And their prediction model helps the doctors inefficient heart disease diagnosis process with fewer attributes.

8- The Methodology:

This methodology was applied to reach for the results of this research:

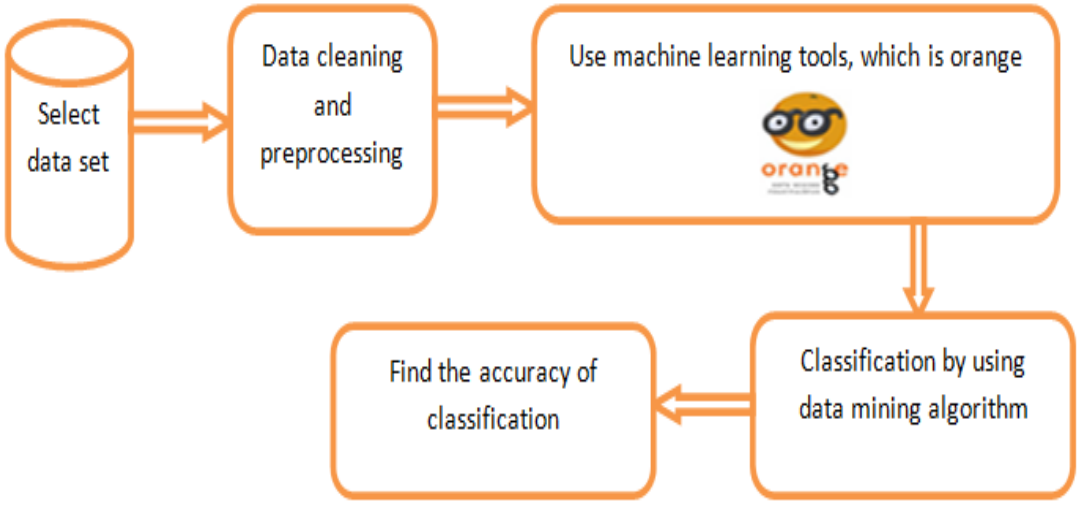

Figure 1: The methodology

8.1 Data set:

We will search and download the dataset related to heart diseases from the GitHub, the Cleveland database and Stat log Heart Data Sets from GitHub Heart Disease Dataset, in the following table we describe the dataset that was used.

Table 1: Dataset Describtion

ID	Attributes Name	Description
1	Age	Age of the individual in year
2	Gender	The gender of the individual
3	Cp	chest pain type
4	Thest bps	Sleeping Blood Force
5	Chloe	Serum Cholesterol
6	Rest Ecg	Resting ECG
7	Fbs	Fasting Blood Sugar
8	Thalach	Greatest heart speed reached
9	Ex ang	Exercise induced angina
10	Old peak	ST despair induced by apply virtual to relax
11	Slope	Slope of the height effect ST Section
12	Ca	Number of key vessels painted by fluoroscopy
13	Thal	Desert Category
14	Class label	Diagnosis of heart disease

8.2 Data cleaning and preprocessing:

Where the data has been processed and cleaned through two basic steps, namely replacing the null values (missing) with the mean for them, and the data has been transferred from string to integer value such as : Normal= 1, Up Normal= 0.

8.3 Using orange:

Orange "is characterized by its simplicity and creative graphical interface, which requires limited knowledge of data mining, and compared to other data mining tools, its strength is the interactive display function that allows beneficiaries to place display marks and then choose data points or nodes directly from the graphs". (Mikut and Reischl, 2011).

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
ISSUE (29), February (2020)
ISSN: 2616-9185
www.mecsj.com

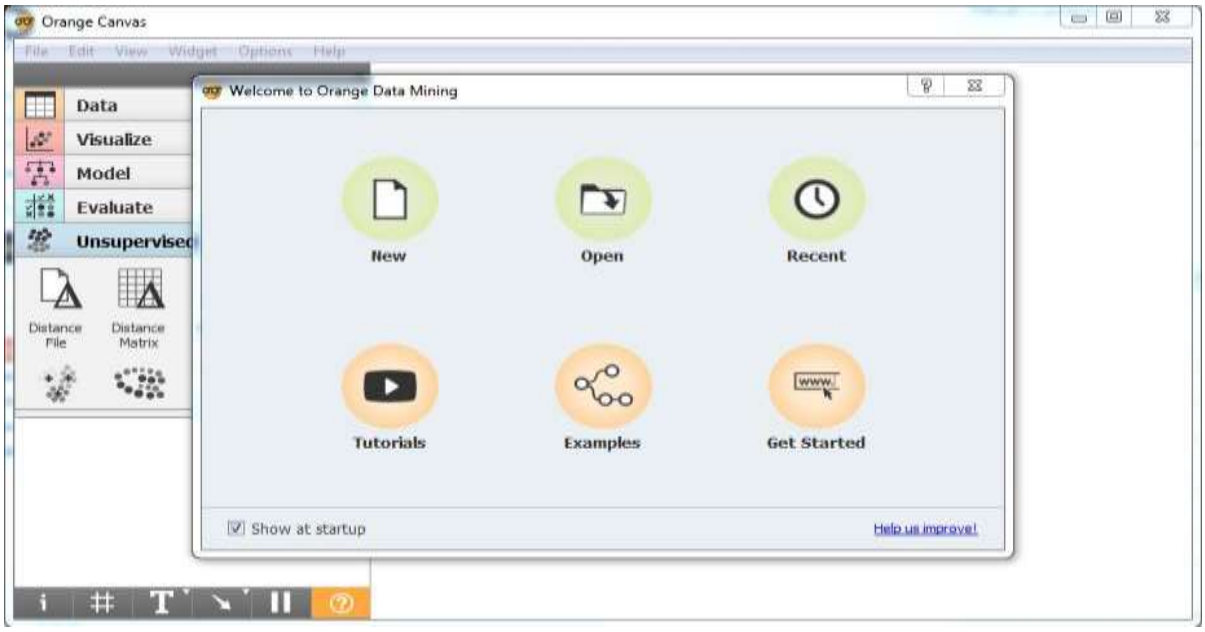

Figure 2: Orange Home page

8.4 Classification by using data mining algorithm:

We use two of data mining algorithms in Orange, which are Naïve bayez NB, and K nearest Neighbor KNN

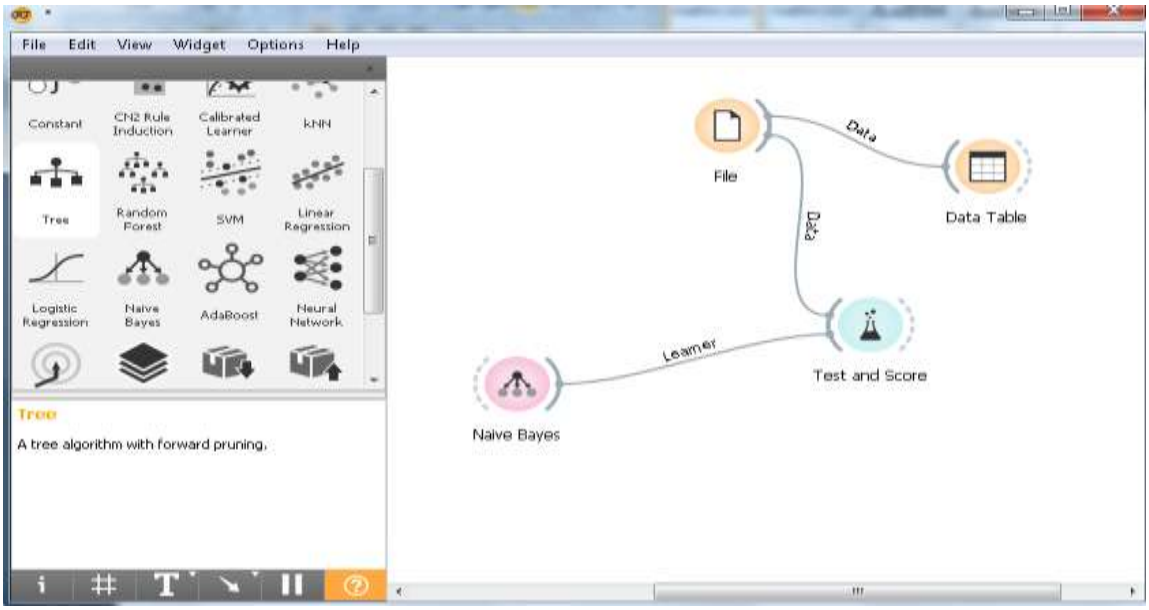

Figure 3: Apply NB classifier

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
ISSUE (29), February (2020)
ISSN: 2616-9185
www.mecsj.com

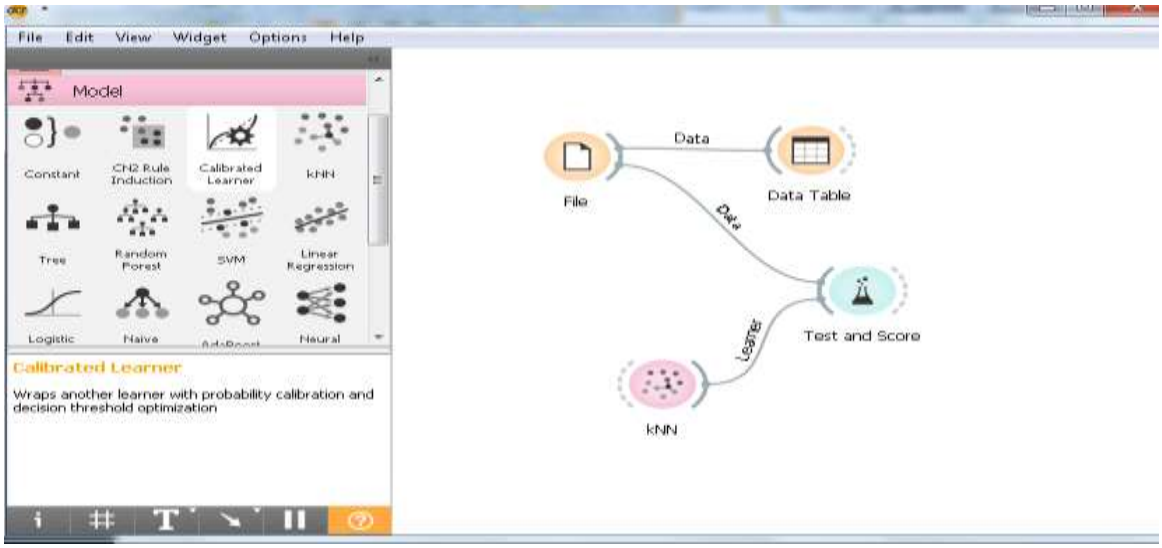

Figure 4: Apply KNN classifier

8.5 Find the accuracy of result

Based on the orange tool, we find this result:

Table 2: Accuracy of classifier

Classifiers type	Orange tool
Naïve Bayes classifier	**0.85**
KNN classifier	**0.88**

Based on table2 describes the accuracy for each classifier in orange tools, and then the result showed, there is no high difference between accuracy, we find the best result of accuracy is KNN is given 88% and another algorithm is a Naïve Bayes classifier which is given 85%.

9- Conclusion:

The machine learning techniques used in many fields, in this research, we use the Orange in the medical file for prediction the diseases, so the main objectives of this research is predicting the heart diseases based on machine learning, we used K Nearest Neighbor, and Naïve Bayes, we applied these techniques on Orange, WEKA tools.

After collecting the data set from Git Hub and cleaning it, we predicted the heart using these techniques to find the accuracy for each technique; we concluded the highest accuracy is KNN which gave 88% accuracy while the other algorithm Naïve Bayes classifier gave 85% accuracy.

References:

1. Alex and Shaji, S. P. (2019, April). Predictionand Diagnosis of Heart Disease Patients using Data Mining Technique. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0848-0852). IEEE.
2. Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data mining algorithms and techniques in mental health: A systematic review. *Journal of medical systems*, *42*(9), 161.
3. Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC public health*, *19*(1), 448.
4. Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving Heart Disease Prediction Using Feature Selection Approaches. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 619-623). IEEE.
5. Chen, Q., Han, L., & Guo, S. (2017). GW28-e0435 Prediction and intervention of coronary heart disease based on data mining. *Journal of the American College of Cardiology*, *70*(16 Supplement), C76.
6. Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia Technology*, *10*, 85-94.
7. Demšar, J., & Zupan, B. (2013). Orange: Data mining fruitful and fun-a historical perspective. Informatica, 37(1).

8. Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 679-684). IEEE.
9. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 Visited: 10/2/2020.
10. Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer science*, *2*(2), 194-200.
11. Khaleel, M. A., Pradham, S. K., & Dash, G. N. (2013). A survey of data mining techniques on medical data for finding locally frequent diseases. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(8).
12. Mikut, R., Reischl, M. (2011). Data Mining Tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (5): 431–445.
13. Soni, R., Bhuyar, G., Jain, S., & Shah, H. (2011). Prevalence of uropathogens in various age groups and their resistance patterns in a tertiary care hospital in central India. *NJIRM*, *2*(4), 7-10.
14. Tarawneh, M., & Embarak, O. (2019, February). Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. In *International Conference on Emerging Internetworking, Data & Web Technologies* (pp. 447-454). Springer, Cham.
15. Yadav, D. K., Kumar, S., Misra, S., Yadav, L., Teli, M., Sharma, P. & Kim, M. H. (2018). Molecular insights into the interaction of rons and thieno [3, 2-c] pyran analogs with SIRT6/COX-2: a molecular dynamics study. *Scientific reports*, *8*(1), 1-16.