**Machine Learning Predicting Human Heart Disease**

Ms. Soeum Zera

M1 Computer Science

Institute Technology of Cambodia

Phnom Penh, Cambodia

soeumzerayha88@gmail.com

*Abstract*— the technical science is very useful and develops from day to day in medical sciences due the computer ability to detect on human heart disease. As the result from the reports of WHO and AIHW showing that this is the most dangerous disease which leads most people death in the world. Today technology defines the disease in timely it is very important to prevent heart failure in the first condition. From traditional approaches for doing detect heart disease is not comfortable and biased. In this case, machine learning is reliable to detect human who are having disease and non-disease in effectively. For this task, we use three-algorithms for solving problem of human heart predicting and evaluate by Logistic-Regression, Random-Forest and K-Neighbors.

# **Keywords- Heart Disease, Prediction, Logistic Regression, Random Forest, KNN, Machine Learning, Cardiovascular**

# **Introduction**

Heart Disease is the type of illness which is showing about the range condition of heart effect. Heart Disease actually often used the same meaning as Cardiovascular Disease [1]. From WHO (World Health Organization) records the cause of death people by cardiovascular with 17.9 million cases in 2019, it means 32 percent of death causes in the world [2]. Also the (AIHW) Australian Institute of Health and Welfare records in 2018 is showing the death of cardiovascular disease 42 percent of all the death cases [3]. Nowadays, researchers are trying to define and develop techniques to recognize the heart disease effectively and timely, as the existing techniques are ineffective in the recognition for the accuracy and timely of computational [4]. Under the heart disease include blood vessel disease, heart rhythm and congenial heart. Generally it refers to condition which is related to block of blood vessel (the main cause of heart attack), chest pain or stroke [1]. In the time of technology modern, the disease has evolution condition so the expert of healthcare define and manage condition of heart disease is very difficult [5]. The patients can be saved by the best of result in diagnosis and treatment. The medical record history of patients, the report of body examination and other are very useful to diagnose of the heart disease. However all these reports from patients are useful, but it is still not enough to define the condition of the heart disease and it spends costly. So we are doing the research and build the model of machine leaning prediction system to solve of these problems faced on. Heart disease is effectively diagnosed using expert decision-making systems based on a classification of machine studies and pseudo-logic, the computation based on values 1 or 0, yes or no. We use the dataset for training, testing data and predicting model of data from Kaggle.com which name is “heart\_disease\_data.csv” [6]. When we use different model of machine learning algorithm, accuracy we get is also different and it can be approved higher. Predicting models can be enhanced by most features in the data related to the cause of the disease. However the machine learning has more models to predicting but accuracy still not good as what we need. Thus, we need do the data processing and data visualization in our dataset such as scaler, min/max and technical to remove all missing values and outlier in the data.

In this paper work, we improve our algorithm by using machine learning algorithm such as Logistic-Regression, Random-Forest and K-neighbor (KNN) and using dataset from kaggle.com which has 13 features and 1 output as number 1 is representing the human is having disease and number 0 is representing the human is none disease. And the 13 features details in dataset we will describe details in age, sex, chest pain types, resting blood, cholesterol, fasting blood sugar, resting electro cardiogram, max hear rate, exercise angina, oldpeak, ST slope.



Figure1 : Human Heart

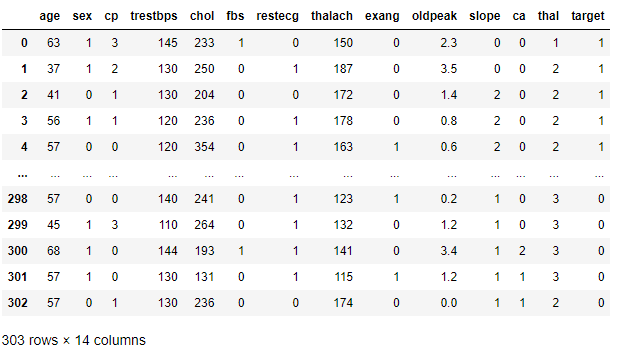
# **Proposed Methods**

Machine learning helps much in computer science. It allow users to compute the large amount of data then it making analysis and making a possible decision based on input data. My paper works with python language and with 3 most popular algorithm of machine learning such as Logistic-Regression, Random-Forest and K-Neighbor.

**Dataset**

The dataset name’s “heart\_disease\_data.csv”, it is from kaggle.com. It contains of 13 attributes and 1 output. And it consists 303 rows and 14 columns.

Figure2 : Dataset details from heart\_prediction\_data.csv

* Age : refers to age of patients occurs in dataset from min of age is 29 and max of age is 77 years old
* Sex : refers to gender of patients which is representing by number 1 is male and 0 is female
* Cp : refers to chest pain types of patients which is representing by 0 is typical angina, 1 is atypical angina, 2 is non-angina pain and 3 is asymptotic
* Trestbs : refers to resting blood pressure values (mmHg)
* Chol : refers to serum cholesterol (mg/dl)
* Fbs : refers to fasting blood sugar values (120mg/dl) which is 1 representing true ( it means the patient has fasting blood sugar 120mg/dl) and 0 is representing false (it means the patient has fasting blood sugar lower than 120mg/dl)
* Restecg : refer to Resting ECG, the result of electro cardio graphic result which 0 is representing normal, 1 is representing having STT-wave-abnormality, and 2 is representing left-ventricular-hypertrophy
* Thalach : refer to max-heart-rate achieved
* Exang : refer to exercise induces angina which is 0 is representing no and 1 is representing yes.
* Oldpeak : refers to ST-depression-induces-by-exercise-relative-to-rest which is in format as float or integer
* Slope : refer to Peak-exercise-ST-segment which is representing 0 is unsloping, 1 is flat and 2 is down-sloping
* Ca : refer to Number-of-Major-Vessels (0-3) color by fluoroscopy which is in format float or integer.
* Thal : refer to thalassemia which is representing 1 is normal, 2 is fixed defect and 2 is reversible defect
* Target: refer to diagnosis of heart disease which is representing by 0 is non disease and 1 is having disease

**Data Processing**

Checking the null values in dataset for 303x14

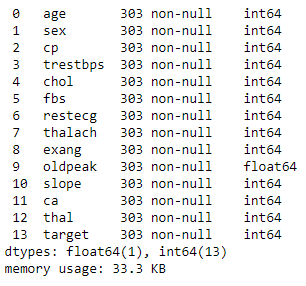


Figure3 : Null values of each columns in data

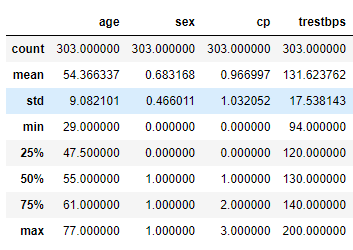
Describe about the dataset in statistic as the result show below:

Figure4 : statistically the dataset is spread

Example in the first columns of age:

* Count : number of all rows (303)
* Mean : average age is 54 years old
* Std : standard deviation is 9.08
* Min : minimum of age in dataset is 29 years old
* 25% : the age younger than 47 years old has 25%
* 50% : the age younger than 55 years old is 50%
* 75% : the age younger than 61 years old is 75%
* Max : maximum age in dataset is 77 years old

Check correlation of features to analyze which feature is negative correlation and which is positive correlation

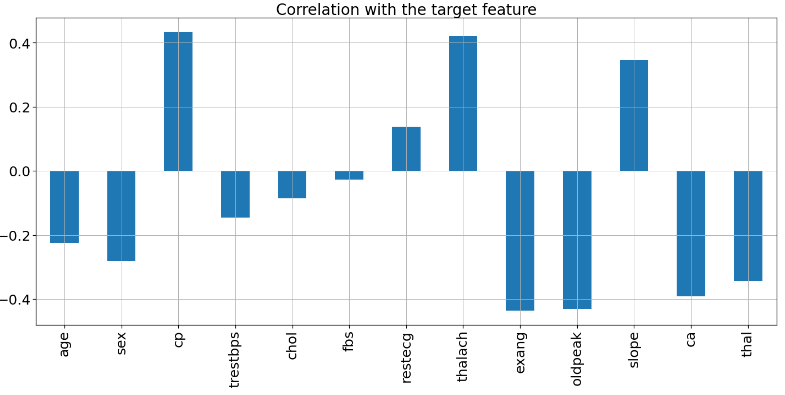


Figure5 : Correlation graph

In this graph show that some features such as cp, restecg, thalach and slope are positive correlation. And other features are negative correlation.

**Data Analysis**

**As this we can see the most age occur in the dataset is 58 years old.**

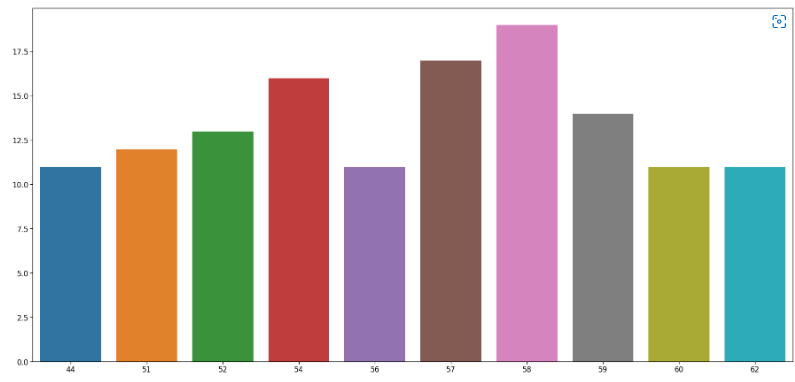


Figure6: Graph age analysis for 10 ages of their counts

We should divide the Age feature into three parts – “Young”, “Middle” and “Elder”:

* Young = 29-39 years old
* Middle = 40-54 years old
* Elder = more than 55 years old

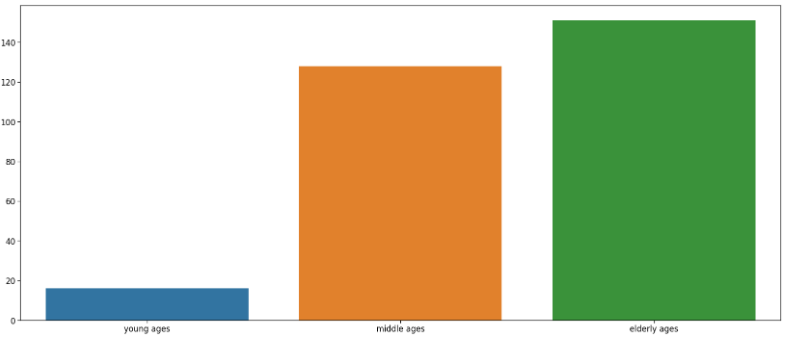


Figure7: We can see the elder is occur most in dataset

**Count the target which 0 is none disease and 1 is having disease**

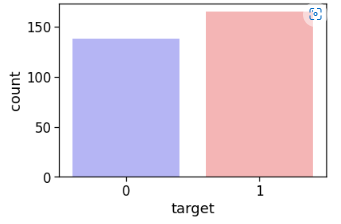


Figure8: show that the patients in dataset having disease more than none disease which represent in the graph 0 is 45% and having disease 56%

**Count the sex column which 0 is female and 1 is male**

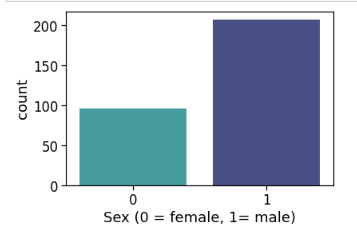
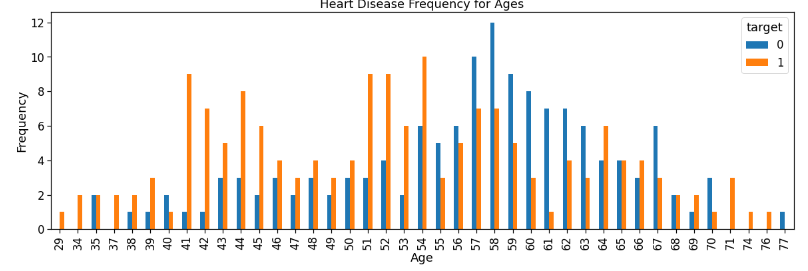


Figure9: show that the male is occur in dataset 69% and female 31%

Figure10: In this graph it is showing that the most disease happening in the age of 54

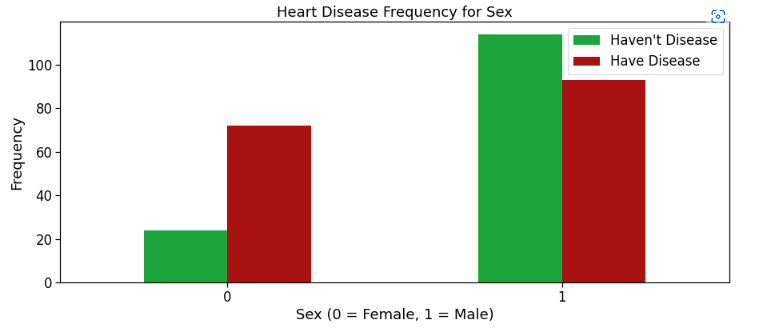
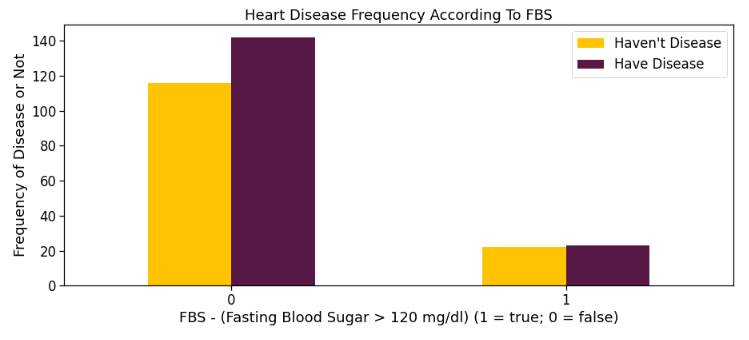


Figure11 : In this graph it is showing that the female is have disease more than none disease but male is having disease less than none but if compare male and female, we see that male is having disease more than female

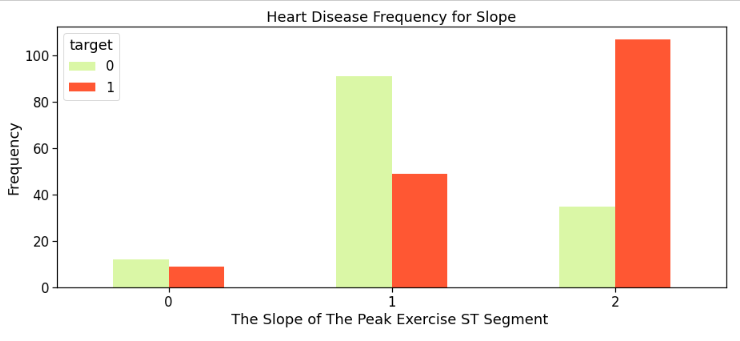


Figure12 : In this graph it is showing amount of 3 kinds of slope is the number 2 is refer to downsloping is having disease much than other

Figure13 : showing the fasting blood pressure which is the patient has fasting blood sugar lower than 120mg/dl is faced on having disease

##### 

Figure14 : show that number 2 which is representing non-angina pain is higher frequency of have disease

##### 

Figure15 : we divide data in 2 parts, the first is categorical values and second is continuous values.

* For categorical values is setting the values less than 10 because it is usually mention​ most only 3 or 4 different types such as sex, exang, ……
* For continuous values is the number from real data such as age, oldpeak, …..



Figure16 : Using standard scaler in column age, trestbps, chol, thalach,oldpeak

We use standard scaler to remove means and scale each feature to unit variance. It is also effect in outlier if it exist in dataset.

**Normalization**

* In our dataset has 13 features and 1 output name is “target”. So when I try to normalize our data I just drop “target” out from data and set it as “x\_data”.
* For “target” column just set it as “y”
* Next take the “x\_data” do normalization to “x” by this formular below:

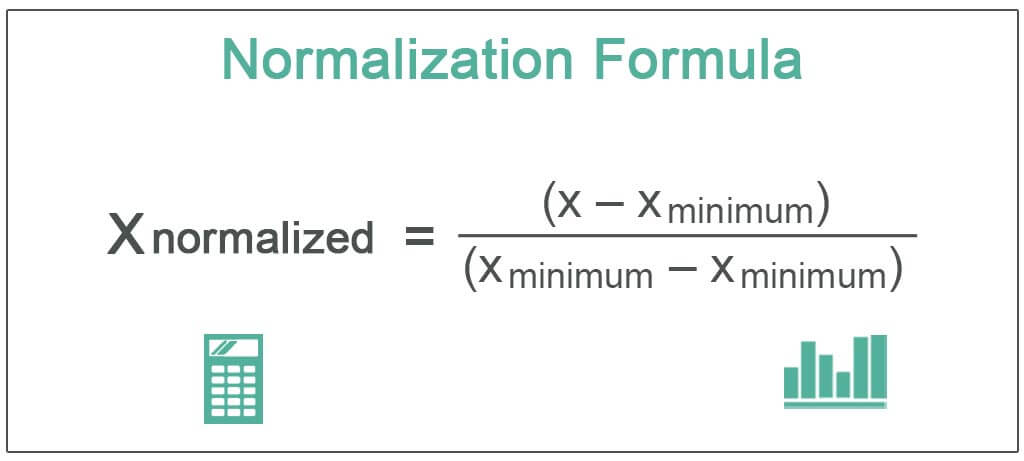


Figure17: Normalize Formula

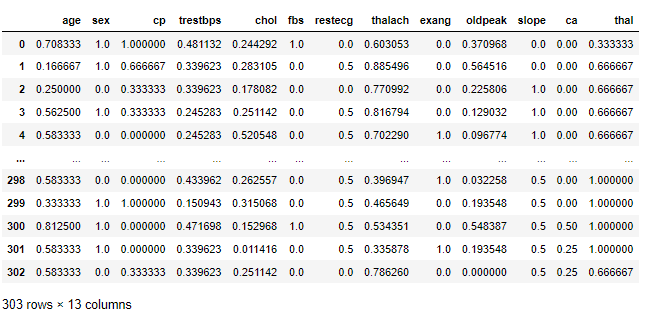


Figure18 : Result of “x” after normalize from “x\_data”

**Split data to train and test**

After we get “x and y” data now split them as 80% is for training data and 20% is for testing data

**Model Algorithm**

* Logistic-Regression

It is used to calculate or predict the probability of binary classification (0/1 or Yes/No)

* K-Neighbor (KNN)

This algorithm makes higher accuracy because it is make high prediction. KNN uses for classification and regression problems. The numbers of nearest neighbor to a new unknown variable for predict is a values set by “k”. In our model we set k=3.

* Random-Forest

It is the similarity of KNN which use for both classification and regression problems. But it builds a decision tree on different samples.

# **Results**

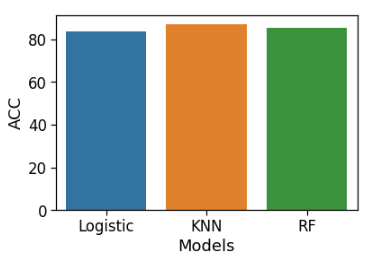


Figure19 :For this results of three algorithms we see KNN is the best model for our data training.

# **Prediction of data**

# From results above, we use the best model (KNN) for making prediction on our heart disease.

# First, just type input data of each feature for making our computer doing predicting in correctly. Example we have 13 features in each columns such as age, sex, chol, exang, restbps, slope,….

* Then in my implement python code, after getting input values so we must change the input data to numpy arrays
* Reshape the numpy array as we are predicting for only instance (1 or 0).
* In this case, we set condition for our prediction. If the result of our predict =1 show “ The Person has Heart Disease” but if our predict =0 show “ The Person dose not have a Heart Disease.”

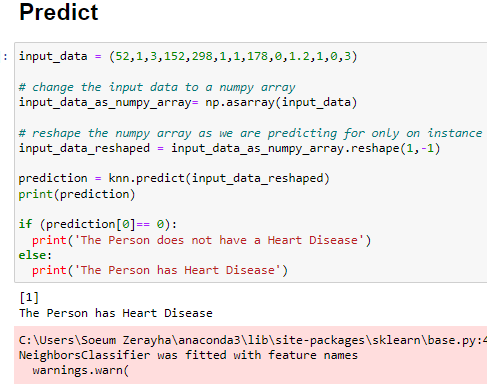


Figure20 : Example of our prediction

# **Conclusion**

From traditional approaches for doing detect heart disease is not comfortable and biased. We build machine learning algorithm to make effectively in predicting data. As we see from the results above we are getting higher accuracy with Logistic-Regression is 83.6%, K-Neighbor is 86.88% and Random-Forest is 85.24%.

For future work, we plan to build other model such as SVM, XGBoots to compare with these 3 algorithm which are achieve higher accuracy. In other hand, we plan to build heart disease predicting system.

##### References

[1] <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc> , 2019

[2] [https://www.who.int/news-room/fact% 20sheets/detail/cardiovascular-diseases-(cvds)](https://www.who.int/news-room/fact%25%2020sheets/detail/cardiovascular-diseases-(cvds))., “World health organization

[3] <https://www.aihw.gov.au/reports/australias-health/coronaryheart-disease>. , Heart Disease 2020

[4] Decision making in advanced heart failure: a scientific statement from the American heart association, 2012

[5] Innovative artificial neural networks-based decision support system for heart diseases diagnosis,2013

[6]<https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data> , dataset of heart Disease