**“Team VIRUS” presents**

**KIDNEY DISEASE ANALYSIS**



Submitted By:

Sushniv Vangari

K. Hrushikesh Koundinya

B. Varun Kumar

S. Sri Ram

K. Sai Gurunadh

**Table of Contents**

1. Title of the Project

1.1 - Introduction

1.2 - Objectives of Research

1.3 - Problem Statement

1. Review of Literature
2. Data Collection
3. Methodology

4.1 - Exploratory Data Analysis

4.1.1 - Figures and Tables

4.2 - Data Modelling

1. Findings and Suggestions
2. Conclusion

Introduction to Machine Learning

* Machine learning is an application of artificial intelligence(AI) that provide system the ability to automatically learn and improve from past experience on its own.
* Machine learning focuses on the development of computer programs that can access data.

Types of Machine Learning:

Supervised Learning:

* It is indicated as a presence of supervisor as a teacher.
* Supervised learning is a learning in which we teach or train the machine using data which is well labeled i.e, some data is already tagged with the correct answer.

Unsupervised Learning:

* It is indicated there is no supervisor as a teacher.
* Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

Semi-supervised Learning:

* In the previous two types, either there are no labels for all the observation in the dataset or labels are present for all the observations.
* Semi-supervised learning falls in between these two. In many practical situations, the cost to label is quite high, since it requires skilled human experts to do that.

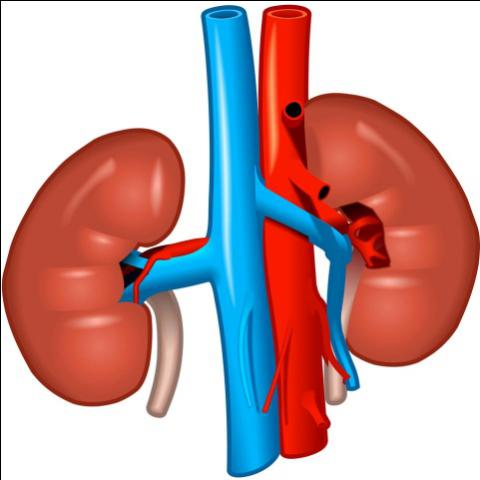
Reinforcement Learning:

* Reinforcement Learning is also a type of Machine Learning.
* It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance.
* Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal.

**KIDNEY DISEASE ANALYSIS**

ABSTRACT

* Our dataset focuses on chronic kidney disease. Dataset consists of both numeric as well as categorical data. Certain features includes Id, age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcv ,wc, rc, htn ,dm, cad, appet, pe, ane and classification.
* So, Initially we have cleaned the missing values using imputation techniques like mean imputation and mode imputation.
* In the next step, we performed Exploratory Data Analysis to analyze the data in a better way. While analysing the data we also drawn distplots, statmodels, boxplots to find the feature which majorly affecting the classification.
* Finally we also applied several models/algorithm like KNN classifier, Support Vector (SVM), Logistic Regression, Support Vector Classification (SVC) on the dataset and found that is the best model.



INTRODUCTION

* WHAT IS KIDNEY DISEASE?

Kidney disease is when your kidneys are damaged and not functioning as they should. When kidney disease is not going away it is called chronic kidney disease or CKD. When the kidney is suddenly injured that is called acute kidney injury or AKI.

* TESTS TO DETECT OR DIAGNOSE KIDNEY DISEASE

1. e-GFR - A blood test called e-GFR (estimated Glomerular Filtration Rate) tells you and your doctor how well your kidneys remove wastes from your body. Normal e-GFR is over 100. An e-GFR of less than 60 may be because of kidney disease.
2. Urine protein - Protein is natural in your body but is not normally found in the urine. A test can be done to check for protein in a urine sample. Too much protein could be a sign of damage to the kidney
3. High blood pressure - High blood pressure may be a sign and a cause of kidney disease.
4. Ultrasound - A picture of your kidneys taken with an ultrasound machine can check the size of your kidneys and for cysts or kidney stones. This can help your doctor tell if blood and urine are flowing through your kidney normally.

* TREATMENT STRATEGIES FOR KIDNEY DISEASE

1. Blood pressure - People with chronic kidney disease should discuss with their doctor the best blood pressure goal for them. For some people, less than 130/80 is recommended to prevent progression of kidney damage.

2. Medicines - Patients with chronic kidney disease should receive an angiotensin converting enzyme inhibitor (ACEI) or an angiotensin receptor blocker (ARB) to slow progression of kidney disease, if recommended by their doctor.

3. Healthy Lifestyle - Patients with kidney disease should not smoke, should be physically active and maintain the weight recommended by their doctor.

4. Managing other complications - Patients with chronic kidney disease should be evaluated and treated for complications related to kidney disease such as anemia and mineral and bone disorders.

**Importing packages**

* Numpy
* Pandas
* Imputer
* Seaborn
* Matplotlib
* Sklearn
  + Logistic Regression
  + SVC
  + Decision Tree Classifier
  + Random Forest Classifier

OBJECTIVES OF RESEARCH:

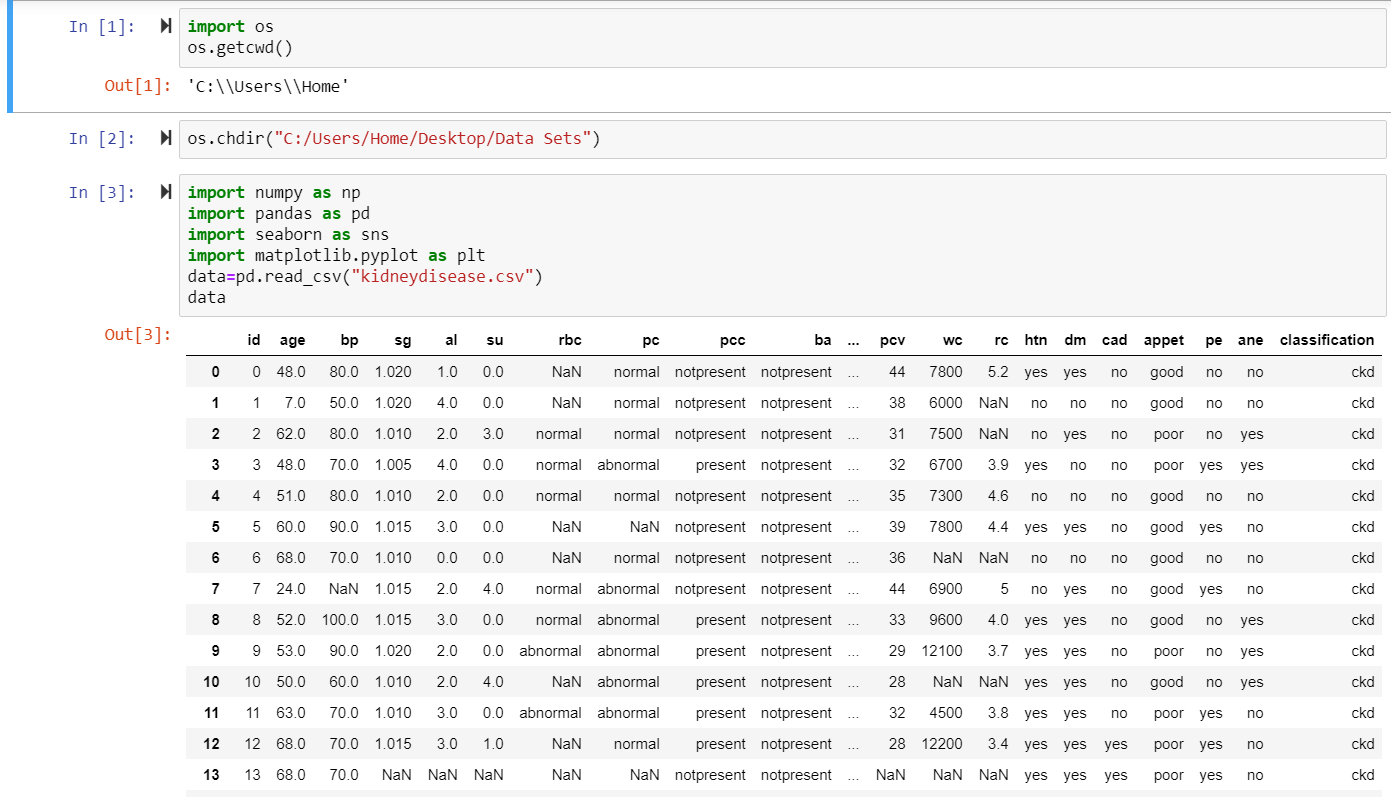
* To identify the classification of the disease based on the features given in the dataset by performing certain techniques and models by drawing plots, tables…etc.
* Enormous data mining techniques are existing for predicting **diseases** namely classification, clustering, association rules, summarizations, regression and etc. The main **objective** of this **research** work is to predict **kidney diseases** using classification algorithms such as Naive Bayes and Support Vector **Machine**.

PROBLEM STATEMENT:

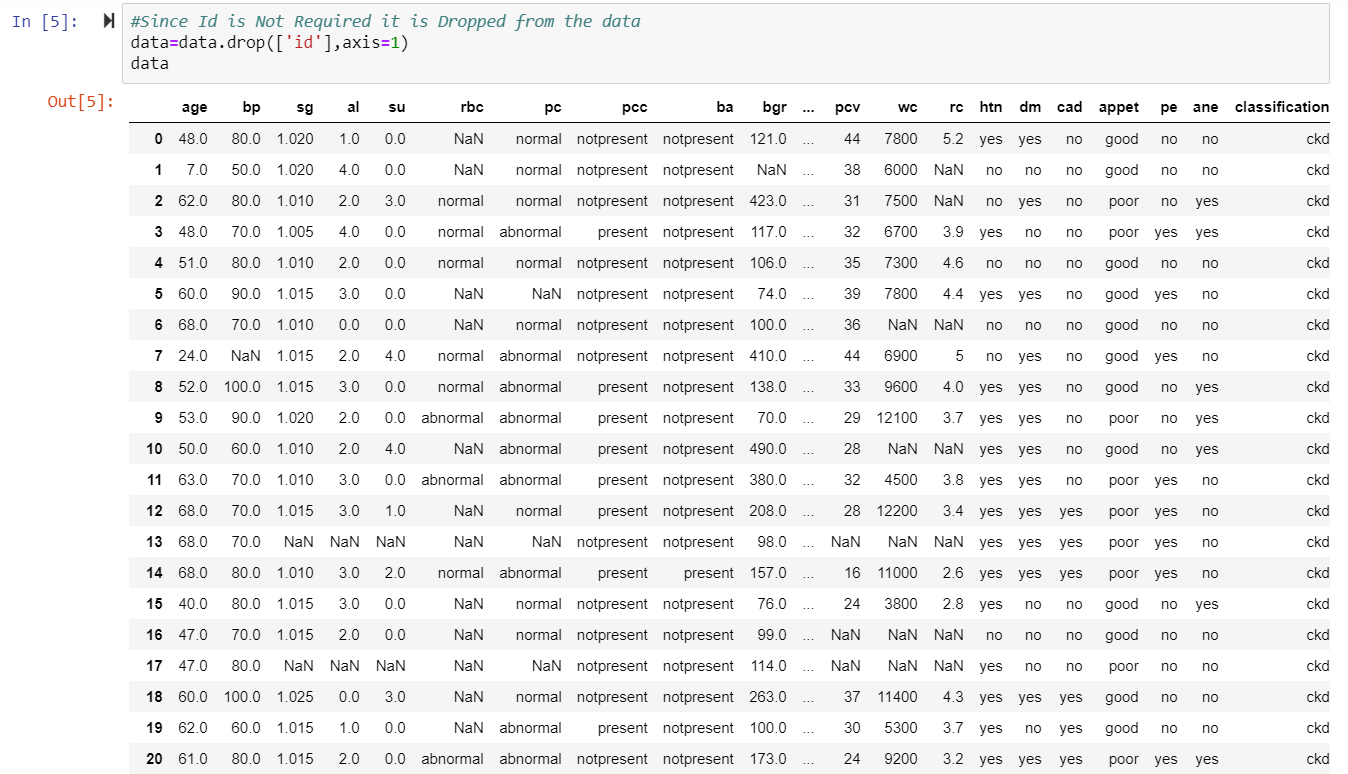
* **Chronic kidney disease (CKD) is one of the major public health issues with rising need of early detection for successful and sustainable care.**
* **There are many factors such as blood pressure, diabetes and other disorders that contribute to gradual loss of kidney function over time. CKD is gradual progression hence, it has many stages for criticality of disease.**
* **Additional data points such as GFR (Glomerular Filtration Rate), current medication and other progressive parameters diagnosis the symptoms, medical conditions, basic lab tests and a class indicating positive CKD.**
* **The goal for this example is to understand the approach that can be applied to many other use cases and real-world data.**
* **The data set may include missing data points; as a part of data preparation, we can remove those attributes or rows with more than 15 percent missing value.**

**Exploratory Data Analysis**

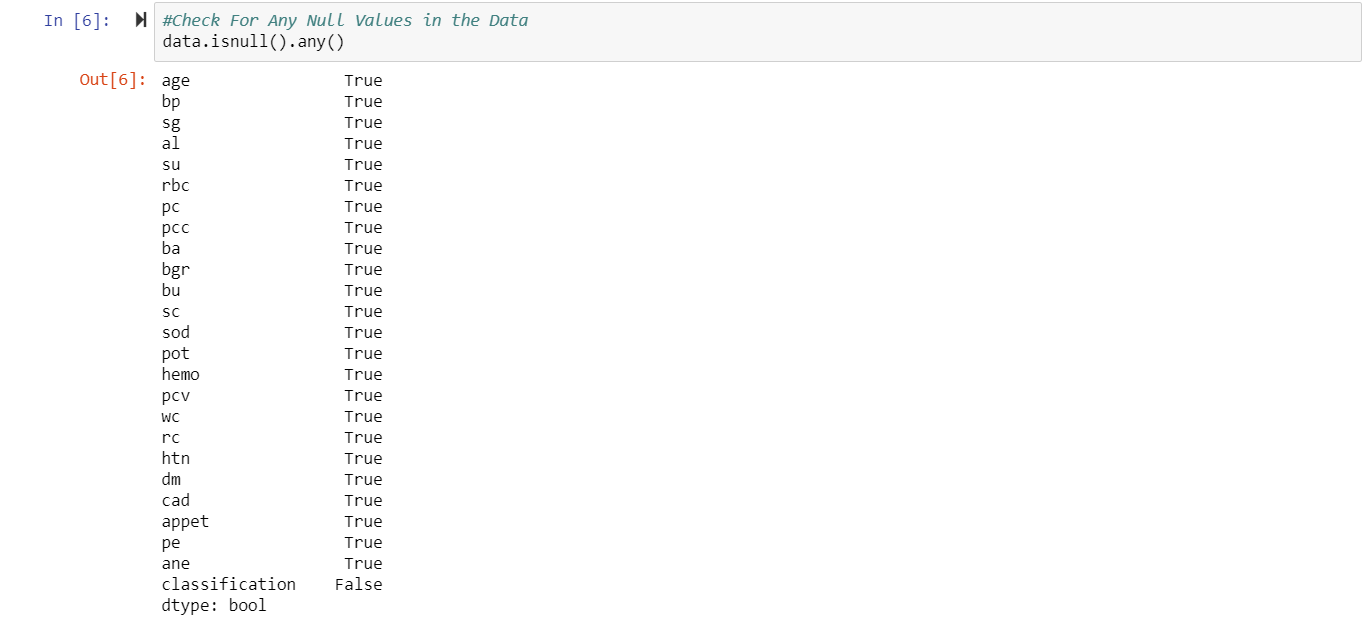
**Reading data from dataset**:-



**Delete ID:-**

****

**Checking null values:**



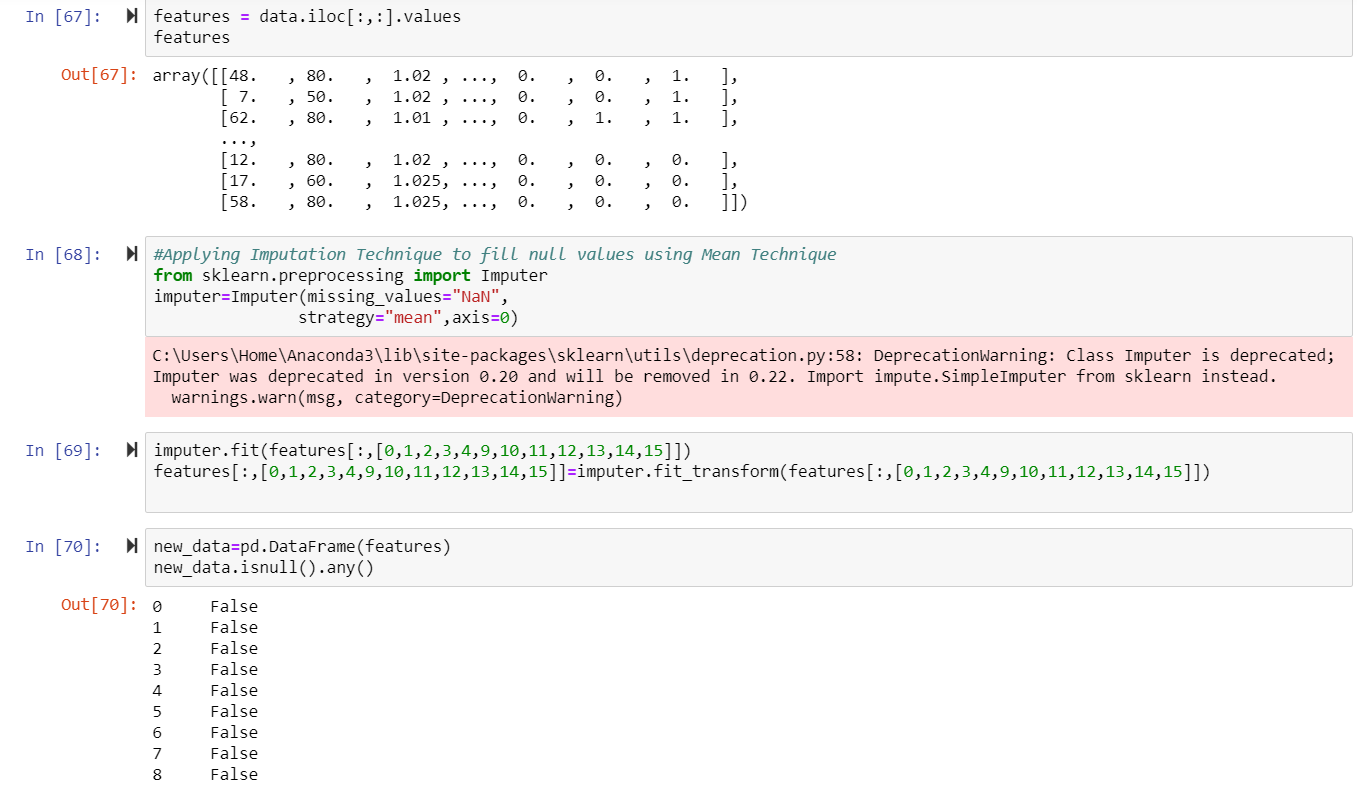
**Filling the null values and replacing:**



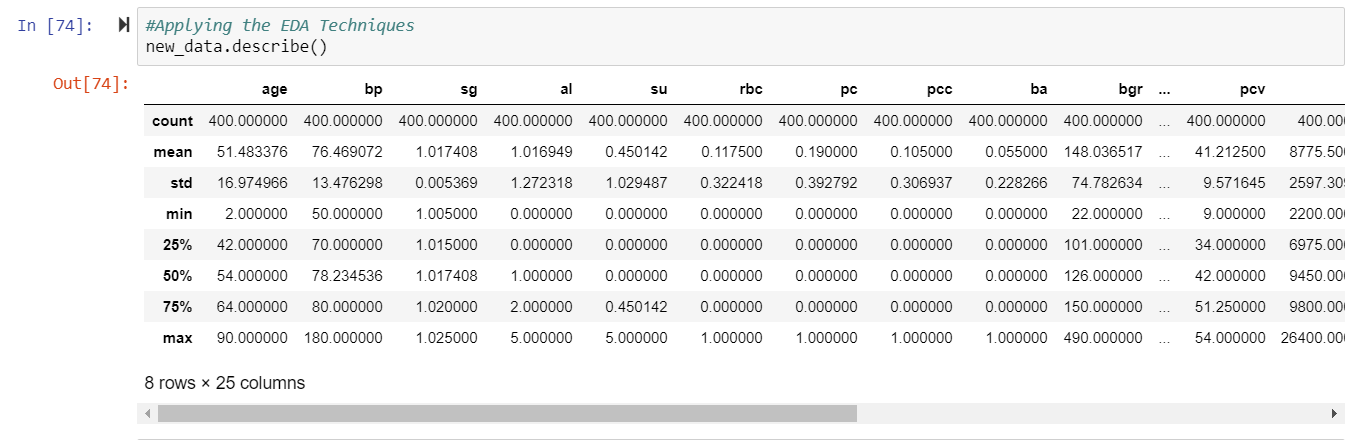
**Converting objects into floats:**



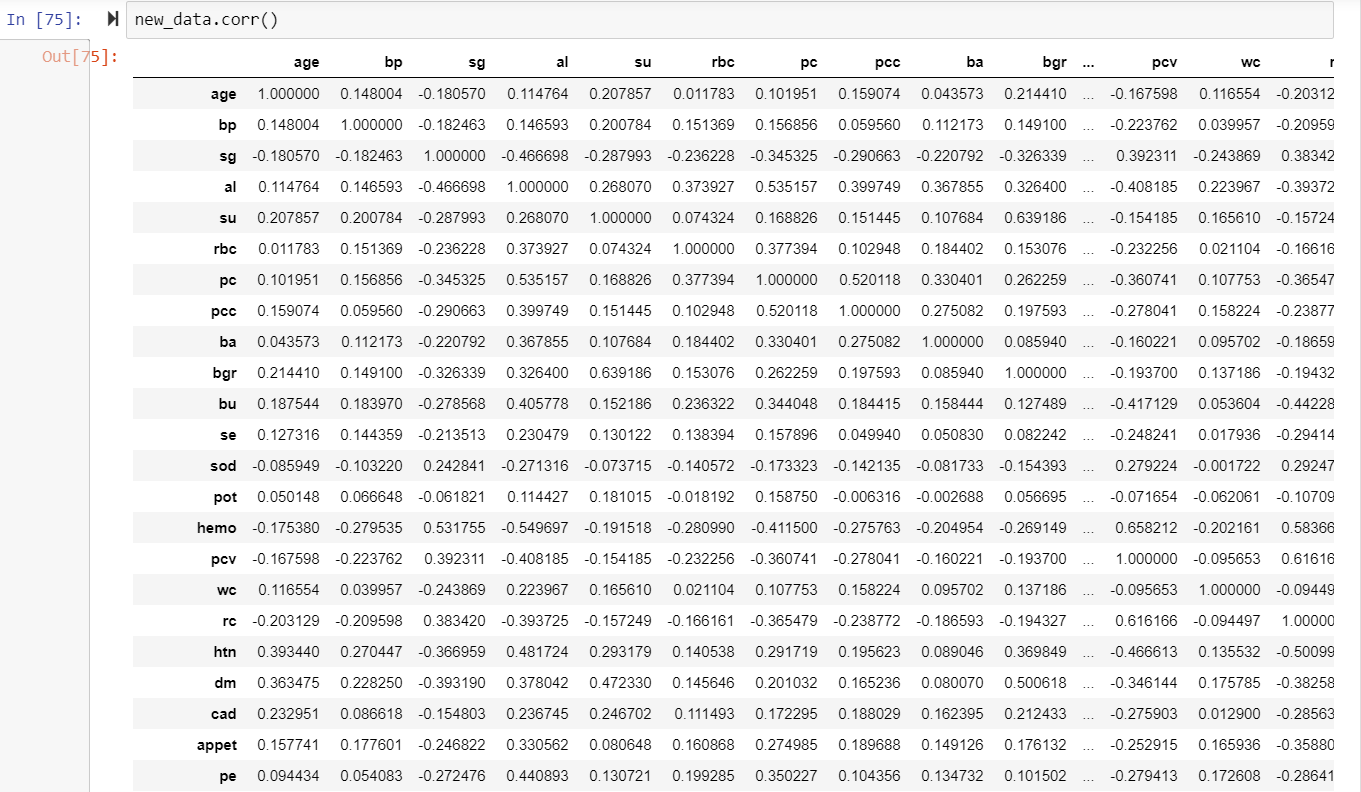
**Filling null values in categorical type using Imputer:**



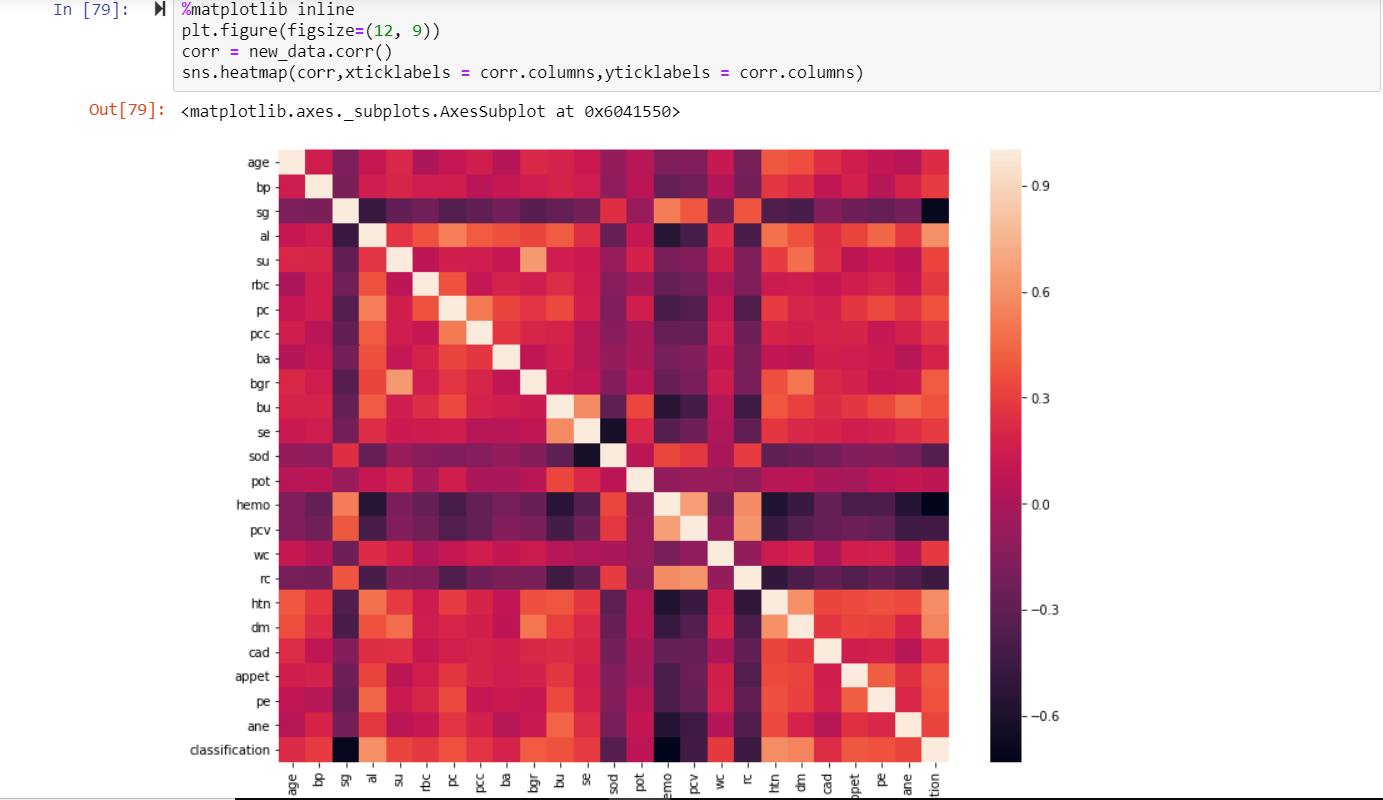
**Describing the new data:**



**Data Correlation:**

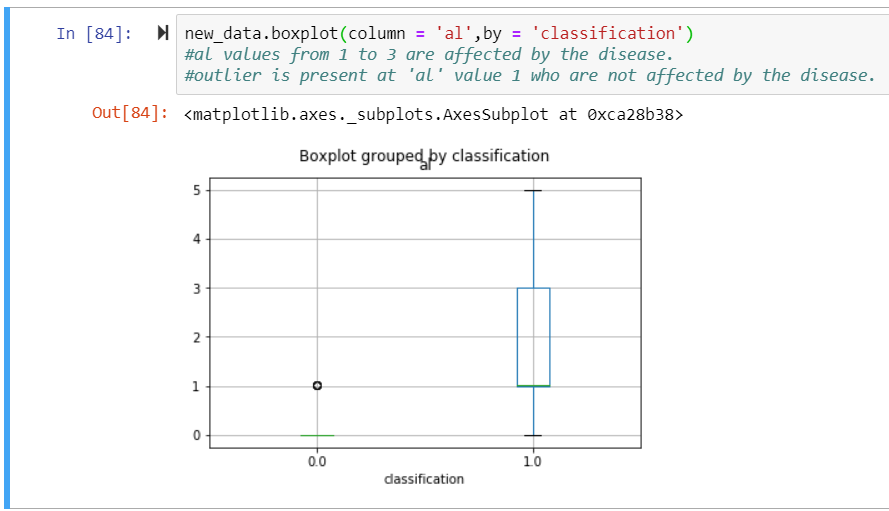
****

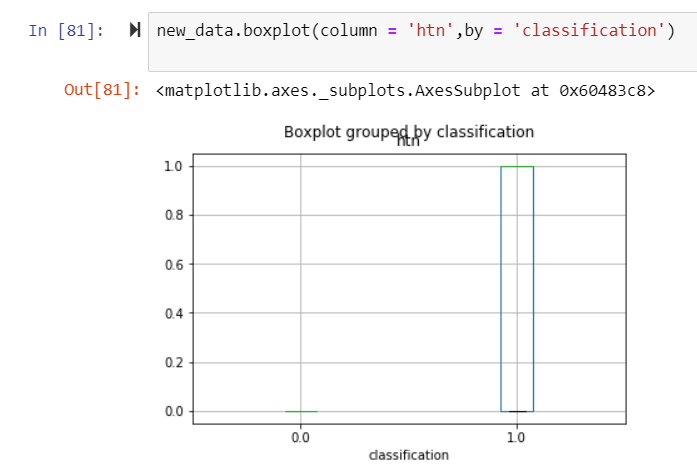
**Heat Map:**

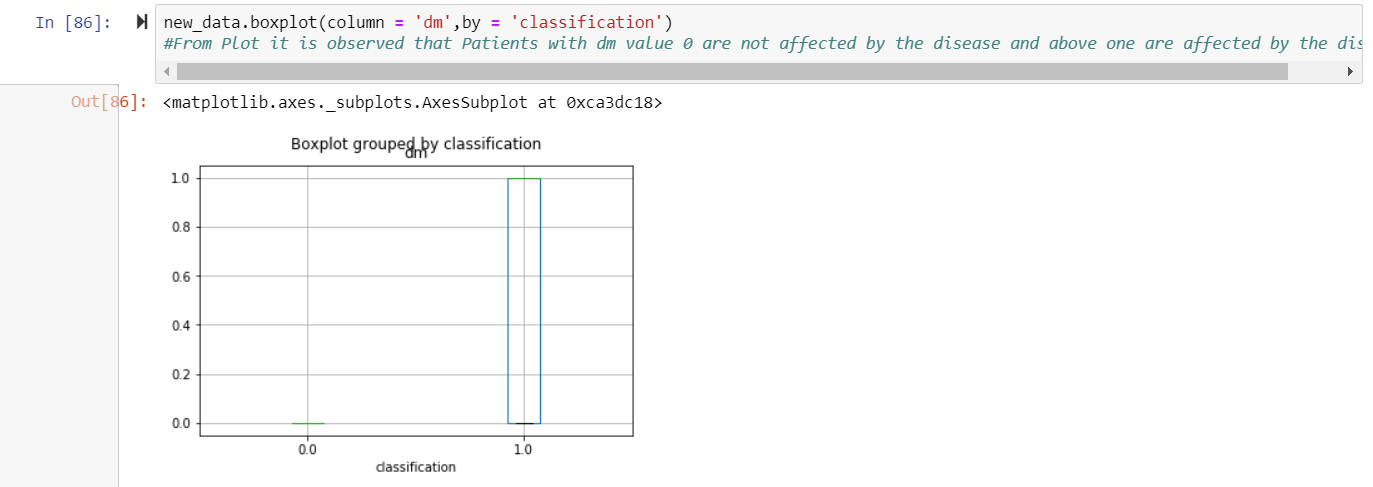
****

* **According to Heat Map It can Be Observed That sg, sod, hemo, pcv, rc are negatively correlated with classification.(i.e With their increase disease decreases)**
* **al, htn, dm are positively correlated with classification.(i.e With their increase disease increases)**

**Boxplot graphs (htn, al, dm, sg):**

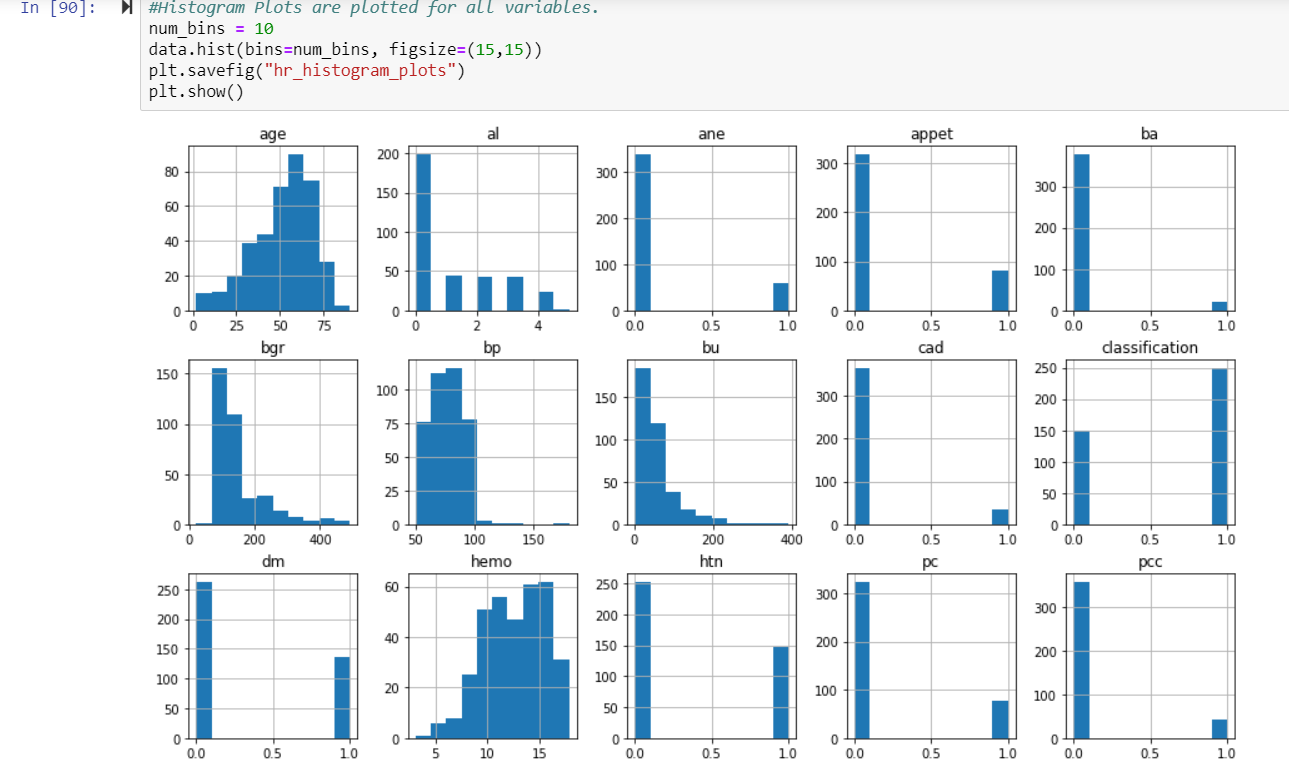
****

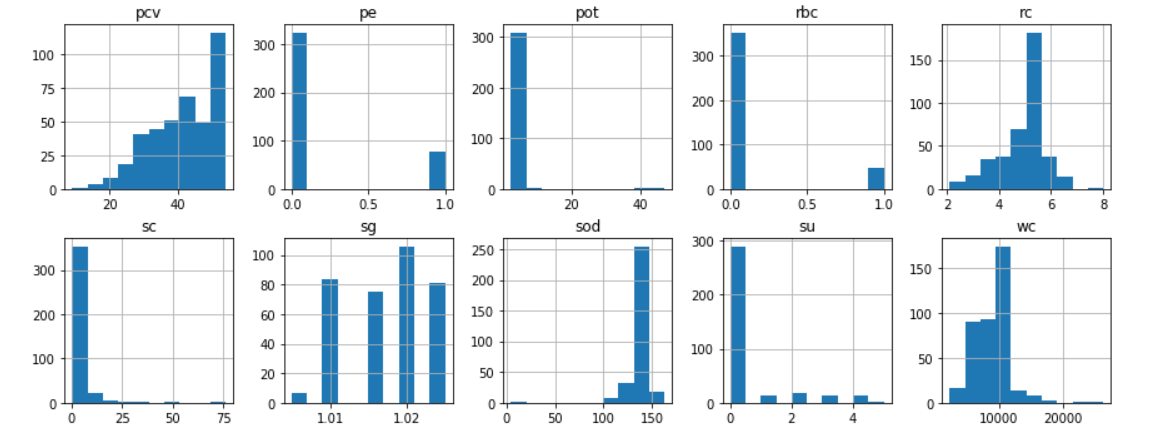
****

****

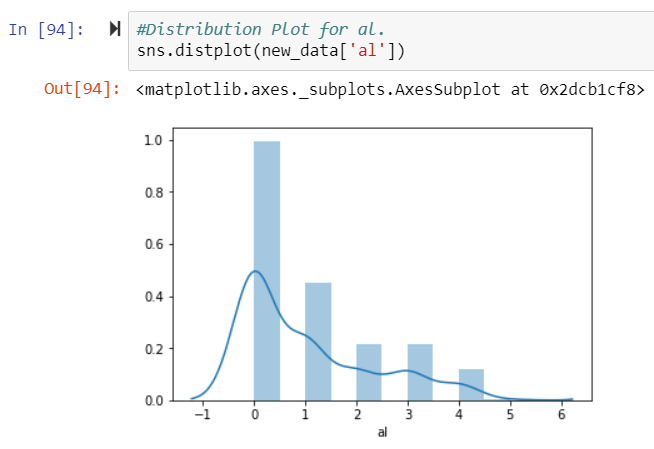
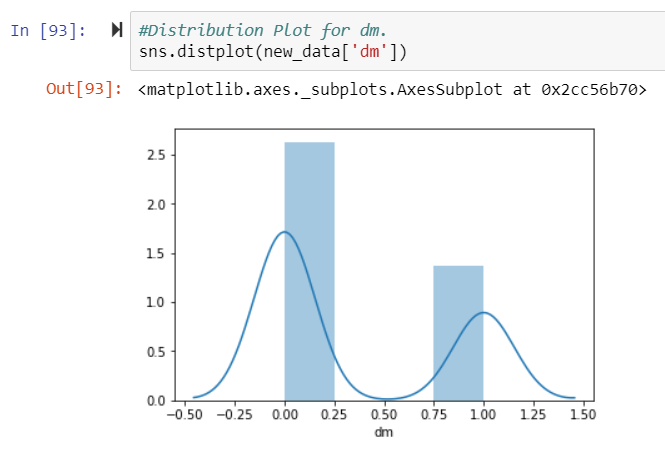
****

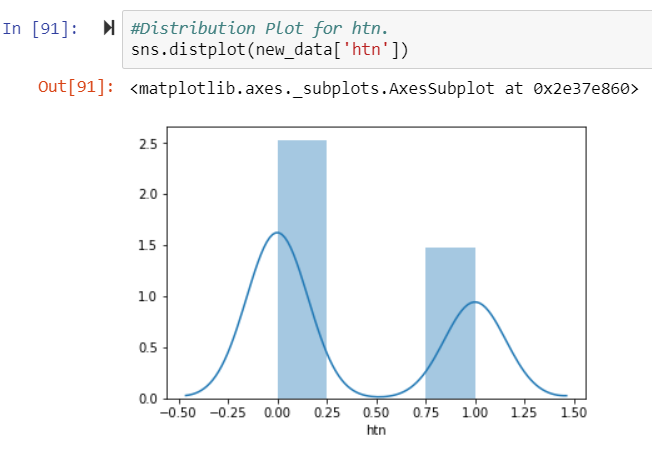
**Histograms:**

****

****

**Distribution Plot (dm, al, htn):**

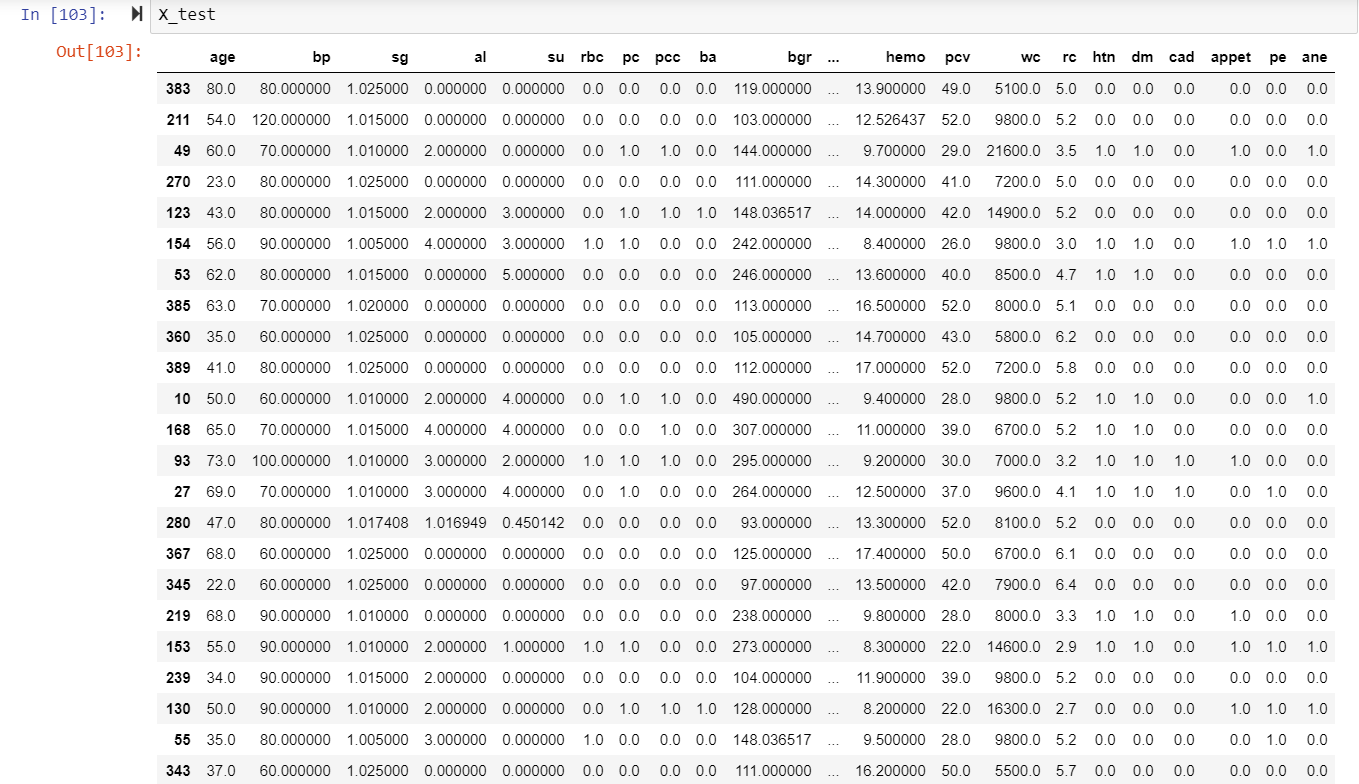
****

****

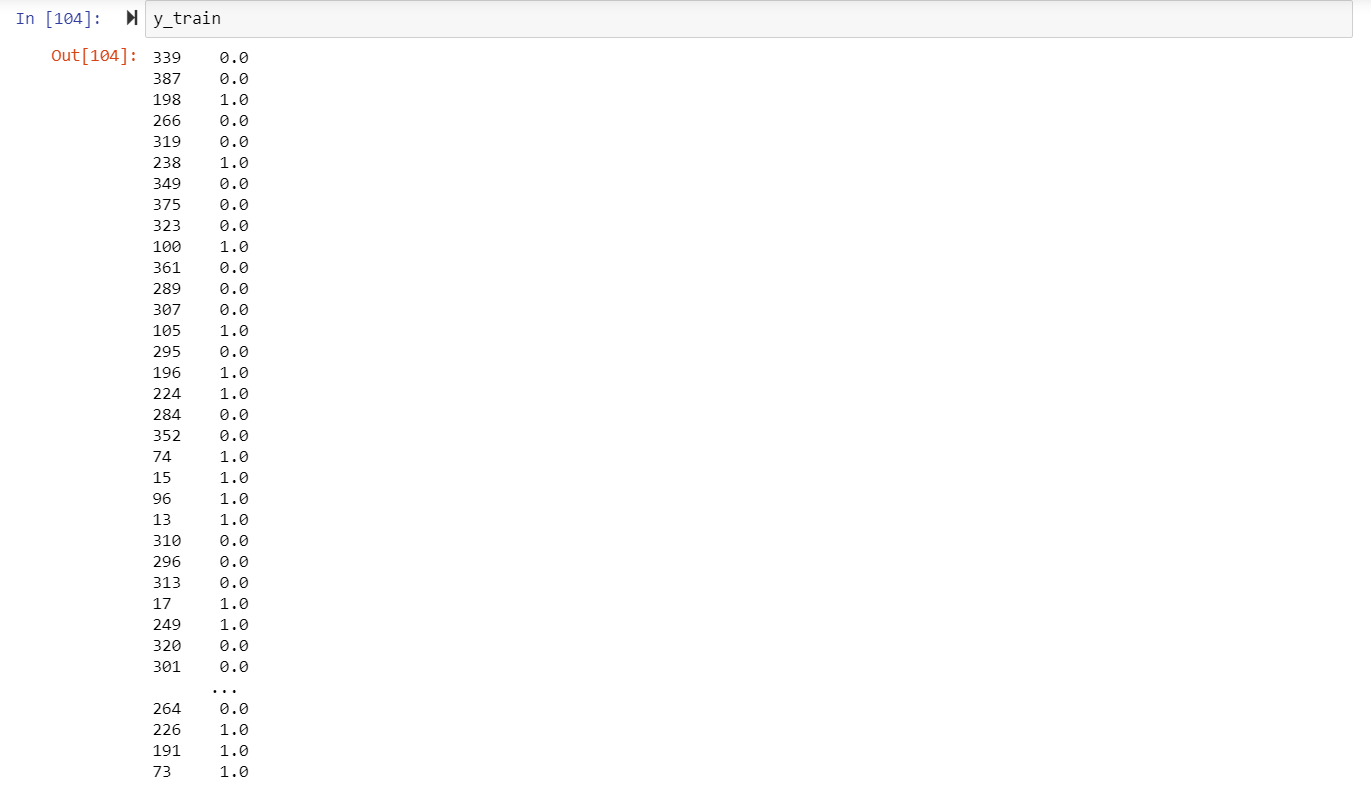
**Normalization of Data:**

****

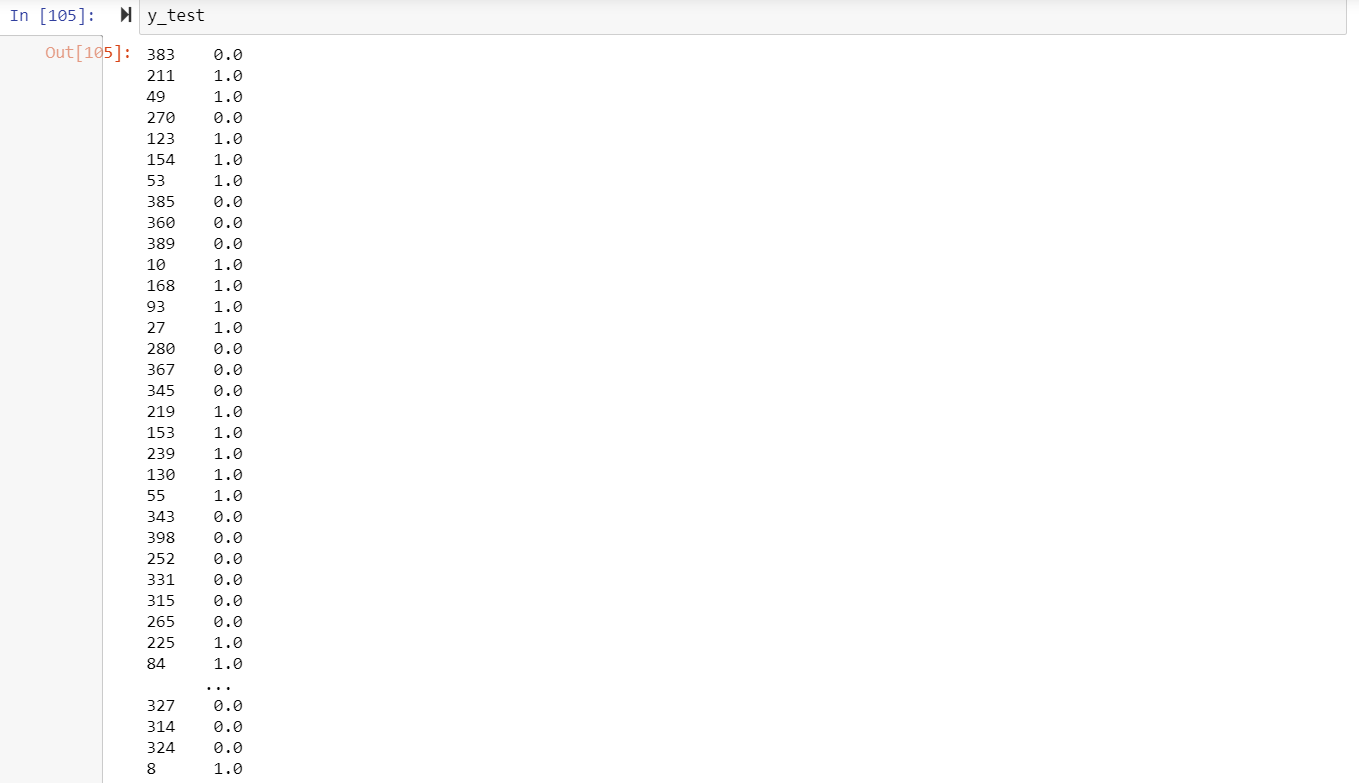
**X-test:**

****

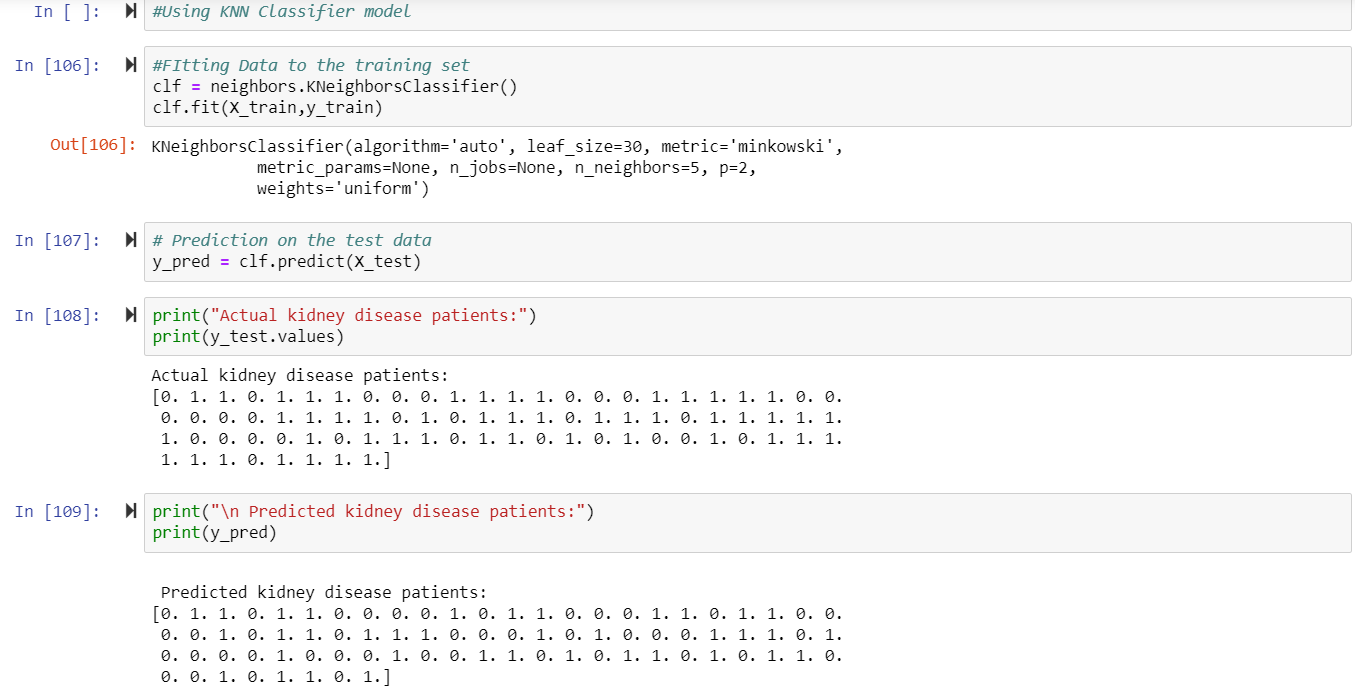
**Y-Train:**

****

**Y-Test:**

****

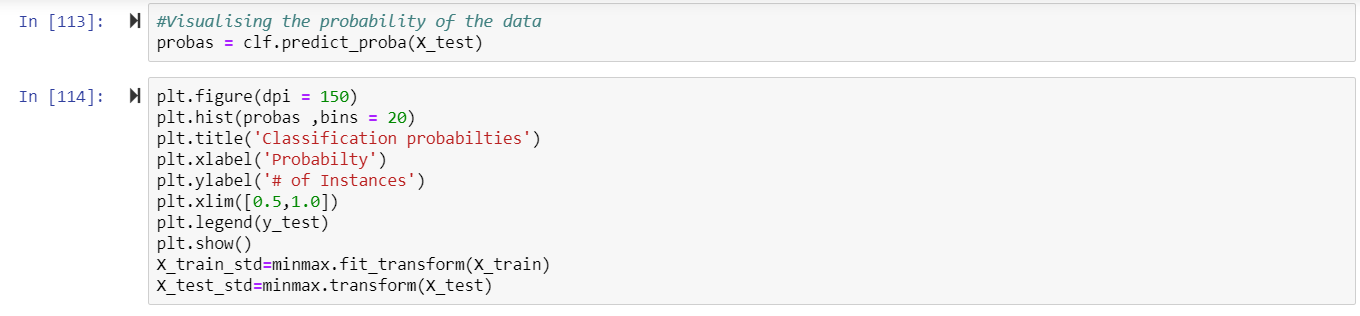
**KNN Classifiers:**

****

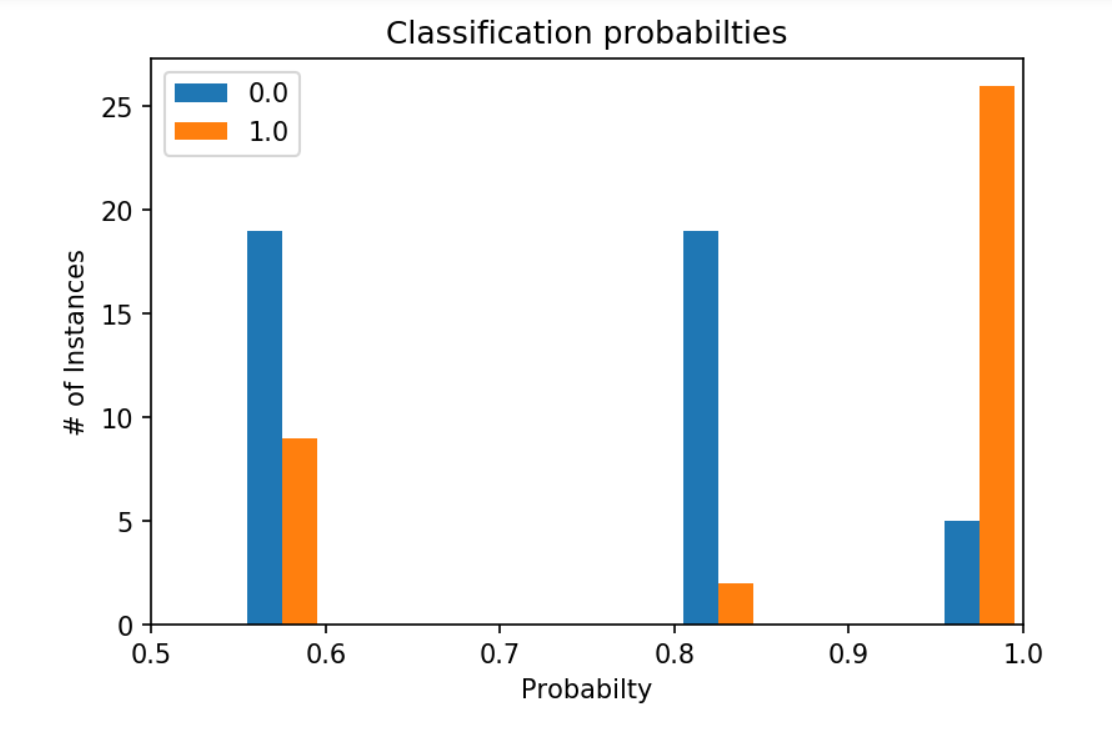
****

* **#26 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#33 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#4 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#17 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

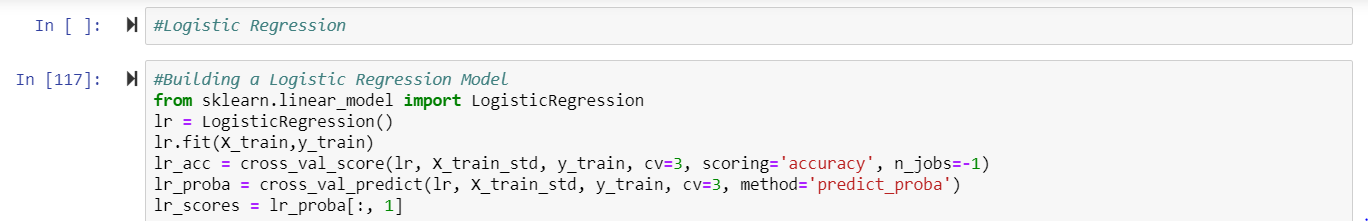
**Probability of Data:**

****

**Classification Probabilities:**

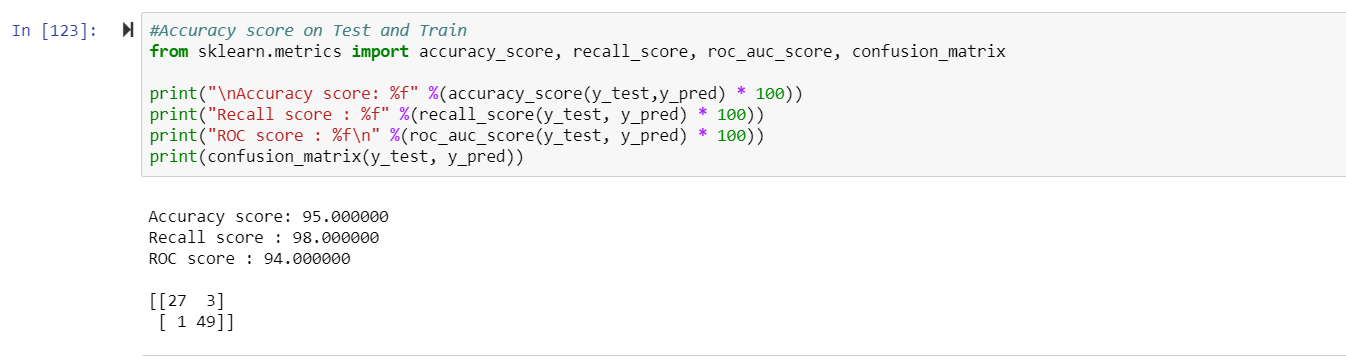
****

**Logistic Regression:**

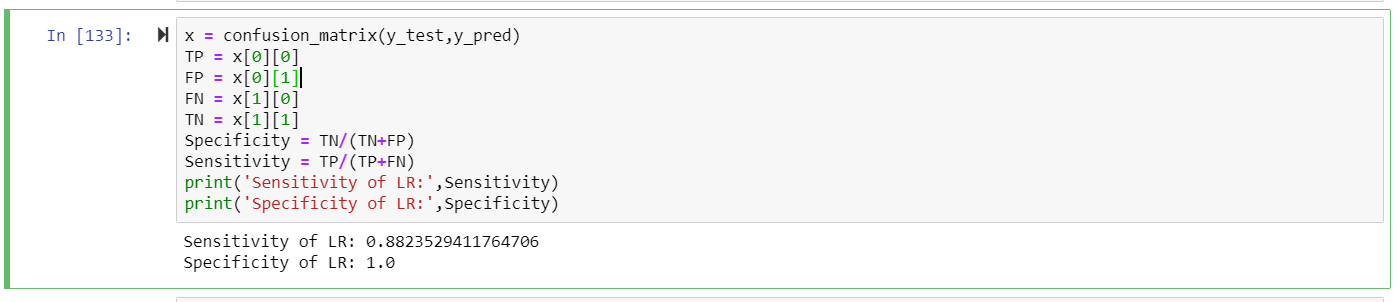
****

* **#27 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#49 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#3 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#1 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

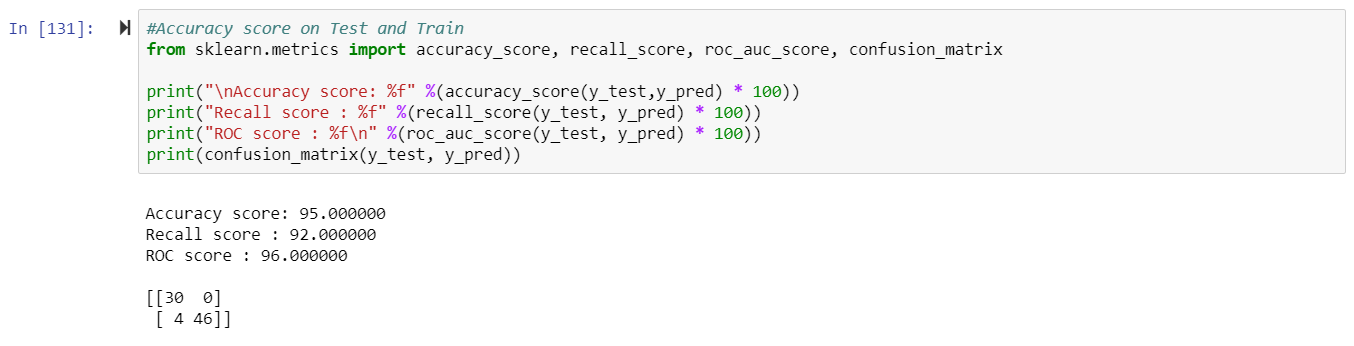
**Conclusion of Logistic Regression:**

****

**Sensitivity and Specificity of Logical Regression:**

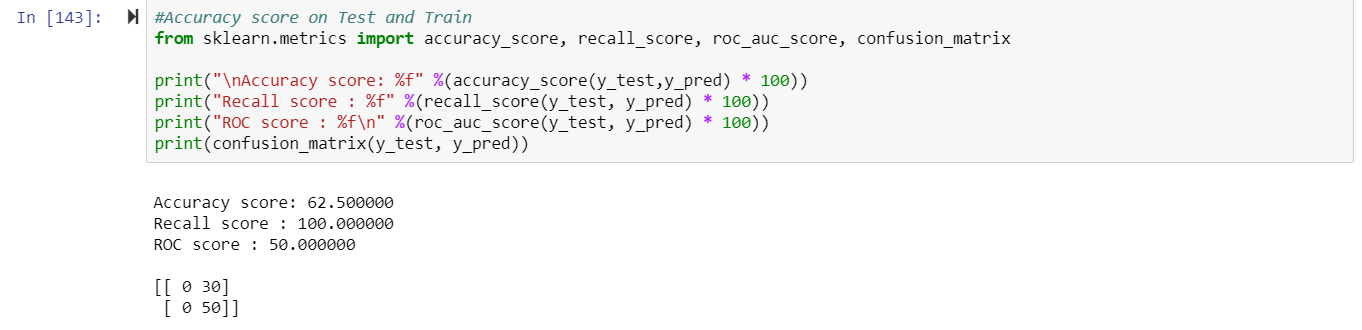
****

**Naïve Bayes:**

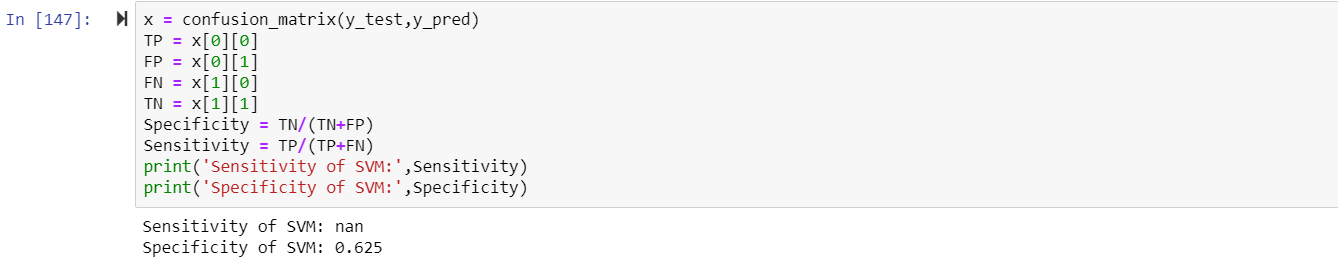
****

* **#30 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#46 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#0 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#4 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

**Support Vector Matrix (SVM):**

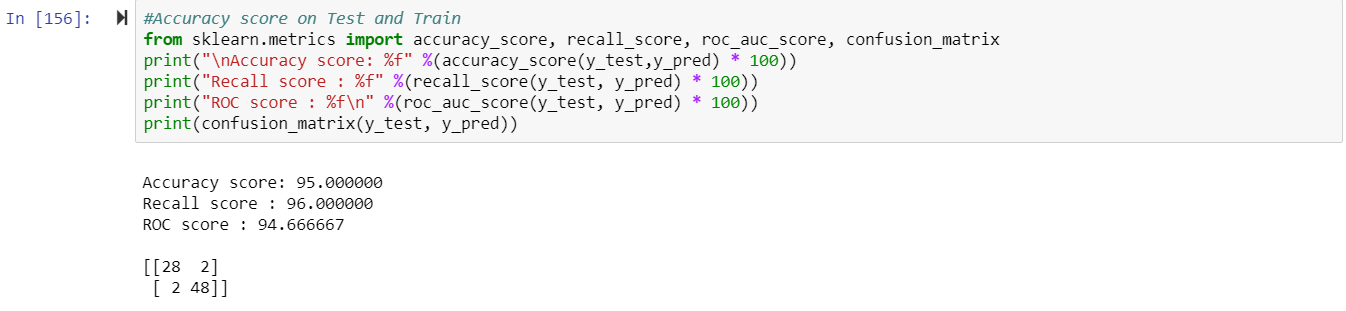
****

**Sensitivity and Specificity (SVM):**

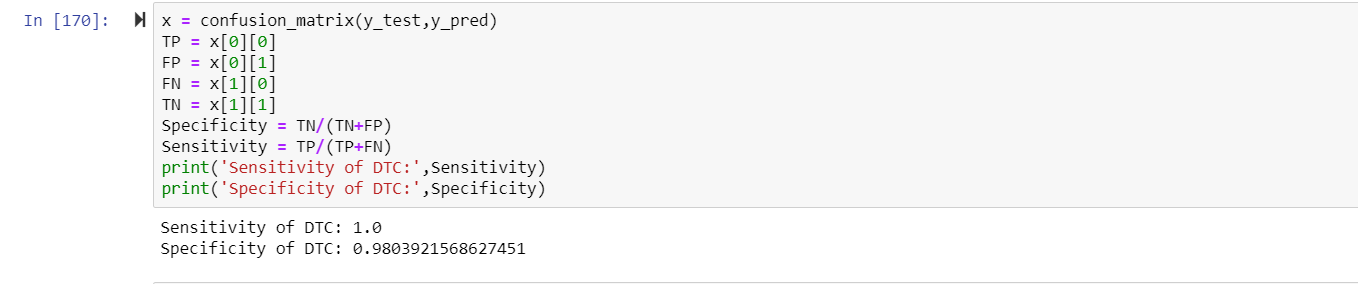
****

* **#0 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#50 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#30 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#0 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

**Decision Tree Classifier (DTC):**

****

**Sensitivity and Specificity (DTC):**

****

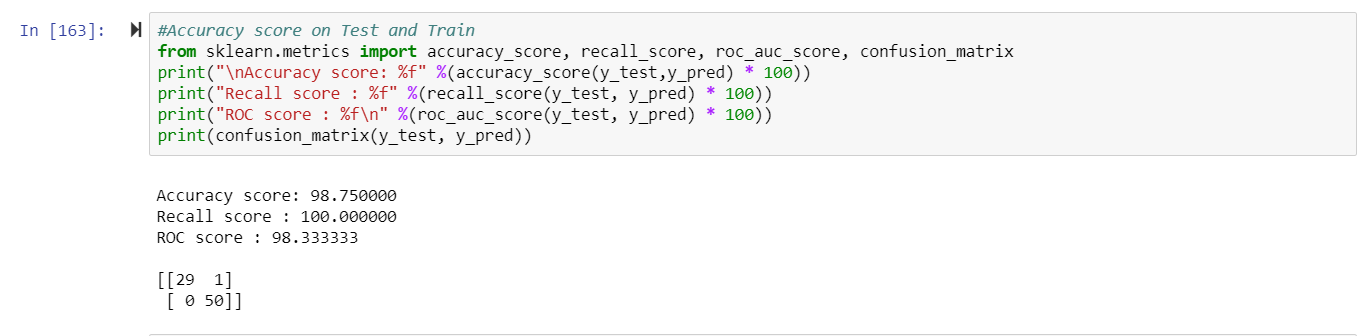
* **#28 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#48 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#2 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#2 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

**Feature Selection Technique:**

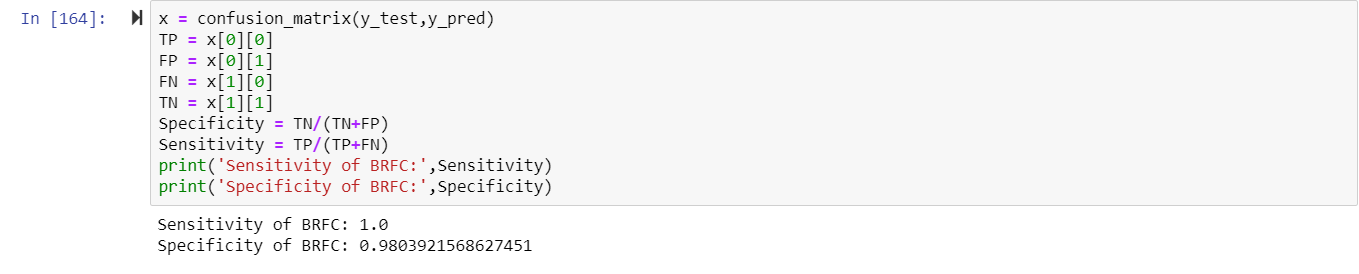
****

* **According to Graph, It can be observed that features 'sg' and 'dm' are relevant towards output Variable.**

**Random Forest Variable:**

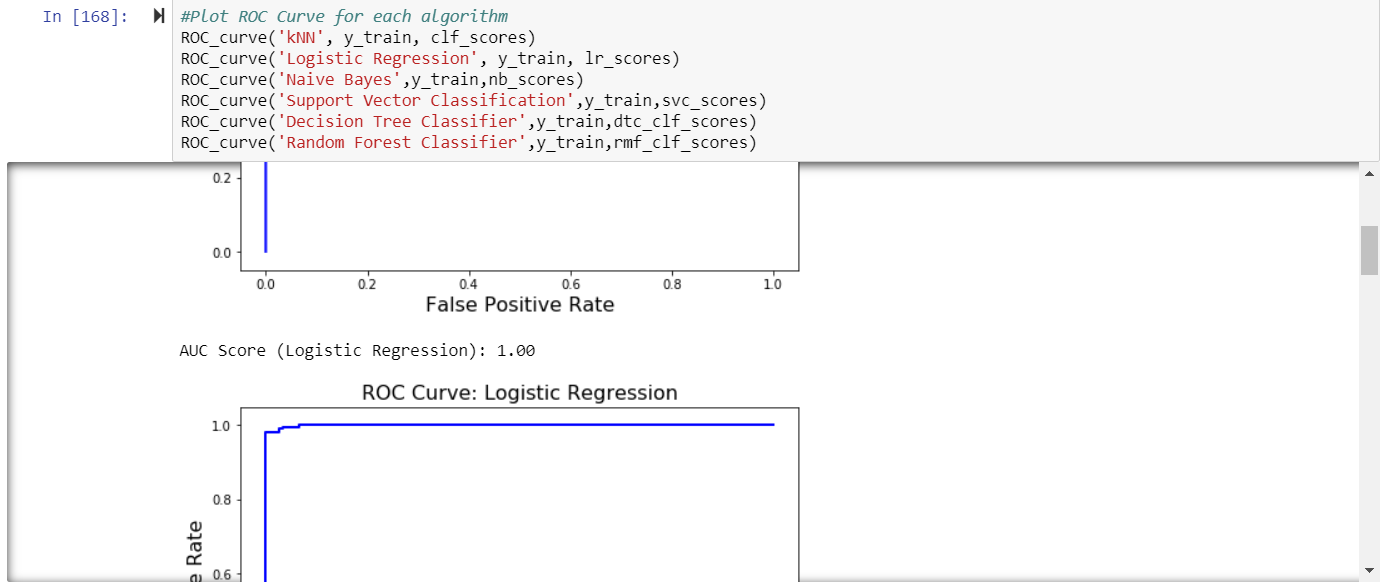
****

**Sensitivity and Specificity (RFV):**

****

* **#29 out of 80 (Doctor's Interpretation)have disease and Model Predicted Disease(True +ve)**
* **#50 out of 80 (Doctor's Interpretation)doesnt have disease and Model Predicted NO Disease(True -ve)**
* **#1 out of 80 (Doctor's Interpretation)Doesnt have disease and Model Predicted yes Disease(Flase +ve)**
* **#0 out of 80 (Doctor's Interpretation)have disease and Model Predicted NO Disease(False -ve)**

**ROC curve:**

****

**KNN**

* **Accuracy:73.75**
* **Recall Score:66**
* **ROC Score:76.33**
* **Confusion Matrix:TP=26,FP=4,FN=17,TN=33**
* **Sensitivity:0.604**
* **Specivicity:0.89**

**Logistic Regression:**

* **Accuracy:95%**
* **Recall Score:98%**
* **ROC Score:94%**
* **Confusion Matrix:TP=27,TN=49,FP=3,FN=1**
* **Sensitivity:0.88**
* **Specivicity:1**

**Naive Bayes Accuracy:95%**

* **Recall Score:92% ROC Score:96% Confusion Matrix:TP=30,TN=46,FP=0,FN=4**

**SVC**

* **Accuracy:62%**
* **Recall Score:100%**
* **ROC Score:50%**
* **Confusion Matrix:TP=0,TN=50,FP=30,FN=0**
* **Sensitivity:NaN**
* **Specivicity:0.625**

**Decision Tree Classifier**

* **Accuracy:95%**
* **Recall Score:96%**
* **ROC Score:94%**
* **Confusion Matrix:TP=28,TN=48,FP=2,FN=2**
* **Sensitivity:1**
* **Specivicity:0.98**
* **Random Forest Classifier**
* **Accuracy:98.75%**
* **Recall Score:100%**
* **ROC Score:98.33%**
* **Confusion Matrix:TP=29,TN=50,FP=1,FN=0**
* **Sensitivity:1**
* **Specivicity:0.98**
* **CONCLUSION: RFC,DTC,LR are the best suited models**