

Healthier Heart Today

For A Safer Tomorrow

predicting early signs of heart diseases

ASHOK BHATRAJU

DANBI KIM

JANHAVI POWALE

TRANG TRAN

KARANSINH RAJ

|  |
| --- |
| **List of Change in the report vs presentation** |
| SVM model was added in the report |

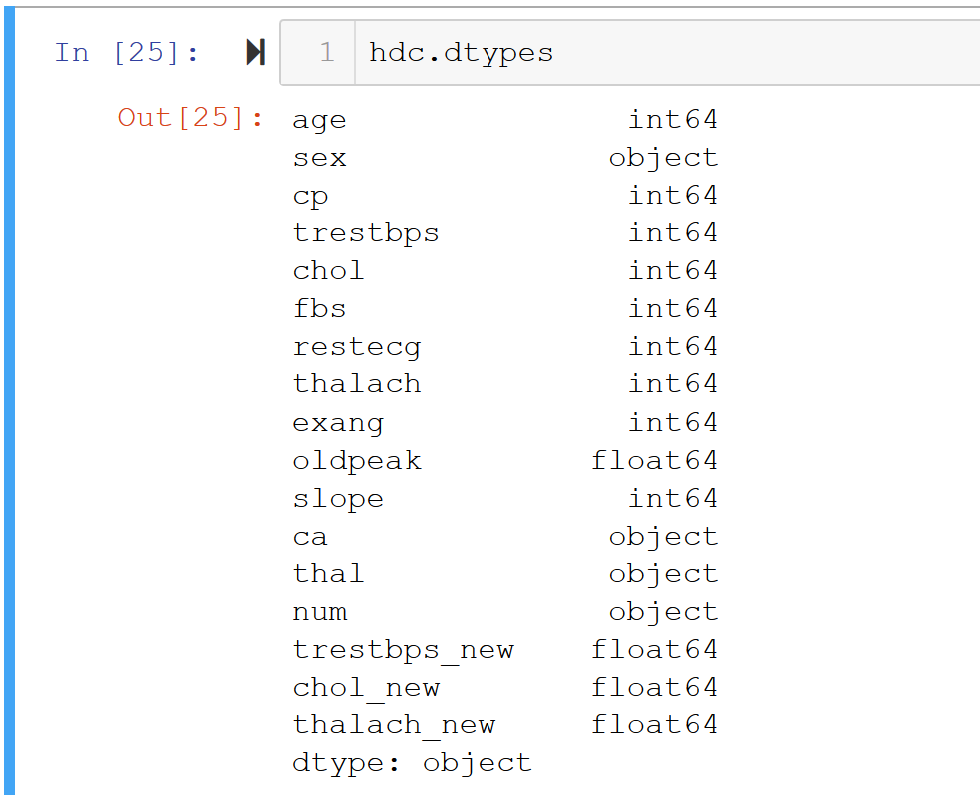
| PROGRAMMING FOR DATA SCIENCE | APRIL 28TH

# Introduction:

Heart disease is a leading cause of death in the USA for both men and women across all ethnicities. It’s responsible for 25% deaths in USA. An American person dies from cardiovascular disease every 37 seconds, leading about 647000 people die in the USA every year. $219 billion dollars have been spent each year from 2014 to 2015 in the USA for health care services, medicines and lost productivity due to heart disease. The disease results commonly in coronary artery diseases and heart attacks. The three risk factors are high blood pressure, high blood cholesterol and smoking [1]. Even though heart disease affects a lot of American lives, some challenges still exist in the medical system. Medical systems have difficulties in distribution of senior clinicians, higher rate of mis diagnosis and longer training period of staffs. Mis- classifications, specifically, can lead to higher expensive for end consumers. Machine learning can be a solution to predict if patients are suspected of heart diseases more accurately than human medical professionals since it is based on supervised learning algorithm from past data. With machine learning, we can get the right medical advice at the right time with no human intervention. Machine learning can improve accuracy of operations and diagnostic efficiencies.

The goal of this paper is to develop an effective Machine Learning predictive algorithm that could help predict the “Presence of Heart Disease” in a patient. In order to develop such an algorithm, a potential patient-level data was needed. This search of data was catered by the ML Repository of UCI. The “Heart Disease Data Set” is used for this project. The models that we built are decision tree model, k-nearest neighbor (kNN) model, support vector machine (SVM), and logistic regression. The models ’performances are evaluated using accuracy, area under curve of ROC curve and recall.

# Dataset:

The dataset which is used here is a subset of the original dataset containing 76 attributes. This dataset consists of 14 features (and referred to as “hdc” in our reports).

The dataset contains attributes describing a range of conditions that affect the heart, this data is collected for people from Cleveland . The detailed description is as follows:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type

Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

1. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
2. chol: serum cholesterol in mg/dl
3. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
4. resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

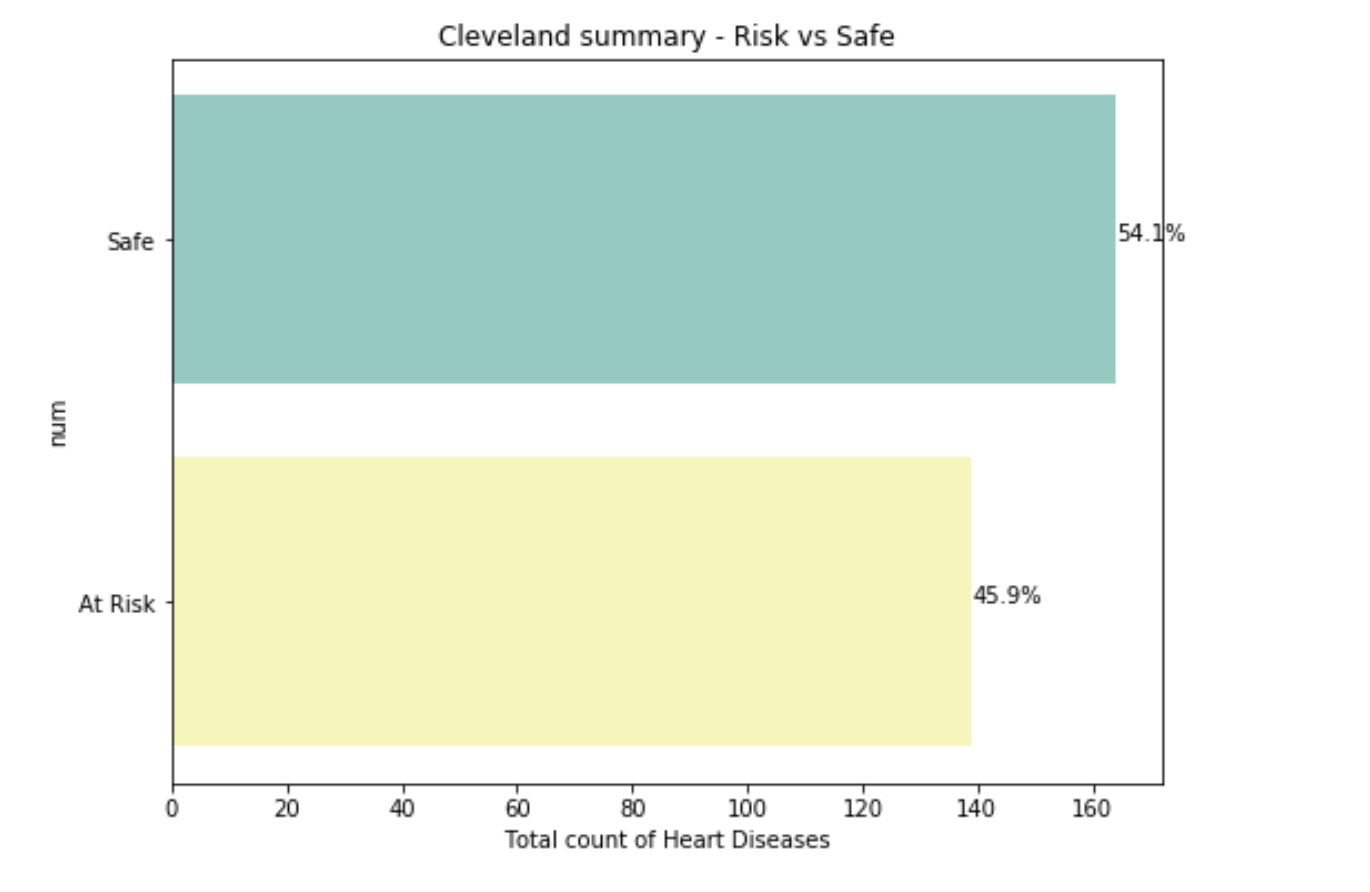
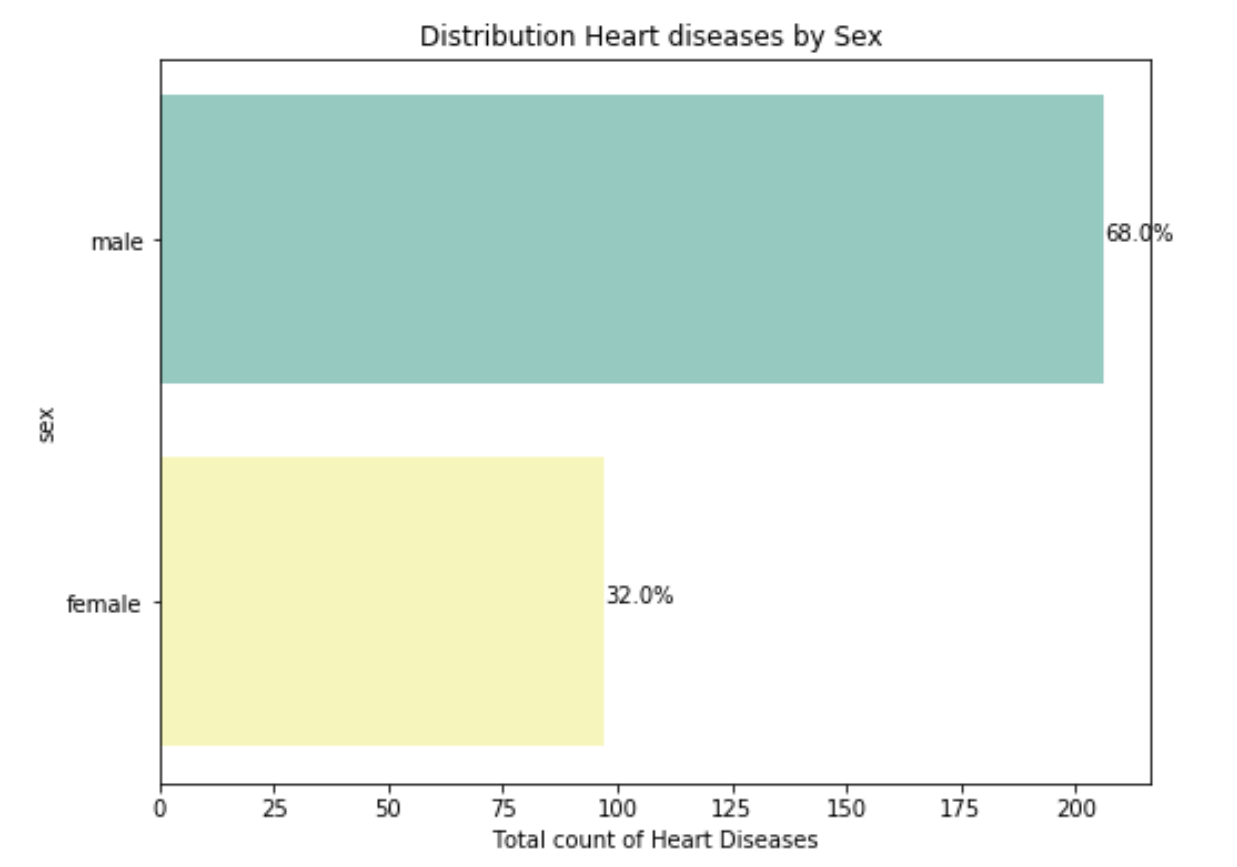
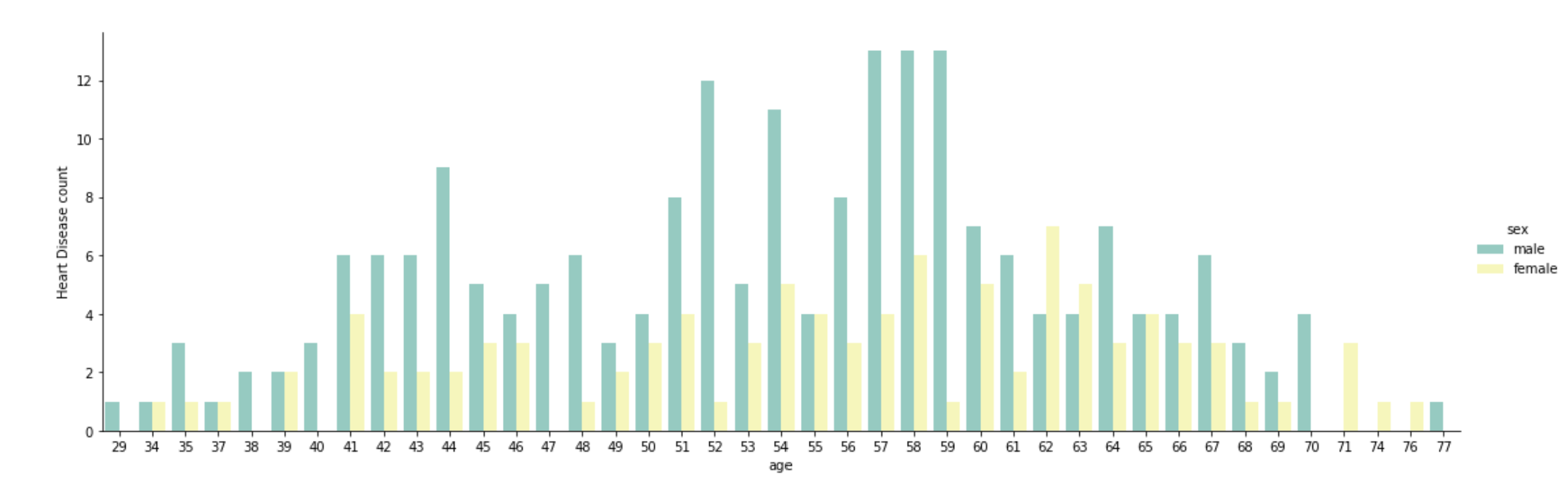
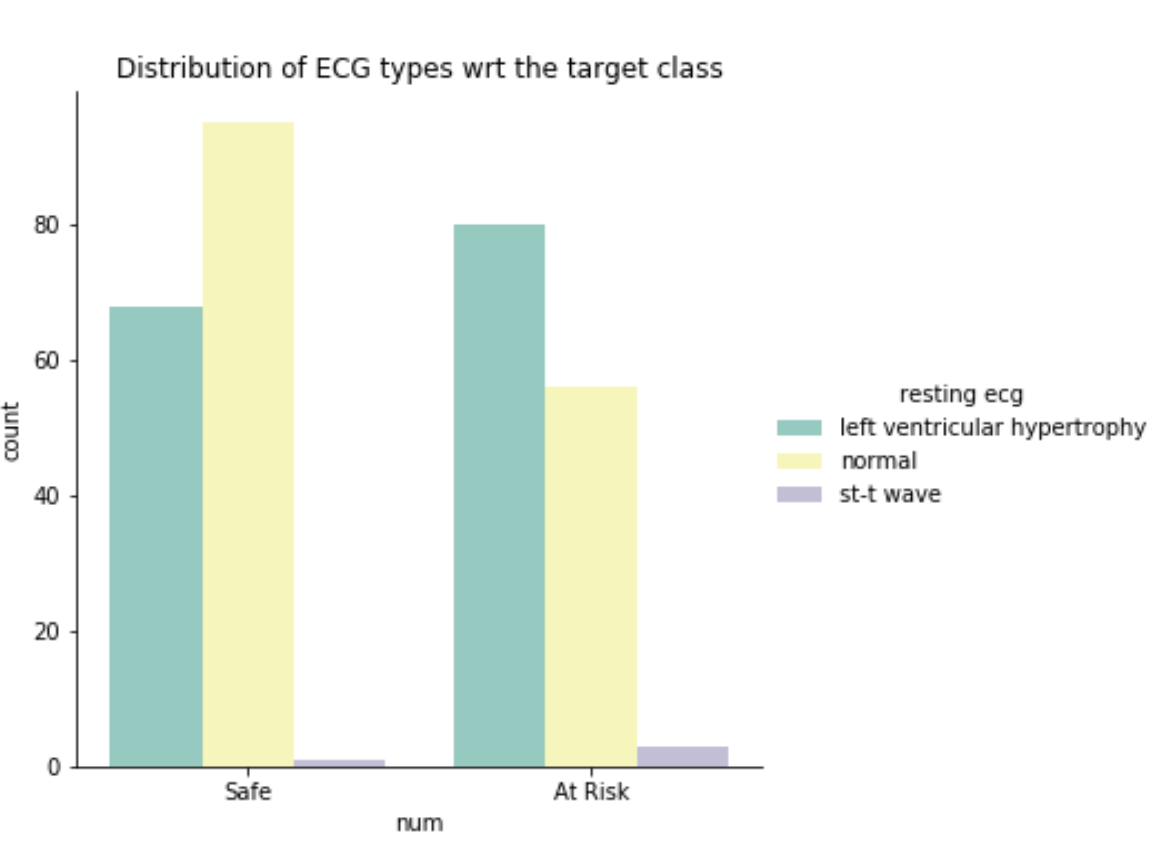
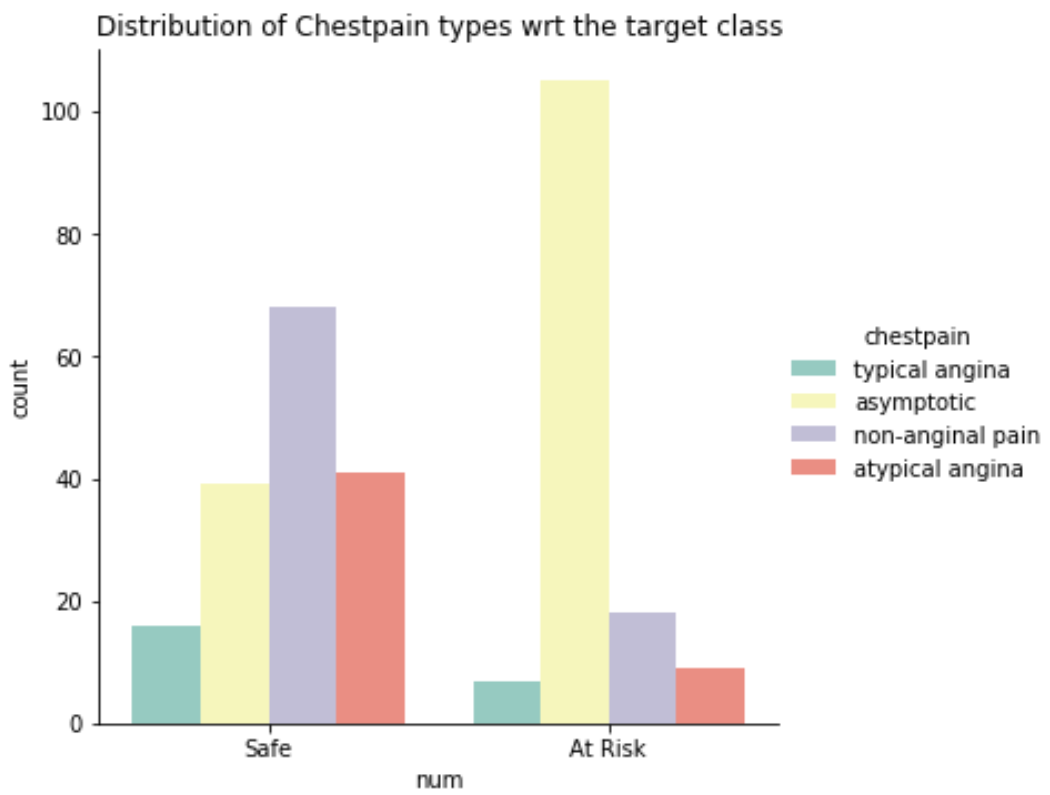
1. thalach: maximum heart rate achieved
2. exang: exercise induced angina (1 = yes; 0 = no)
3. oldpeak = ST depression induced by exercise relative to rest
4. slope: the slope of the peak exercise ST segment (1 Upslope, 2 flat, 3 Downslope)
5. ca: number of major vessels (0-3) colored by fluoroscopy
6. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
7. num: diagnosis of heart disease (angiographic disease status - and the target variable)

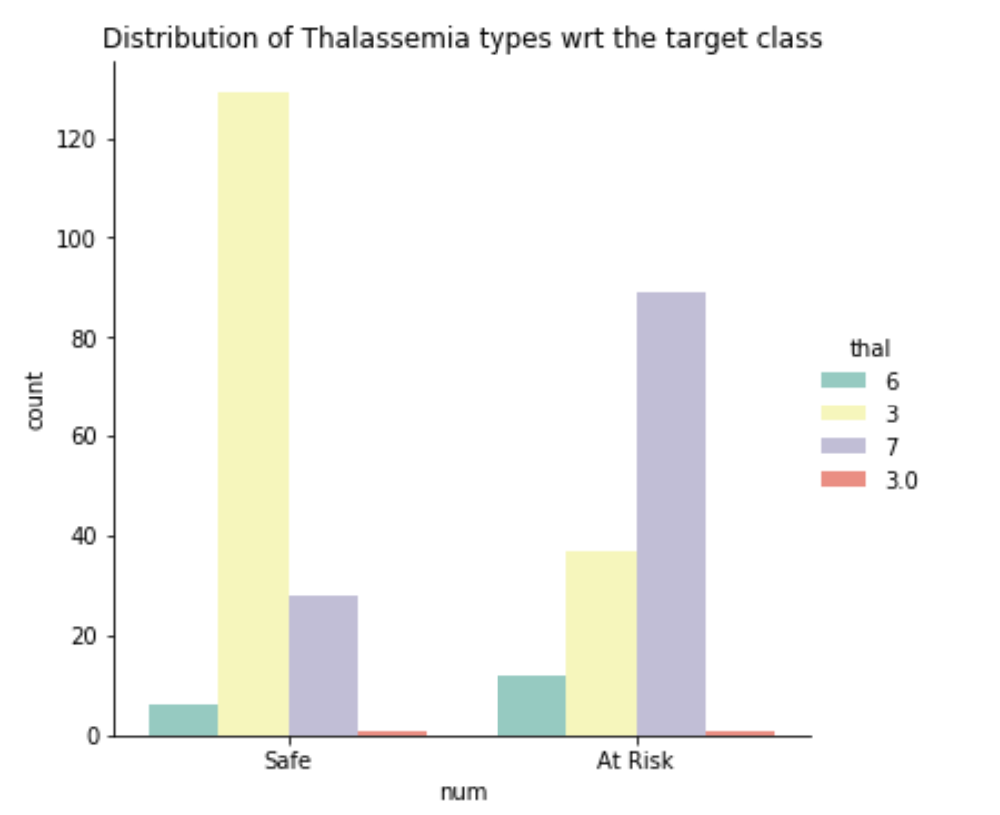
Value 0: < 50% diameter narrowing

Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)

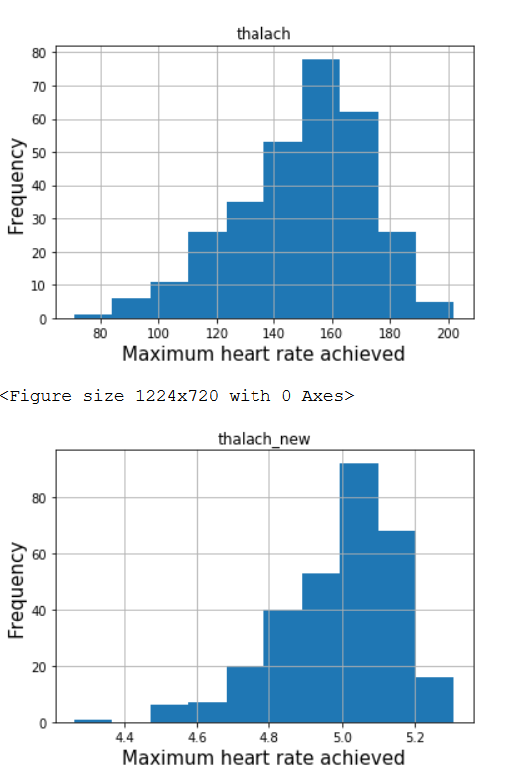
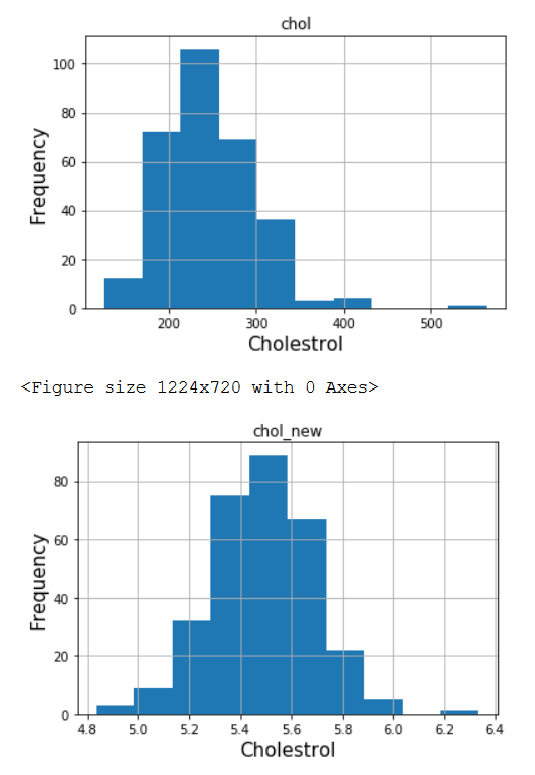
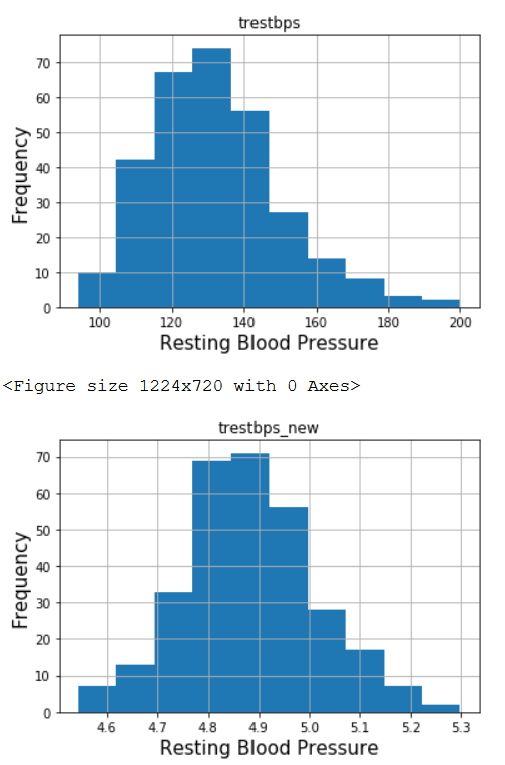
# Data exploration:

Using the above dataset, it was possible to get following insights by doing bivariate analysis:

1. **Class balancing:** From the following plot it is observed that, the used multivariate dataset is balanced in reference to target varia
2. **Checking the distribution of Heart diseases for male and female class**: It is clear that Males are having nearly double the quantum of registered heart disease.
3. **Age distribution**: Here we get the distribution of heart diseases by males and females as per their age. We see that the ages 57-59 yrs are most prone to heart diseases for male category and female risk peaks at 58 and 62yrs. 
4. **Symptom analysis for Heart Disease:** We see that patients with asymptomatic chest pain type have the most propensity to get heart diseases. This is usually risky because one can never understand till the very end and condition of the patient can deteriorate quickly. Likewise, patients with left ventricular hypertrophy (thickening of heart's wall) results in high risk for heart diseases.



1. **Checking for Normality:** Further, before applying the linear models it is important to check for the normality of numeric attributes. On checking the attributes, we found that some of them needed normalization (log normalization was done). It has effectively helped us tackle the skewness in first two attributes



# Model

## DECISION TREE

Decision tree is one of predictive models commonly used in data mining to predict target outcomes from input variables. Decision tree is a complex version of “if-then” algorithm. Based on the number of different predictors, the algorithm will generate various “if-then” rules to split data from top to bottom direction. The accuracy and model performance depends on amount of split and purity of those splits. Decision tree can be used for classification problems. In this project, we try to predict if a patient is at risk of heart disease which is a classification problem that decision tree method is suitable.

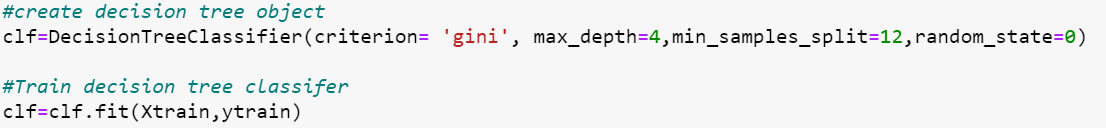
In decision tree model in Python, some parameters can affect the performance of the model such as “criterion”- a function to measure the quality of a split, “max\_depth”- the maximum depth of the tree, and “min\_samples\_split”- the minimum number of samples required to split an internal node. In decision tree model, criterion can be measured by Gini or Entropy for information gain. Gini index measure of a node is the degree or propability of a particular variable being wrongly classified when it is randomly chosen:

Entropy measures the degree of disorder to determine which feature/ attribute have maximum about the class. The formula of entropy is:

In order to determine parameters that give good performance for the dataset, a grid search was performed with cross validation of 5. Criterion can be tested with Gini and Entropy. Maximum depth of tree is varied with range from 3 to 15. Minimum samples split ranges from 2 to 20 with increment of 2 per step. The decision tree model that gives the best performance has parameter as following:

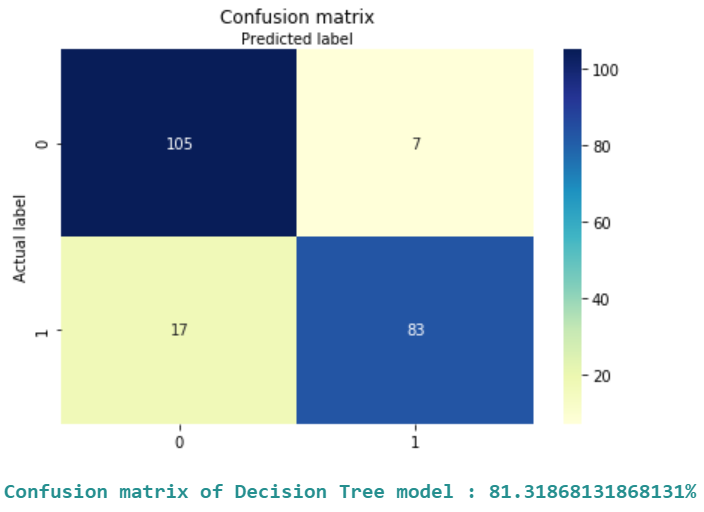
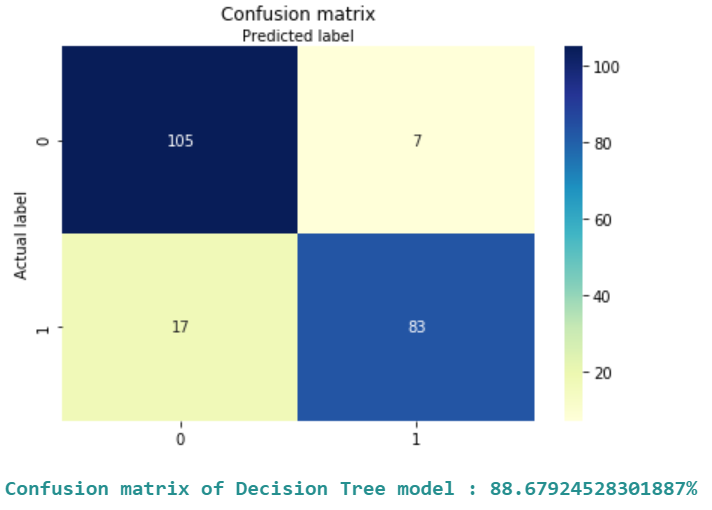
|  |  |
| --- | --- |
| Criterion | Gini |
| Max\_depth | 4 |
| Min\_samples\_split | 12 |

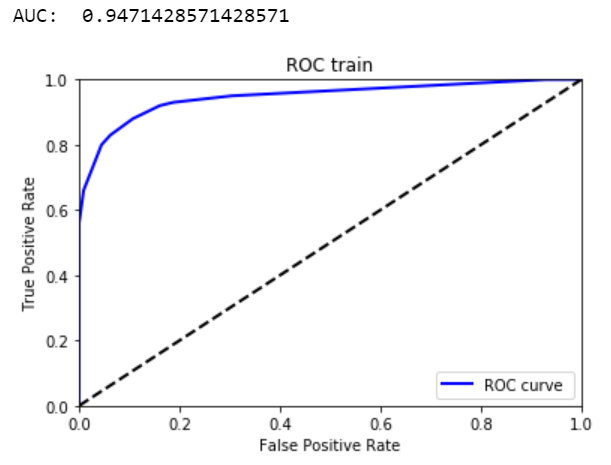
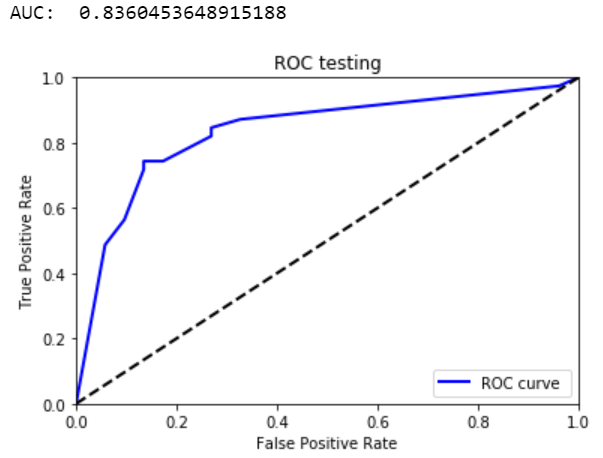
With the decision tree that give best performance, we perform those parameters on our dataset:



On a training set, the accuracy of the model is 88. 67% with the recall for class 1 (patients at risk) to be 83%. Area under curve for ROC of training set is around 94.71%, which means that the model can well distinguish between ‘0’ and ‘1’ – patients not at risk and patients at risk of having heart disease. However, on the test set, the accuracy of the model is 81.32% and the recall for class 1 (patients at risk) to be 74%. Area under curve for ROC of test set is 83.6%. Even though the accuracy and the area under curve for ROC of test set is relatively high, the recall is low (74%). For our problem, recall for class 1 (patients at risk) is an important evaluator since it indicates how sensitive the model is to predict class 1 and minimize the false prediction for patients who should have been at risk.

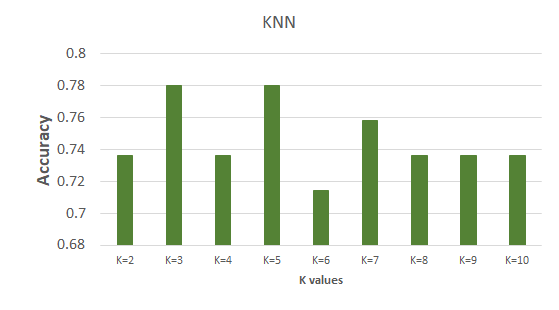
Following is the confusion matrix and ROC curve for training and testing set of the dataset:

## K-NEAREST NEIGHBORS

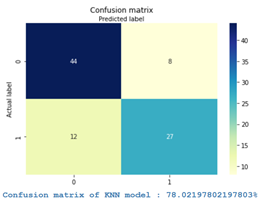
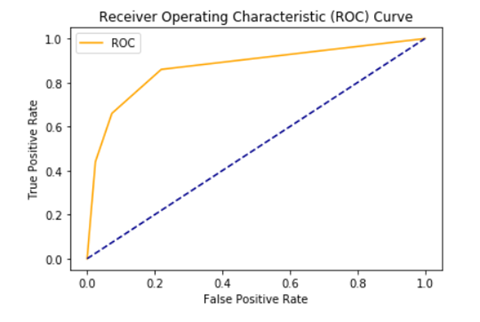
The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, a data has the same label to its k nearest neighbors.



We used multiple K values to check which value gives us the best accuracy. In our case K = 3 and 5 gives the best possible accuracy for this model. This accuracy is 78%. Therefore, we will pick k=3 for further evaluation in ROC and confusion matrix.

**ROC and Confusion Matrix KNN:**

ROC curve explains how good our model is. Higher the Area under curve, better the model.



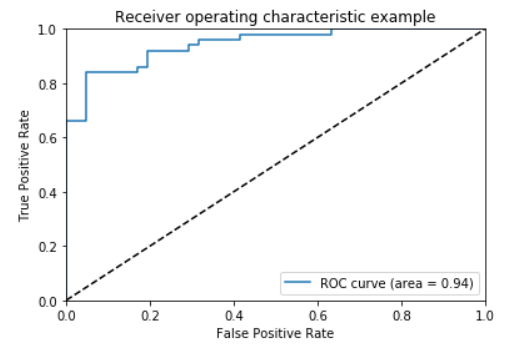
**KNN model predicts heart disease accurately by 78 percent**

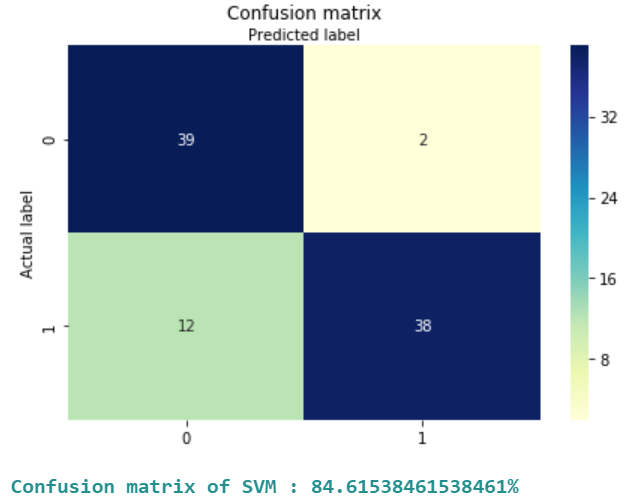
The AUC for the ROC curve for KNN model is 84%. Accuracy of the model is 78% and the recall for class 1 (patients at risk) is 69%. Even though the AUC of the ROC curve for KNN method is not too low, the accuracy and the recall for class 1 for KNN model is not high, indicating that KNN is not a good predictive model for our problem

## SUPPORT VECTOR MACHINE

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they’re able to categorize new data in testing set.

**ROC and Confusion Matrix KNN:**



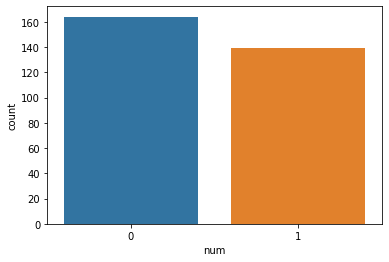


The AUC for the ROC curve for SVM model is 94%. Accuracy of the model is 85% and the recall for class 1 (patients at risk) is 76%. Even though the AUC of the ROC curve and accuracy for SVM method are high, the recall for class 1 for SVM model is low. In our problem, recall is the important indicator of a good model since it is more important to predict accurately people who have high risk of heart disease. For that reason, SVM may not be an ideal model for prediction in our problem

## LOGISTIC REGRESSION

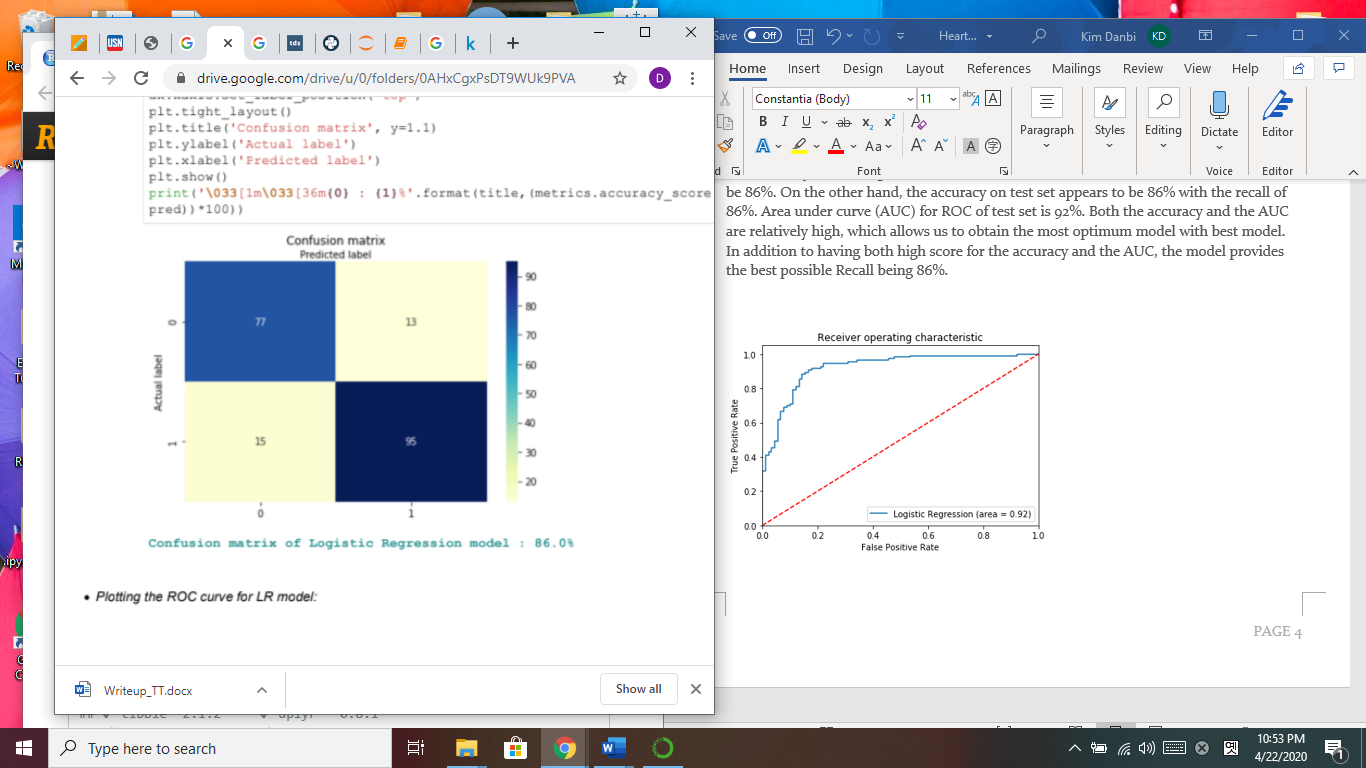
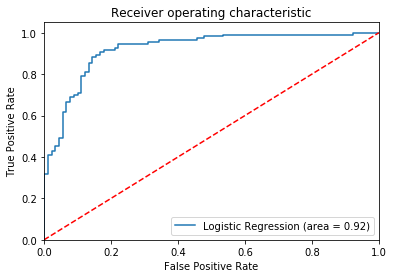
Logistic regression model is one of the basic Machine Learning classification algorithms that is fast and relatively uncomplicated. It is mainly used to predict and calculate the probability of categorical dependent variable. The method is for binary classification, so the dependent variable has to be a binary variable that contains data coded as 0 or 1, ‘No’ or ‘Failure’ and ‘Yes’, ‘Success’, respectively.

In this project, the response variable or dependent variable is ‘num’. We converted the variable ‘num’ into binary type, 0 and 1. We remapped the values of ‘num’, 1, 2, 3, and 4 into 1, which represents the presence of heart diseases.



In this graph, the number of 0, ‘No Presence’ is shown to be 164 while 1, ‘Presence of Heart Diseases’ is 139, meaning there are 164 patients with no heart disease and 139 with risk of heart disease.

The accuracy on training set comes out to be 84.75% with the recall for ‘Patients at risk’ to be 86%. On the other hand, the accuracy on test set appears to be 86% with the recall of 86%. The confusion matrix shows 77+95 = 172 correct predictions and 13+15=28 incorrect ones. Area under curve (AUC) for ROC of test set is 92%. Both the accuracy and the AUC are relatively high, which allows us to obtain the most optimum model with best model. In addition to having both high score for the accuracy and the AUC, the model provides the best possible Recall being 86%.

The images below are the ROC curve and the confusion matrix for test set:

In conclusion, the model predicted with 86% accuracy with the recall for ‘Patients at risk’ to be 86%. Also, the AUC for ROC curve is 92%, located towards the top left corner of the plot, where the sensitivity and specificity are at best levels. model is more specific than sensitive as well as very satisfactory.

# Conclusion

|  |  |
| --- | --- |
| **Model** | **Recall Percentage** |
| Decision Tree | 74% |
| KNN | 69% |
| SVM | 76% |
| Log Regression | 86% |

|  |  |
| --- | --- |
| **Model** | **AUC** |
| Decision Tree | 84% |
| KNN | 84% |
| SVM | 85% |
| Log Regression | 92% |

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Decision Tree | 81% |
| KNN | 78% |
| SVM | 94% |
| Log Regression | 86% |

In general, machine learning can help us better predict if patients are high risk of heart disease. After building models in data from the past, machine learning models can predict unseen data with high accuracy and recall and two classes are well separable. Logistic regression provides the highest recall and AUC among Decision tree, KNN and SVM models that we have built. SVM has the highest accuracy and logistic regression is the second highest for accuracy. KNN performed the worst among other models in term of recall and accuracy.

In order to find the best model, we have looked at the “Recall” measurement for class 1 (patients at high risk) of model prediction since it is more important to predict accurately patients with high risk of disease. “Recall” for class 1 is the model performance to predict positive class perfectly – class that has heart disease. We have found that the best model for our case is Logistic Regression. Logistic Regression provides 86 percent of the best possible Recall for “Patients at risk” with 92 percent of AUC for the ROC curve. Logistic Regression not only provides the best percentage of Recall, but also has the lowest model complexity, which facilitates in lower probability of overfitting. Therefore, in our problem, Logistic regression can be used to better predict if a patient is at risk.

# References

|  |  |
| --- | --- |
| [1] | National Center for Chronic Disease Prevention and Health Promotion, "Centers for Disease Control and Prevention," 2 December 2019. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed 29 April 2020]. |