Title: Prediction of Heart Disease using Data Mining and Machine Learning Algorithms

Abstract: Nowadays health diseases are increasing day by day due to various reasons i.e. our life styles, smoking, drinking [1]. Peoples of all over the world are suffering from various health diseases and heart attack is one of them. Nowadays heart disease is the biggest cause of death in the worldwide. This research paper focuses on developing a system that can predict the possibilities of heart attack disease with the help of data mining and machine learning algorithms. The outcome of this system provides the possibilities rate of heart disease in percentage. The datasets that are used in the system are classified in the terms of medical parameters. The datasets are processed in Python Programming using two Machine Learning (ML) Algorithms namely K-Nearest Neighbors (KNN) Algorithm and Random Forest Algorithm which shows the best accuracy level of heart disease prediction.

Keywords: Heart diseases, machine learning algorithms, python programming, data mining.

I. **Introduction**: Heart disease is now the leading cause of death worldwide. Every year many peoples die cause of coronary artery disease or heart attack. Blood pressure, pulse rate, and cholesterol are the main reason for heart disease. Family history, life style, high blood pressure, poor diet, smoking and drinking alcohol are also the risk factor for heart disease. Smoking is the reason for 40% of death of heart disease [1]. Nowadays all the cases of hospitals are recorded on computers. So, it becomes easy to collect data for medical issue. As we have huge resource of medical data, we can use it easily for researching

on medical science. For processing medical data certain techniques are used and Data Mining is one of them.

Data mining is the process extracting information and knowledge from large dataset. Data mining is mainly used to discover or extract the hidden information from a large dataset. It is also called Knowledge Discovery Database (KDD). Data mining is mainly essential in various fields like medical field, business field, educational field to extract the useful information from large amount of database. Data mining techniques can be implemented through machine leaning algorithms and python programming. In this project, a heart disease data set is used to develop a system that predict heart disease. The aim of this project is to predict the possibilities of occurring heart disease rate in percentage.

II. **LITERATURE SURVEY** [1] [2]**:** This paper predicts heart disease for both male and female patients using ML Algorithms and data mining classification techniques. Different types of studies have been done due to focus on prediction of heart disease more accurately. Various data mining techniques are used in this paper to diagnosis and achieved different accuracy level for different methods to predict. The detailed information about heart disease, its common types and main risk factors have been described in this paper. KNN algorithm and random forest algorithm are main data mining techniques and through this techniques heart disease is predicted in this system.

In this paper, the information about data mining and heart disease has been gathered. Description of the dataset are also given in this paper. Important things that are related with heart disease have been described in this paper.

III. HEART DISEASE [2] [3]: A heart is the most important organ of human body. If the heart does not perform its operation properly, it influences the other organ of the body i.e. kidney, brain etc. According to WHO (World Health Organization), heart disease or cardiovascular diseases (CVDs) is the number one cause of death globally. About 17.9 million people die each year from heart attack or CVDs which is an estimated 31% of all death worldwide. According to WHO Individuals at risk of heart attack or CVDs may demonstrate raised blood pressure, glucose, and lipid as well as overnight and obesity.

Risk factors of Heart Disease [2] [3]:

1. High Cholesterol
2. High Blood Pressure
3. Diabetics
4. Being overnight or obese
5. Smoking
6. Consuming too much Alcohol
7. Poor diet
8. Hereditary (Family history of coronary illness)

Symptoms of Heart Attack [2] [3]:

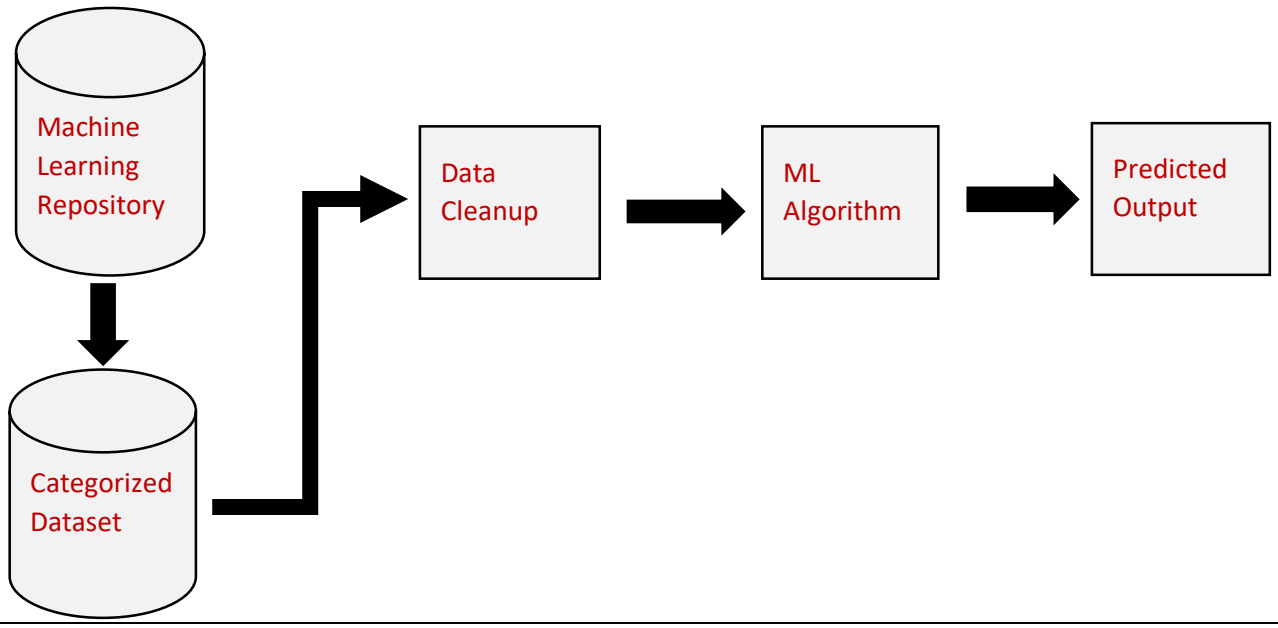
1. Shortness of breath
2. Cold sweat
3. Fatigue
4. Pressure, tightness, pain in chest or arms
5. Heart burn or abnormal pain
6. Pain may spread to left or right arm or to neck, jaw, back, or stomach
7. Lightheadedness or sudden dizziness

Types of Cardiovascular Disease [2] [3]:

1. Coronary artery disease
2. Cardiac arrest
3. Heart attack
4. Heart failure
5. Heart valve disease
6. Pericardial disease and more…..

IV. **DATA DESCRIPTION** [1]**:** The dataset that is used here for predicting heart disease is taken from Kaggle Dataset. Kaggle is the world's largest data science community which has powerful tools and resource that helps people to achieve their data science goals. All the data have been cleaned in python and processed before applied in the ML algorithm. We applied the dataset in two different ML algorithm to differentiate between the two algorithm and to check which one can predict better. The dataset consists of 1025 data with 14 clinical parameters. Commonly used attributes are age, sex, blood pressure, pulse rate, cholesterol, chest pain etc.

V. **ARCHITECTURE DIAGRAM** [2]**:**



vi. **DESCRIPTION OF USING ALGORITHMS:** In this system we have used two ML Algorithms namely (a) K-Nearest Neighbors Algorithm, (b) Random Forest Algorithm.

(a) K-Nearest Neighbors Algorithm (KNN): K-nearest neighbor algorithm is a non-parametric method. It is used for classification and regression. The output of a KNN classification is a class membership. If k=1, then the object is assigned to the class of the single nearest neighbor. The output of a KNN regression is the property value for the object.

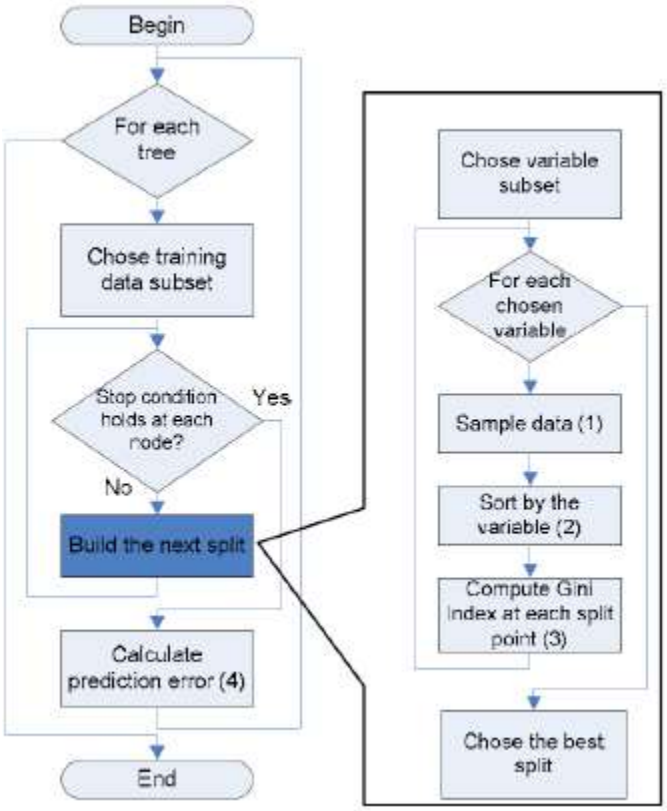
Suppose, we have pairs (X1,Y1), (X2,Y2) ,…….,(Xn,Yn) taking values in R^d*{1,2} where Y is the class label of X, so that X|Y=r ≈ Pr for r=1,2 (Probability distributions Pr).

$$\arg\max_i \sum_{j=1}^{k} sim(D_j \mid D) * \delta(C(D_j), i)$$

(b) Random Forest Algorithm: Random forest is a classification algorithm consisting of many decisions' trees. Formula of random forest algorithm is:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Flow chart of random forest algorithm:



vii. **RESULT AND DISCUSSION:** The main target of this project is to confirm that whether the patient has heart disease or not. The records in the dataset are divided into two parts, one is training set another is test set. After cleaning and processing the data, data classification techniques namely KNN algorithm and Random Forest algorithm were applied in the dataset through python programming. Here some figures are added which are related with the implementation.

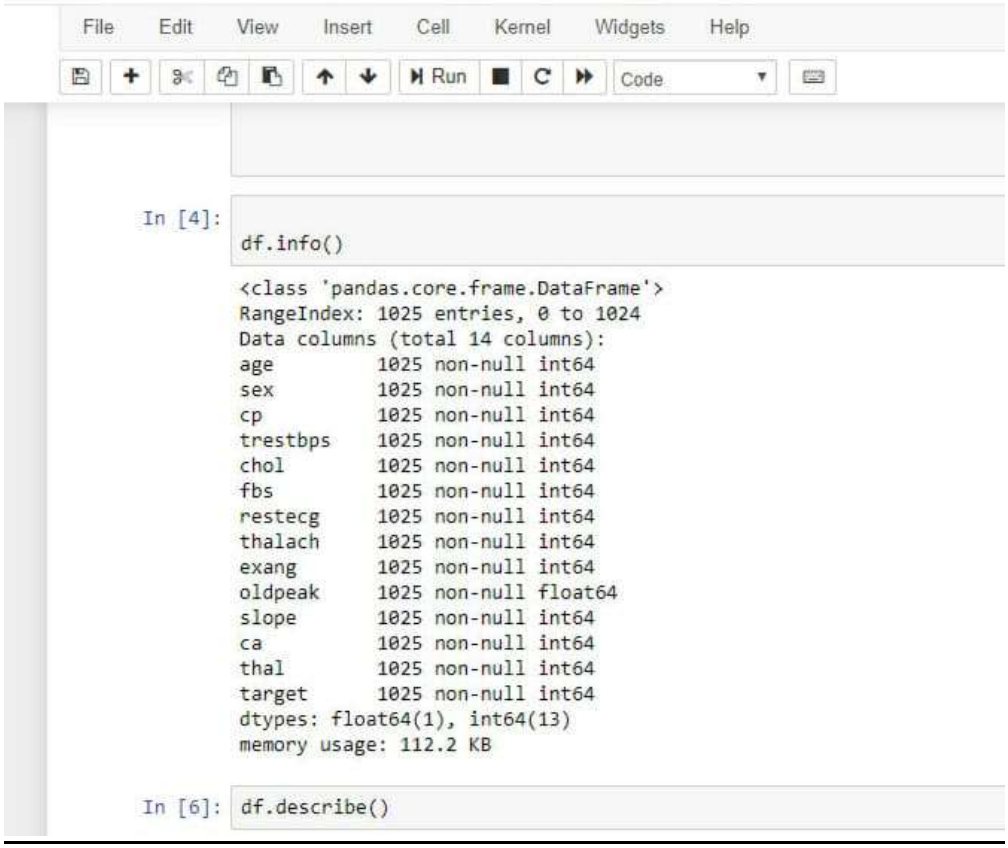


Fig:01

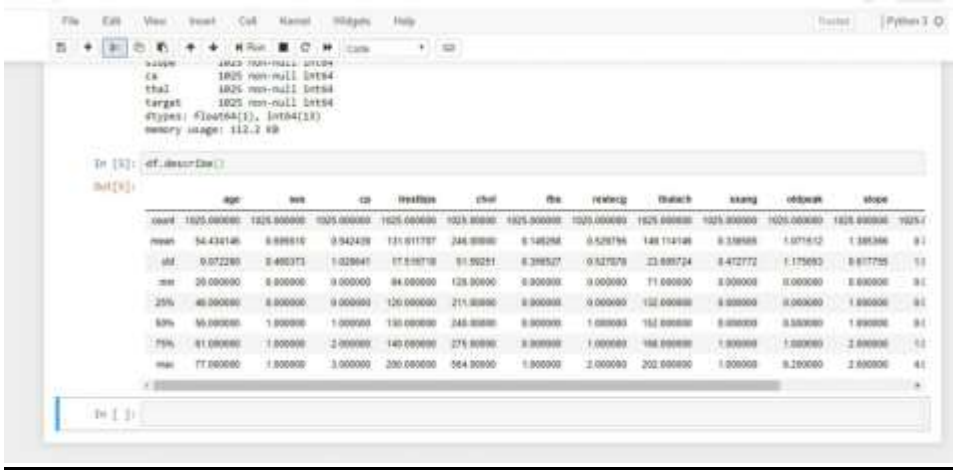


Fig:02

Fig-02 shows the result of mean, std, min, and max of all data in clinical parameters.

Fig:03

Fig-03 shows the statics of all the data.

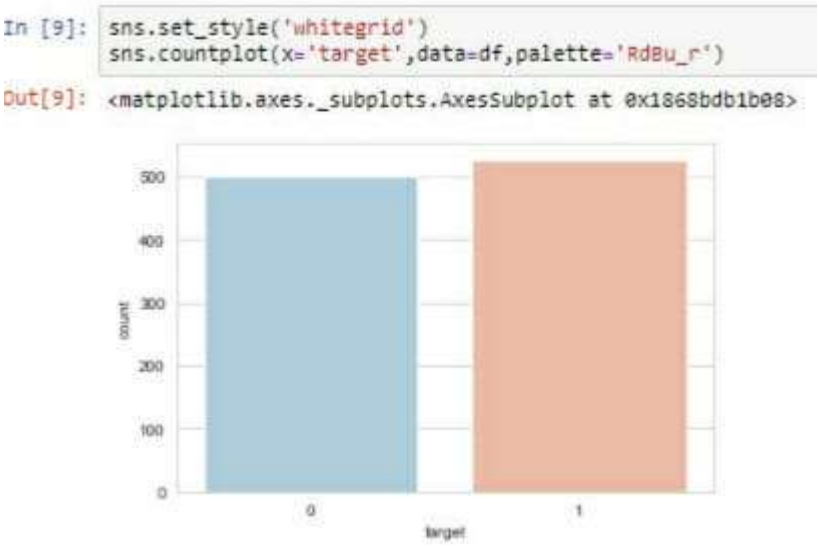


Fig:04

Fig-04 shows the target value where 0 indicates death and 1 indicates well.

Fig:05

Fig-06 shows the target for Random Forest classifier. Seaborn library function is used in python code.

Fig:06

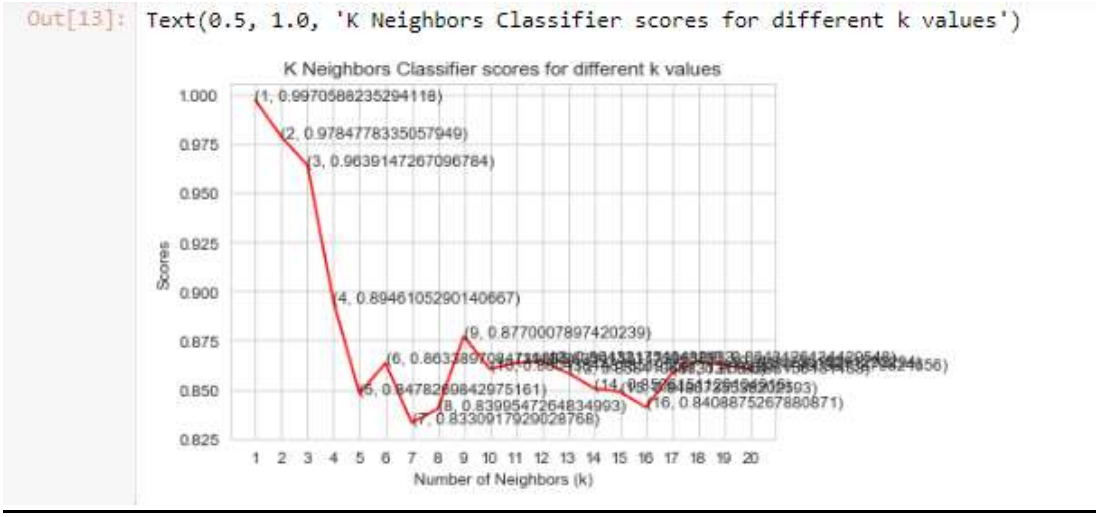


Fig:07

Fig-07 shows the graphical view of KNN classifier scored for different k values.

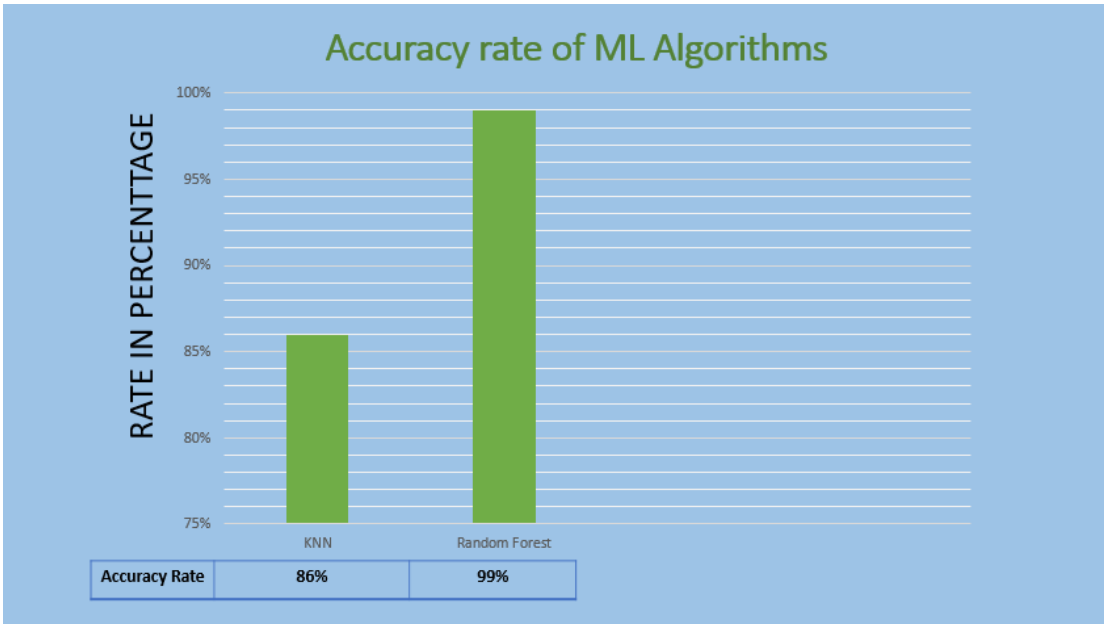


Fig:08

Fig-08 shows the comparison of two models where Random forest showing the highest accuracy rate.

viii. **CONCLUSION:** The main motive of this paper is to provide a system that can predict heart disease occurring rate in percentage using Data Mining and Machine Learning Algorithms. In this paper, two ML Algorithms namely K-Nearest Neighbors Algorithm and Random Forest Algorithm was applied on the dataset. These two algorithms are applied to the same dataset so that we can analyze the best algorithm in terms of accuracy. The K-Nearest Neighbors Algorithm has predicted the heart disease patient with an accuracy level of 86% and Random Forest Algorithm has predicted the heart disease patient with an accuracy level of 99%. This is showing that Random Forest Algorithm is better than K-Nearest Neighbors Algorithm and is best to handle medical dataset. In future, this work can be extended or improved by using other Machine Learning Algorithms and can be developed an automatic heart disease prediction system that both Doctor and patient can monitor live health condition.

References

[1] J. T. Theresa Princy. R, "Human Heart Disease Prediction System using Data," in *2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]*, Bangalore, India-560060. , 2016.

[2] D. Mr.Santhana Krishnan.J, "Prediction of Heart Disease Using Machine Learning Algorithms.".

[3] D. S. N. S. Monika Gandhi, "Predictions in Heart Disease Using Techniques of Data Mining," in *2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management*, Noida, India , 2015.

[4] Chaitanya Suvarna, Abhishek Sali, Sakina Salmani," Efficient Heart Disease Prediction system using Optimization Technique," in Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), Mumbai, Maharashtra., 2017.

[5] John R. Finnegan Jr., Ph.D.,*,2 Hendrika Meischke, Ph.D.,† Jane G. Zapka, Sc.D.,‡ Laura Leviton, Ph.D.,§ Angela Meshack, Dr.P.H.¶ Ruby Benjamin-Garner, M.A.,¶ Barbara Estabrook, M.S.P.H.,‡ Nancy Johnston Hall, M.A.,* Sheri Schaeffer, B.A.,‡ Cassandra Smith, M.P.H.,‡ Elissa R. Weitzman, Ph.D.,¶ James Raczynski, Ph.D.,** and Elaine Stone, Ph.D.††, "Patient Delay in Seeking Care for Heart Attack Symptoms: Findings from Focus Groups Conducted in Five U.S. Regions," in Preventive Medicine 31, 205–213 (2000) doi:10.1006/pmed.2000.0702, available online at http://www.idealibrary.com on, Bethesda, Maryland.

[6] M.A. Rahmana, M.M. Zamanb, " Smoking and smokeless tobacco consumption: Possible risk factors for coronary heart disease among young patients attending a tertiary care cardiac hospital in Bangladesh," Dhaka-1205, Bangladesh,2008.

[7] A.K.M. Monwarul Islam a, A.A.S. Majumder, "Coronary artery disease in Bangladesh: A review," Dhaka 1207,2013.

[8] World Health Organization(WHO).

[9] Wikipedia.