**FUNDAMENTALS OF MACHINE LEARNING**

**FINAL ASSIGNMENT REPORT NAME: NITHIN VARMA**

**EMAIL:** [**nchinta2@kent.edu**](mailto:nchinta2@kent.edu)

**OVERVIEW**

In this project, we will work closely with the prediction of heart disease. To do this, we will examine the heart disease dataset. From this dataset, we will derive various insights that help us understand the importance of each feature and how they relate to one another, but this time, our only goal is to determine whether a person is likely to experience a heart disease problem or not.

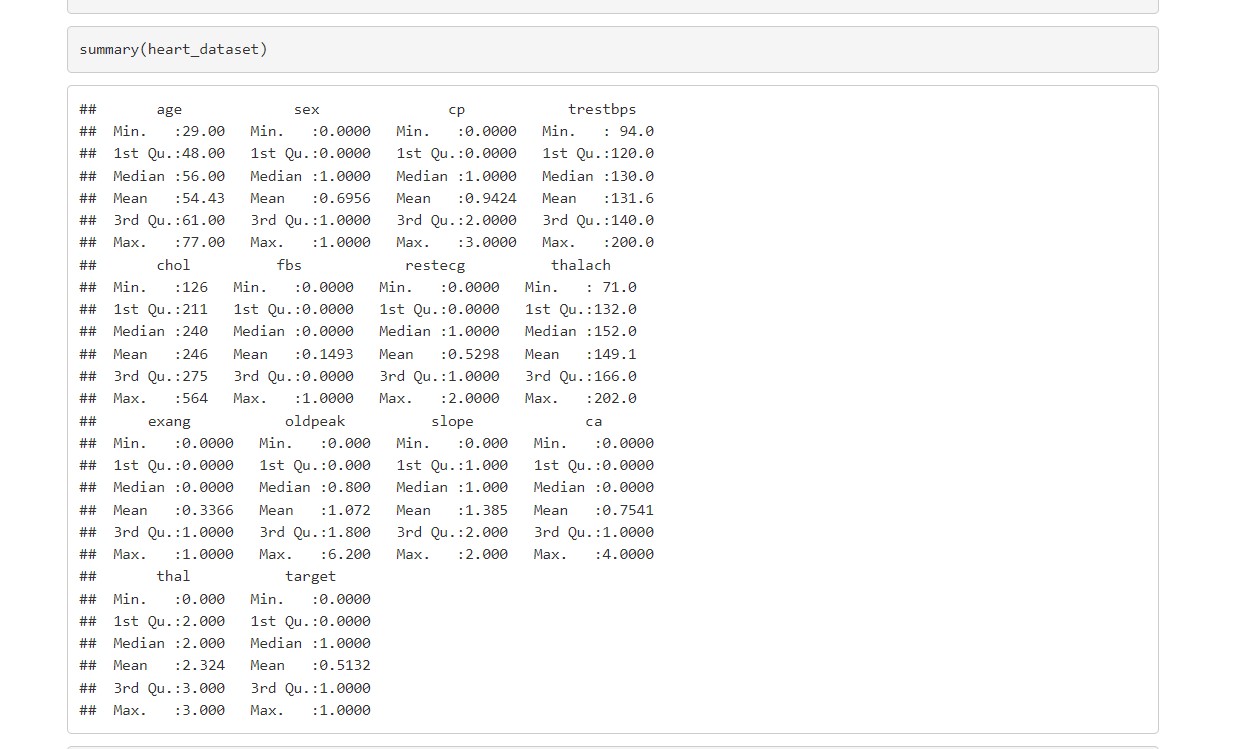
**OBJECTIVES**

* Machine learning may be crucial in determining whether heart diseases, and other conditions are present or absent. If foreseen far in advance, such information can offer valuable insights to clinicians, who can then customize their diagnosis and course of care for each patient.
* To know which model accuracy is perfect for the dataset and predict the heart disease
* The objective is to determine whether the patient is suffering with heart disease or not . I used classification techniques from machine learning to fit a function that can predict the discrete class of new input to accomplish this.

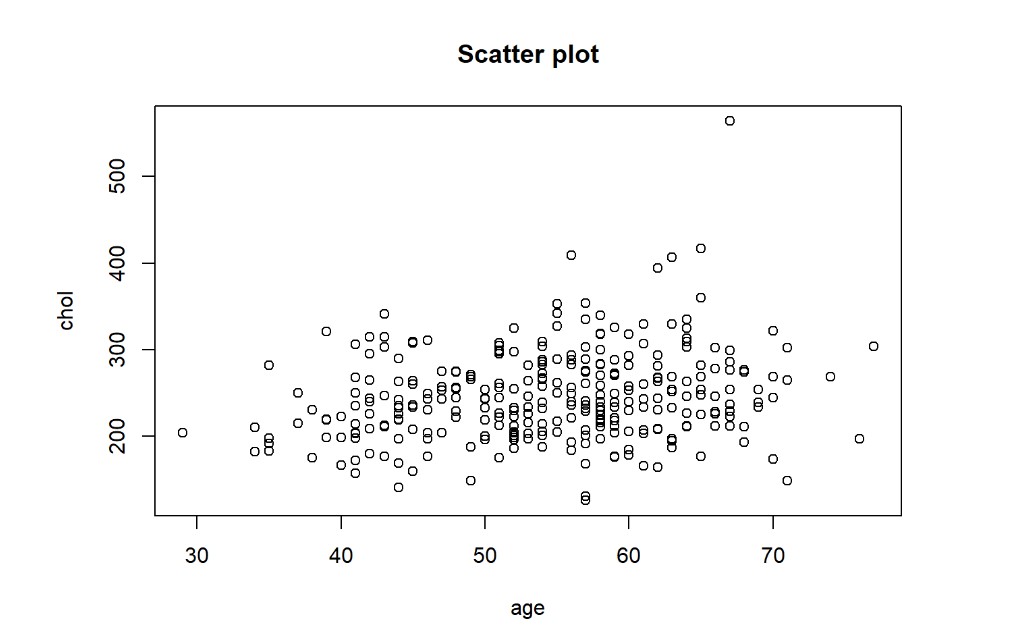
**ATTRIBUTES IN THE DATASET**

1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
   * 0: Typical angina: chest pain related decrease blood supply to the heart
   * 1: Atypical angina: chest pain not related to heart
   * 2: Non-anginal pain: typically oesophageal spasms (non heart related)
   * 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholesterol in mg/dl
   * serum = LDL + HDL + .2 \* triglycerides
   * above 200 is cause for concern
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
   * '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
   * 0: Nothing to note
   * 1: ST-T Wave abnormality
     1. can range from mild symptoms to severe problems
     2. signals non-normal heart beat
   * 2: Possible or definite left ventricular hypertrophy
     1. Enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak  - ST depression induced by exercise relative to rest looks at stress of heart during excise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    * 0: Upsloping: better heart rate with exercise (uncommon)
    * 1: Flat sloping: minimal change (typical healthy heart)
    * 2: Down sloping: signs of unhealthy heart
12. ca - number of major vessels (0-3) coloured by fluoroscopy
    * coloured vessel means the doctor can see the blood passing through
    * the more blood movement the better (no clots)
13. thal - thallium stress result
    * 1,3: normal
    * 6: fixed defect: used to be defect but ok now
    * 7: reversable defect: no proper blood movement when exercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

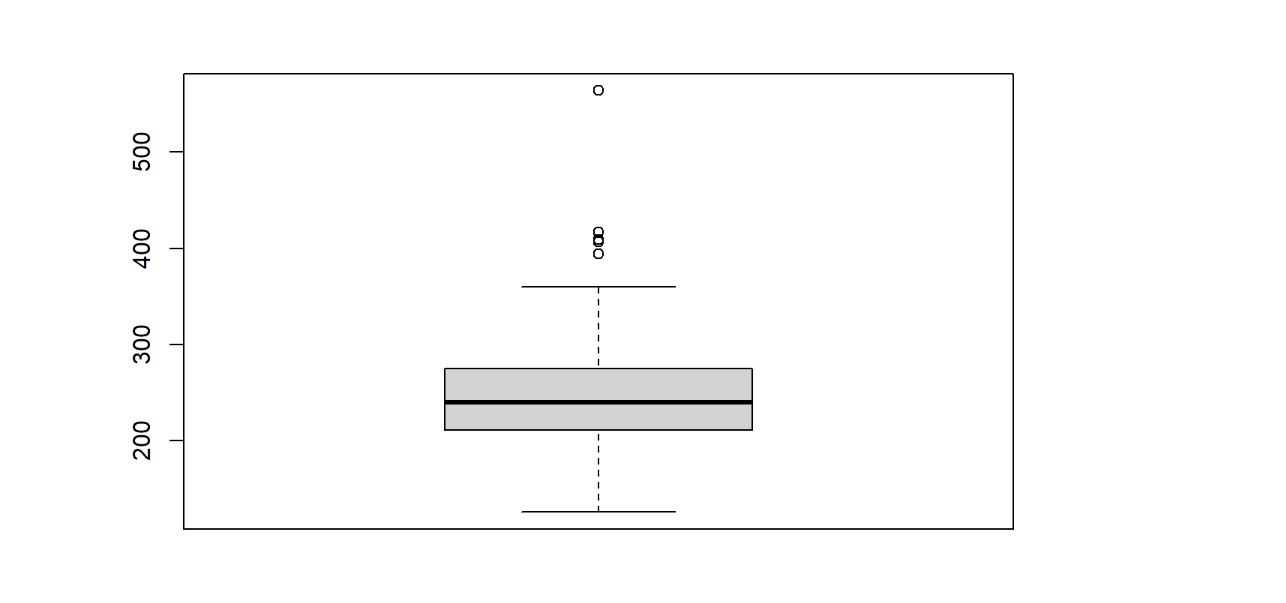
**SUMMARY OF THE DATASET**



**GRAPHICAL PRESENTATION**



This is the Scatter plot showing age wise cholesterol in a person .



This is the boxplot of the variable chol (cholesterol) . The black line is the median between the 1st quartile and 3rd quartile range and the small circles above are the possible outliers .

**ANALYSIS OF THE PROJECT**

DATA PARTITION



Once there are no missing values in the data, we divide it into training and test sets using the CARET tool in R. "0.8" is the partition index, which refers to the training set, which makes up 80% of the total dataset, and the test set, which makes up 20%

DETAILS OF MODELLING STATEGY

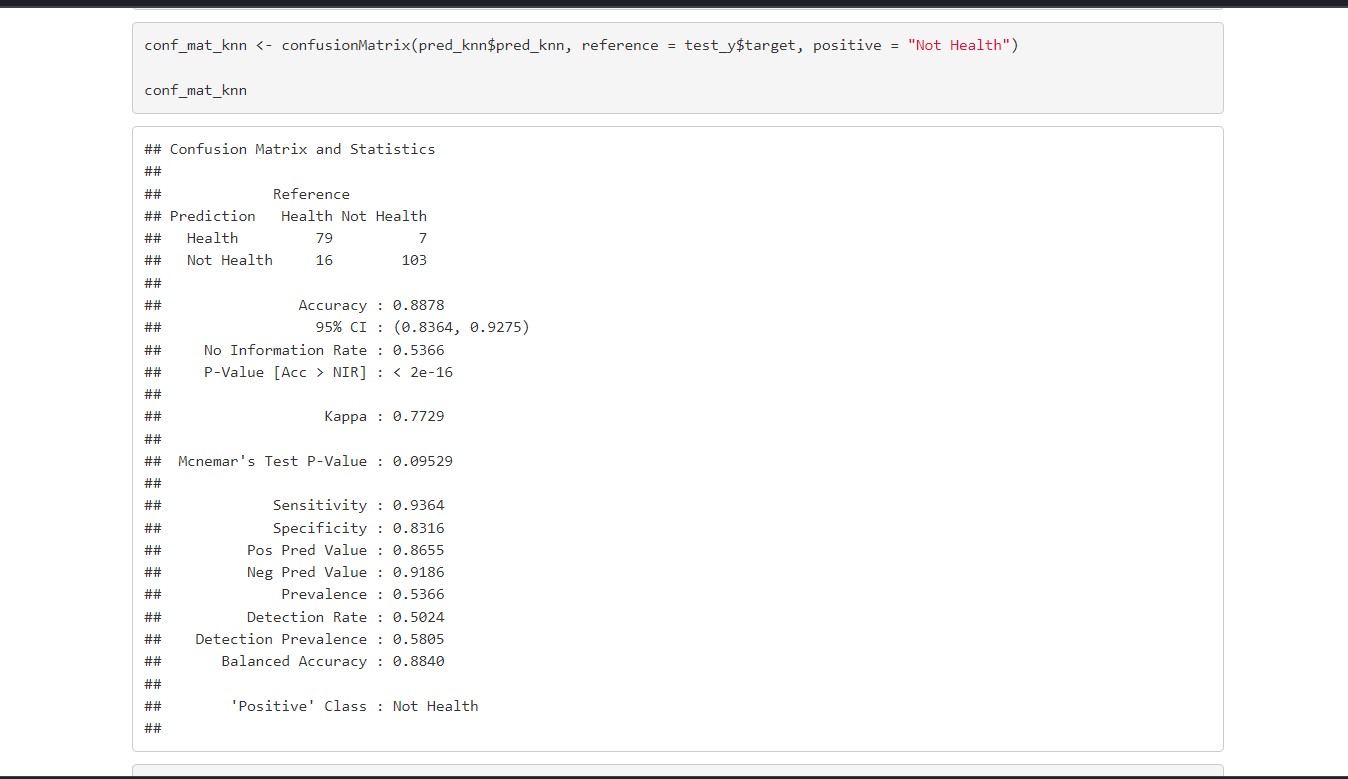
It is defined as the method used to extract the important variables and comprehend how the data should be comprehended in order to obtain the best analysis from the data.

As a sample for the training and validation sets, which are 80 and 20 respectively, we divided the given data into two equal halves. The dataset was then evaluated using Knn and decision trees.

BUILDING A Knn MODEL



CONFUSION MATRIX FOR Knn MODEL



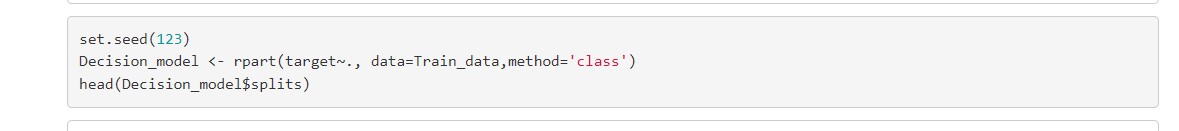
\* Summary of the Confusion matrix of the Knn model

1)Accuracy = 0.8878

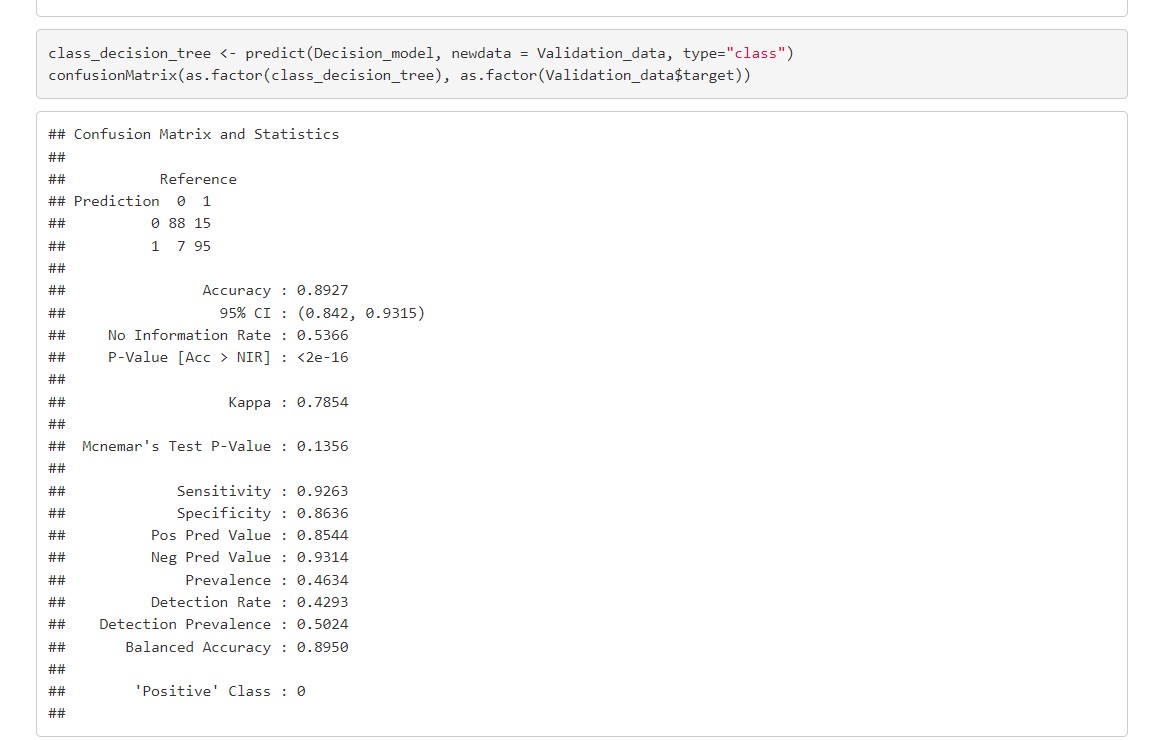
2)Sensitivity = 0.9364

3)Specificity = 0.8316

BULDING A DECISION TREE MODEL



CONFUSION MARTIX FOR THE DECISION TREE MODEL



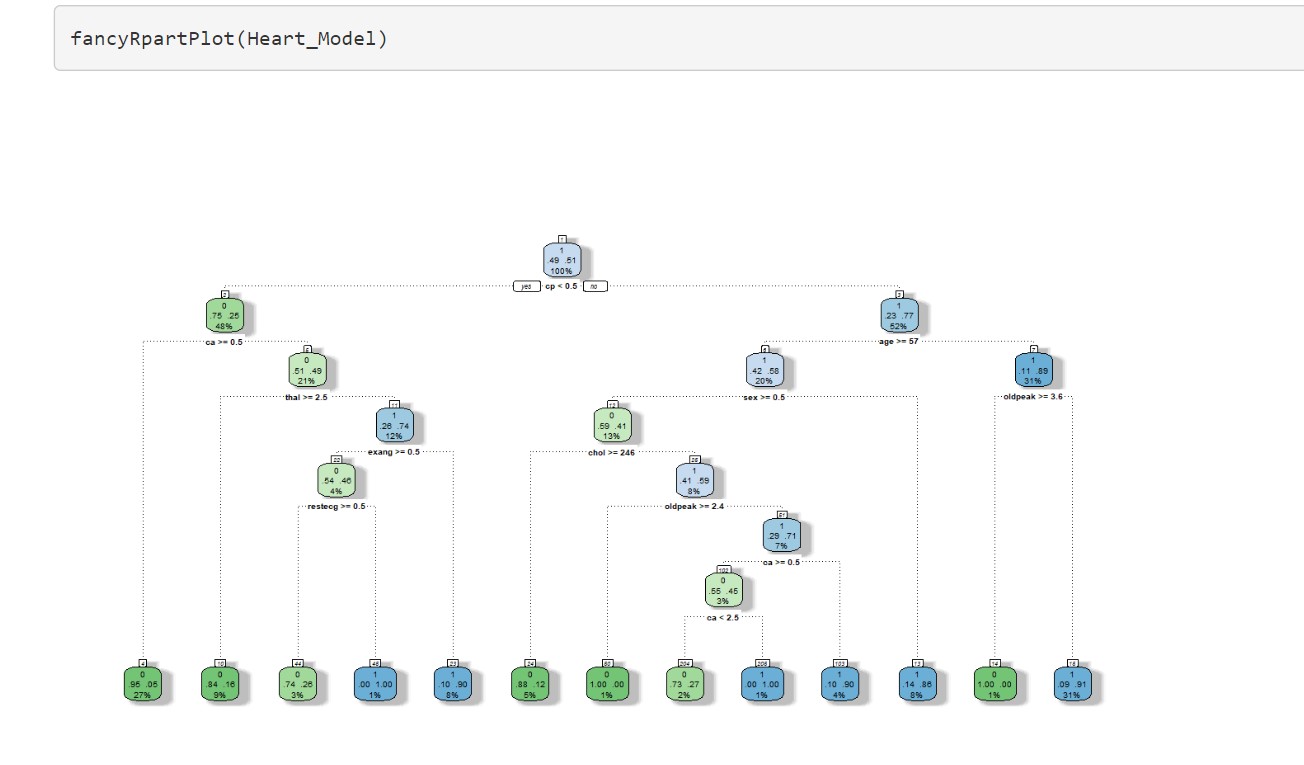
\* Summary of the Confusion Matrix of Decision Tree model

1) Accuracy = 0.8927

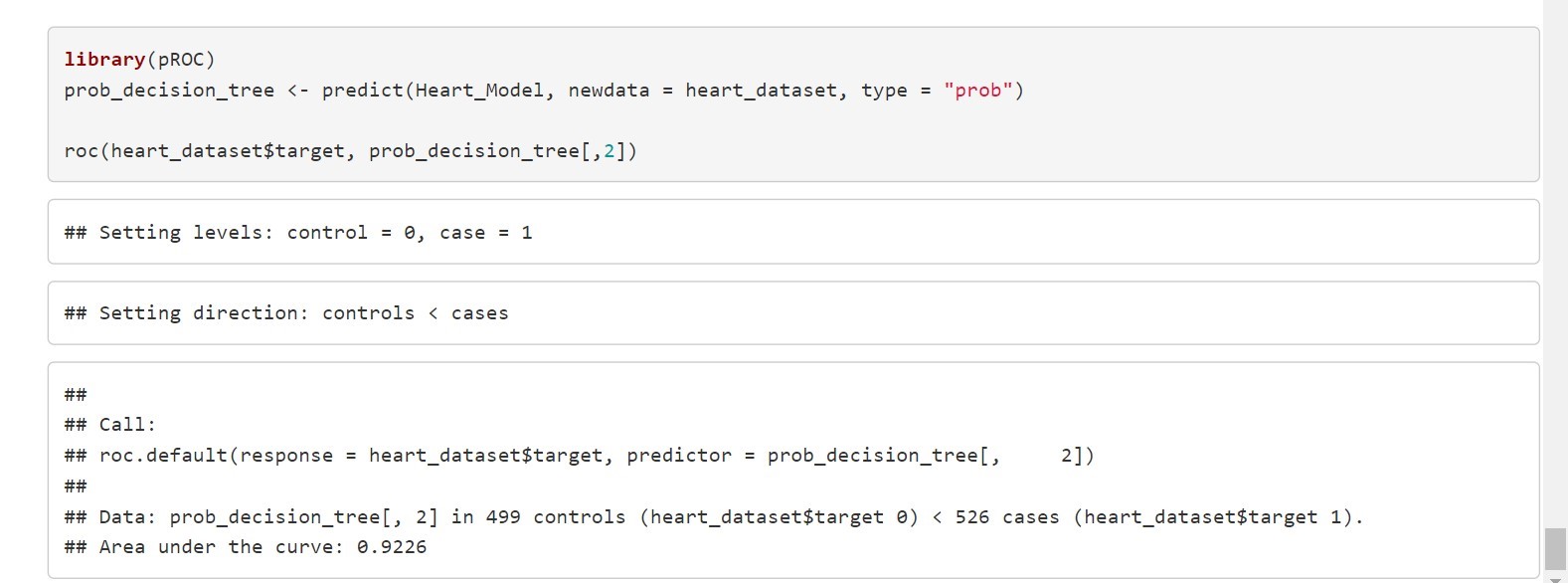
2)Sensitivity = 0.9263

3) Specificity = 0.8636

DECISION TREE FANCY RPLOT



pROC FOR DECISION TREE MODEL



**So , The value of AUC (Area under the curve) is 0.9226 for the decision tree model**

**THE RESULTS**

* Based on our tolerance for false negatives and desire for genuine positives, we apply the model to the test data to produce Accuracy in order to assess the algorithm's efficacy and choose the optimal threshold.
* Therefore, the greater the Accuracy , the more confidence we are in the prediction power of our model.

|  |  |  |  |
| --- | --- | --- | --- |
| Knn Model | Accuracy = 0.8878 | Sensitivity= 0.9364 | Specificity=0. 8316 |
| Decision tree model | Accuracy=0.8927 | Sensitivity=  0.9263 | Specificity = 0.8636 |

**INSIGHTS AND CONCLUSIONS**

Therefore by this, In terms of Accuracy and Specificity, Decision Tree is the best model to apply for the dataset .The prediction of the heart disease is more accurate when we use the decision tree model . So we choose Decision Tree model to apply on the test data as the model has the highest accuracy among others.

If a doctor wanted to give a better treatment for people diagnosed with heart disease and thus who do not, I will use a method with a better Accuracy value.

**REFERENCES**

**THIS IS THE LINK WHERE I DOWNLOADED THE DATASET**

<https://www.kaggle.com/code/faressayah/predicting-heart-disease-using-machine-learning/data>

**THIS IS WHERE I TOOK THE REFERENCE OF THE CODE**

<https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c#:~:text=Machine%20Learning%20can%20play%20an,and%20treatment%20per%20patient%20basis>.