Heart Disease Analysis and Prediction

Kavita Soni

Department of Physics and Computer Science,
Wilfrid Laurier University,
Waterloo, Ontario – N2M 5C8

{ soni1197@mylaurier.ca }

Abstract:
Preventing heart disease is important. Good data-driven systems for predicting heart disease can improve the entire research and prevention process, making sure that more people can live healthy lives. Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men. Coronary heart disease (CHD) is the most common type of heart disease, killing over 370,000 people annually. Heart disease is the leading cause of death for people of most ethnicities in the United States, including African Americans, Hispanics, and whites. For American Indians or Alaska Natives and Asians or Pacific Islanders, heart disease is second only to cancer. Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack. Here the paper strives to understand various traits given as attributes the data that help to predict if a person is likely to have a heart disease or not. Paper explores the data and builds three models for prediction and chooses the most accurate one.

I. INTRODUCTION

Machine learning and artificial intelligence is going to have a dramatic impact on the health field. Heart disease or Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. Cardiovascular diseases are the leading cause of death globally. This is true in all areas of the world except Africa. Together CVD resulted in 17.9 million deaths (32.1%) in 2015. Deaths, at a given age, from CVD are more common and have been increasing in much of the developing world, while rates have declined in most of the developed world since the 1970s.
This paper focuses on predicting the presence of heart disease in person based on attributes such as blood pressure, cholesterol levels, heart rate, and other characteristic attributes, patients will be classified according to varying degrees of coronary artery disease.

II. OBJECTIVE

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis. The objective is to predict potential Heart Diseases in people using Machine Learning algorithms. This paper utilizes a dataset of 303 patients and distributed by the UCI Machine Learning Repository. We apply three machine learning algorithms, namely KNN (K Nearest Neighbor), Logistic Regression and Decision Tree. Each model is trained on the training set and then tested on the test set. These models consider the history of a patient's records and predict whether each individual patient will be suffering from heart disease or not, thereby, alerting the health care system and potentially triggering preventive actions.

III. DATA AND METHODS

A. Detailed Data Description and Analysis

The dataset has been taken from Kaggle. This dataset gives several variables along with a target condition of having or not having heart disease. We will try to use this data to create a model which tries predicting if a patient has this disease or not. This dataset has 14 columns (i.e. 13 feature columns and 1 target column) and 303 rows. The various categories of medical factors, along with the number of factors are shown in the Table 1. 80% of that set forms the *training set* – used for training algorithms – and the remaining 20% is designated as the *test set* and used exclusive for evaluating the performance of the algorithms. Also, there are no missing values so we don't need to take care of any null values.

AGE: (age in years)
SEX: (1 = male; 0 = female)
CP: (chest pain type)
TRESTBPS: (resting blood pressure (in mm Hg on admission to the hospital))
CHOL: (serum cholestoral in mg/dl)
FPS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
RESTECH: (resting electrocardiographic results)
THALACH: (maximum heart rate achieved)
EXANG: (exercise induced angina (1 = yes; 0 = no))
OLDPEAK: (ST depression induced by exercise relative to rest)
SLOPE: (the slope of the peak exercise ST segment)
CA: (number of major vessels (0-3) colored by flourosopy)
THAL: (3 = normal; 6 = fixed defect; 7 = reversable defect)
TARGET: (1 or 0)

Table 1: Data Description

Also 45.54% of patients (count: 138) do not have heart disease and 54.46% patients (count: 165) suffer from heart disease, shown in Fig. 1:

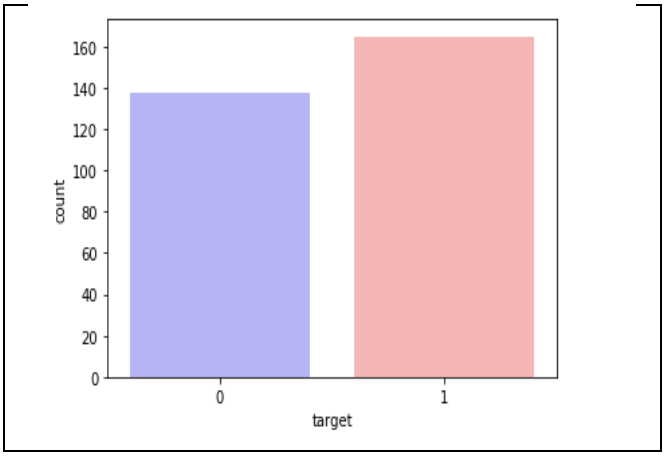

Fig 1: Heart Disease patient count

Out of 96 females - 72 have heart disease and 24 do not have heart disease. Similarly, out of 207 males - 93 have heart disease and 114 do not have heart disease.

Interpretation of correlation coefficient with different features and target values can be seen in Table.2 and through heat map in Fig.2. We can see that there is no variable which has strong positive correlation with target variable since there are no values close to +1. Also we can see that there is no variable which has strong negative correlation with target variable. There is no correlation between 'target' and 'fbs'. Also 'cp' and 'thalach' features are mildly positively correlated.

Feature	Correlation coefficient
Target	1.00000
Cp	0.433798
Thalach	0.421741
Slope	0.345877
Restecg	0.137230
Fbs	-0.028046
Chol	-0.085239
Trestbps	-0.144931
Age	-0.225439
Sex	-0.280937
Thal	-0.344029
Ca	-0.391724
Oldpeak	-0.430696
exang	-0.436757

Table 2: Correlation coefficient

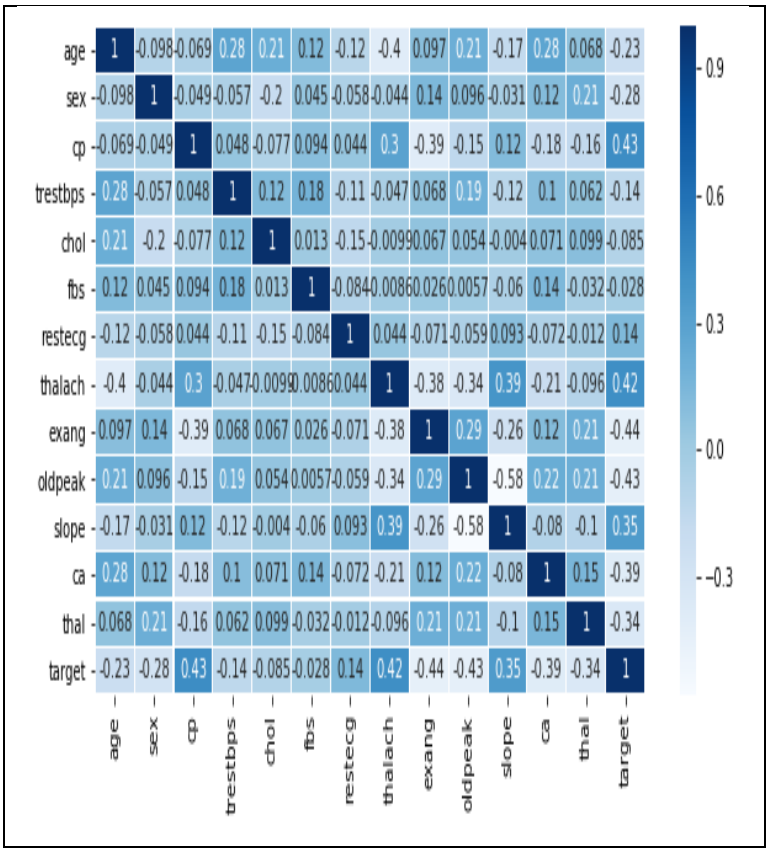

Fig 2: Heat Map

Analyzing 'cp' and 'thalach' we can see that those people suffering from heart disease (target = 1) have relatively higher heart rate (thalach) as compared to people who are not suffering from heart disease (target = 0). Fig.3 shows the frequency of heart diseases based on the type of chest pain.

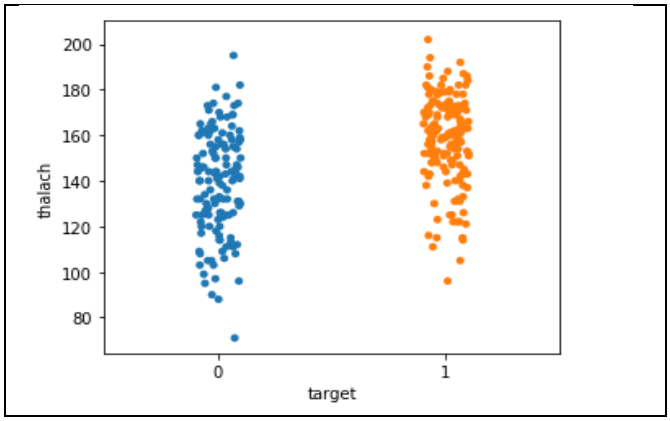

Fig 3: Frequency distribution of 'thalach' vs 'target'

Fig.4 shows the frequency distribution of 'cp' vs 'target'. Statistics for people suffering with heart disease with different types of chest pain experienced (Value 1: typical angina- 39, Value 2: atypical angina- 41, Value 3: non-anginal pain- 69, Value 4: asymptomatic- 16).

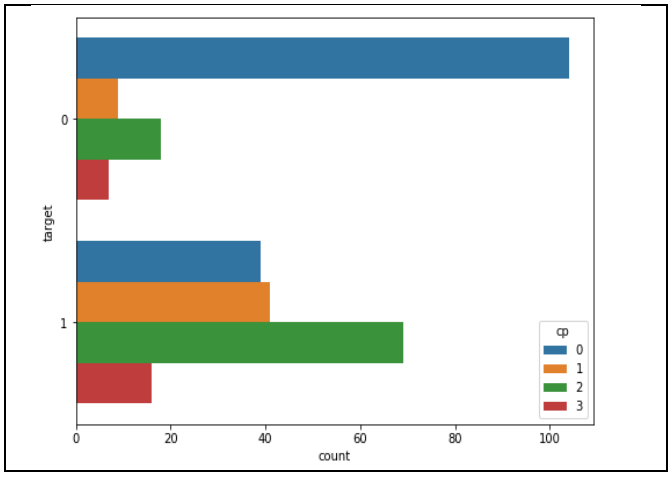

Fig 4: Frequency of 'cp' vs 'target'

The Bivariate plot shown in Fig.5 between cholestrol levels and target suggests that the Patients likely to suffer from heart diseases are having higher cholestrol levels in comparison to the patients likely to not suffer from the heart diseases.

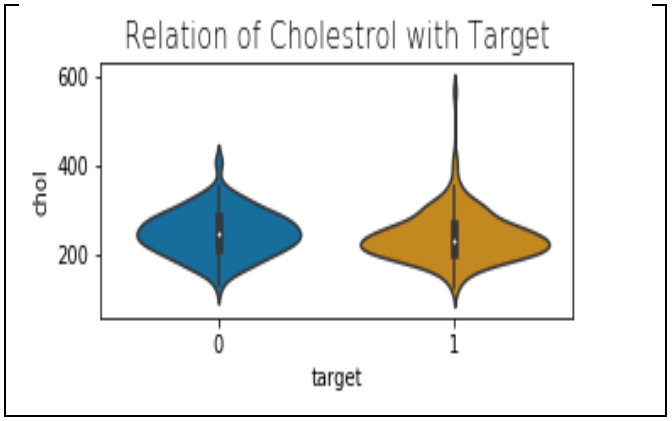

Fig 5: Frequency distribution of 'cholestrol' vs 'target'

B. Model Building

Here three algorithms for model building have been used, and their accuracy is compared. The target column is dropped from the database, then the values are normalized using the following Equation.1.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Equation 1. Normalization equation

The first method to be used is KNN Model. This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied. I varied them from 1 to 20 neighbors and calculated the test score in each case. Then a line graph is plotted for the number of neighbors and the test score achieved in each case, as shown in Fig.6.
It is observed that the highest accuracy is achieved at n = 3 with maximum score of 88.52%.

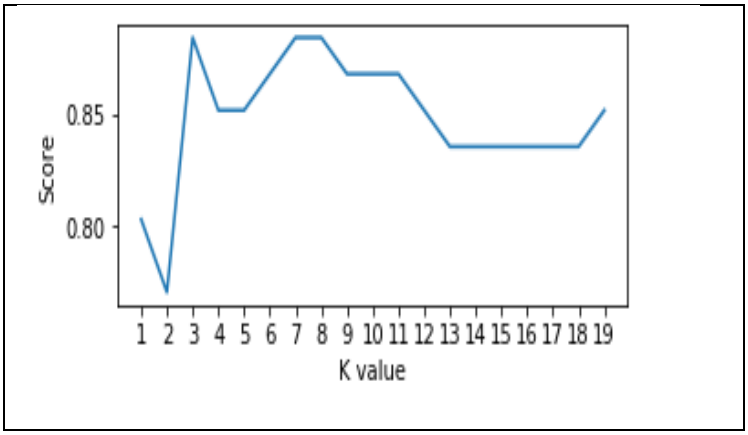

Fig 6: KNN model scores for different K values

The Roc curve for KNN model is as shown in Fig.7, KNN model gives an excellent Roc AUC value of 0.9172.

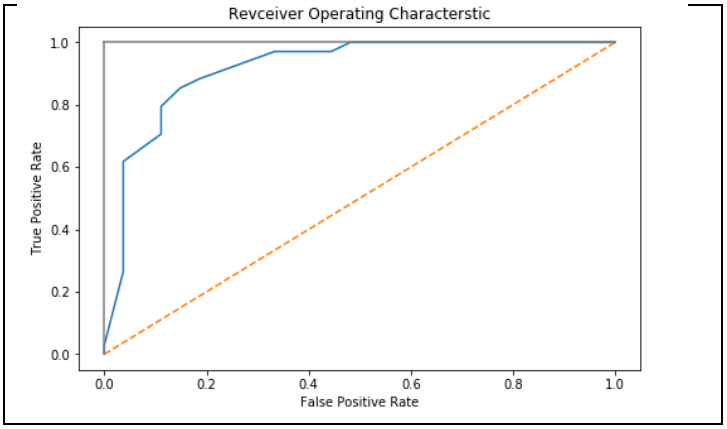

Fig 7: ROC Curve for KNN Model

The second method used is Decision Tree algorithm. A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. This gives us accuracy of 80.33% with Roc AUC value of 0.8.

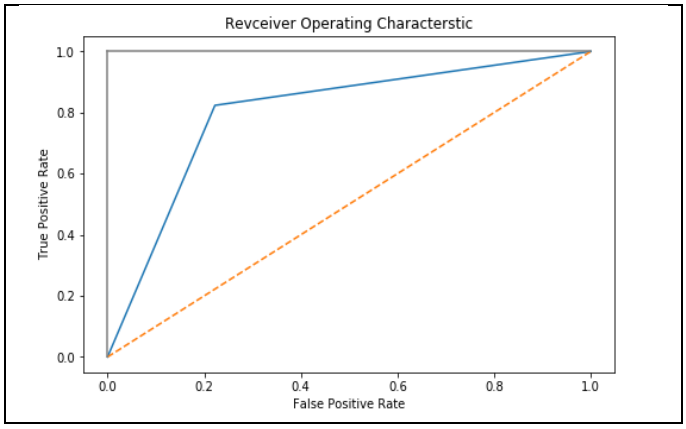

Fig 8: ROC Curve for Decision Tree Model

Accuracy can be improved by tuning the parameters in the Decision Tree Algorithm. Optimization of decision tree classifier is performed by only pre-pruning. Maximum depth of the tree can be used as a control variable for pre-pruning. By applying pre-pruning, on the same data with max_depth=3 and using selection measure as entropy and accuracy was increased to 81.97%.

The third method used is Logistic regression. Logistic Regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic/sigmoid function. The output is binary or in the form of 0/1 or-1/1. Accuracy of 86.89% is achieved through this model. Fig.9 shows Roc curve for this method, the value of Roc AUC is 0.9204.

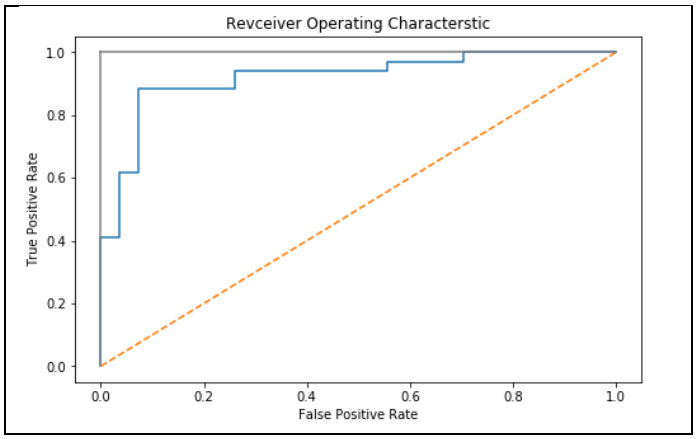

Fig 9: ROC Curve for Logistic Regression

C. Comparison of Models

Three approaches are compared by bar-plot, as shown in Fig.10 and confusion matrix for each model.

Fig 10: Comparison bar-plot

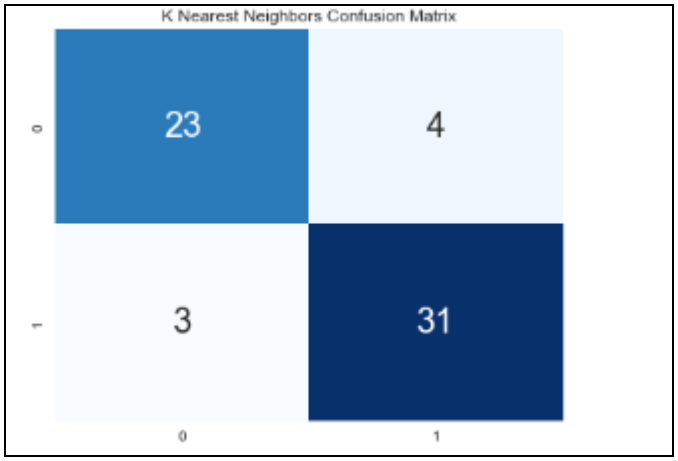

Fig 11: K Nearest Confusion Matrix

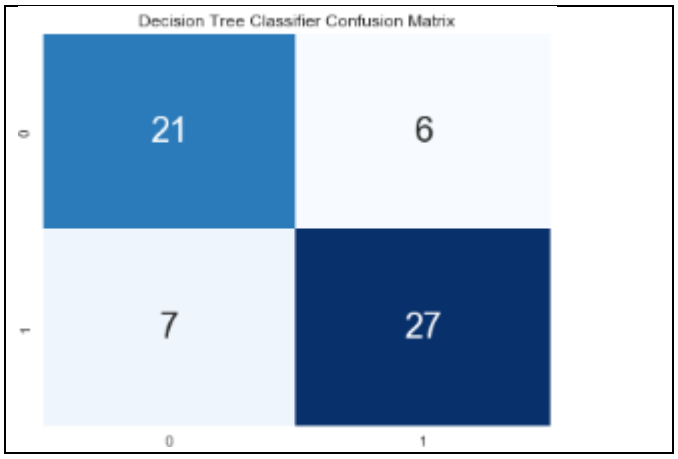

Fig 12: Decision Tree Confusion Matrix

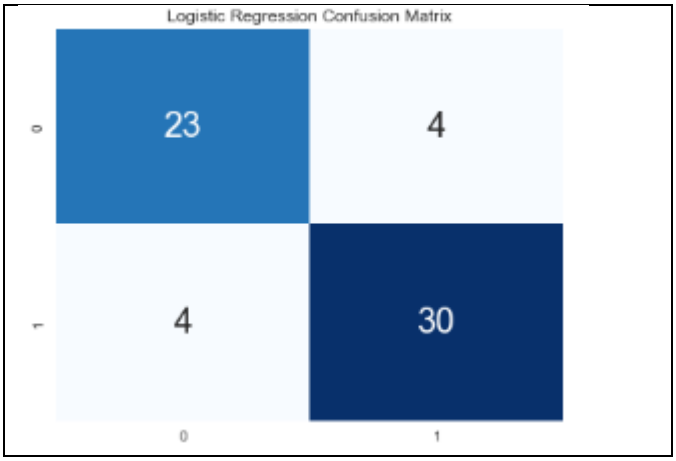

Fig 13 : Logistic Regression Confusion Matrix

IV. CONCLUSION

The paper involved analysis of heart disease patient dataset with proper data exploration and analysis. Then three models were trained and tested with maximum score as follows :

1. KNN Model : 88.52%
2. Decision Tree : 80.33%
3. Logistic Regression: 86.89%

Also the three main features contributing towards decision of model are ‘thalach’, ‘cp’ and ‘cholestrol’.

V. REFERENCES

1. https://seaborn.pydata.org/generated/seaborn.heatmap.html
2. https://medium.com/greyatom/logistic-regression-89e496433063
3. https://www.datacamp.com/community/tutorials/decision-tree-classification-python
4. https://stackabuse.com/understanding-roc-curves-with-python/