



**PROJECT TITLE: KIDNEY DISEASE ANALYSIS**

Submitted By:

A.Keerthi Reddy

Ch. Lakshmipriyanka

A.L.S.D.Ruchitha

M.Rohitha Reddy

D.Sarath

***CERTIFICATE***

This is to certify that this project based Class report entitled **“KIDNEY DISEASE ANALYSIS”** is a bonafide work done by A. Keerthi Reddy, Ch.Lakshmipriyanka, A.L.S.D.Ruchitha, M.Rohitha Reddy, D.Sarath in partial fulfillment of the requirement for the award of **internship conducted by smart bridge** during the summer 2019.

***DECLARATION***

We hereby declare that this project based lab report entitled ”**KIDNEY DISEASE ANALYSIS”** has been prepared by us in partial fulfillment of the requirement for the award of **internship conducted by smart bridge** during the summer 2019.

We also declare that this project based class report is of our own effort and it has not been submitted to any other university for the award of any degree.

**Date: 01-06-2019**

**Place: Hyderabad**

Submitted By:

A.Keerthi Reddy

Ch. Lakshmipriyanka

A.L.S.D.Ruchitha

M.Rohitha Reddy

D.Sarath

**TABLE OF CONTENTS**

1.Abstract…………………………………..

2.Introduction………………………………

3.Importing Packages………………………

4.Exploratory Data Analysis……………….

4.1 Data Cleaning…………………….

Figure:-4.1A(checking data)

Figure:-4.1B (datatype)

Figure:- 4.1 C (Mode imputation for categorical cleaning)

Figure:- 4.1 D (Mean imputation for numerical data)

4.2 Data Description

Figure:-4.2 A(Data Description)

4.3 Data Correlation

Figure :- 4.3 A(Data Correlation)

4.4 Heat Map

Figure:- 4.4 A

5.Plots and Models………………………….

Figure: 5.1A (Albumin)

Figure: 5.1B (Hyper Tension)

Figure: 5.1C (Diabetes Mellitus)

6.Algorithms………………………………..

6.1A figure(KNN)

6.1B figure(LR)

6.1C figure(SVM)

6.1D figure(DTC)

6.1E figure(BRFC)

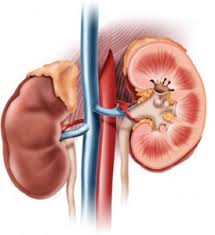
7. NodeRed……………………………………

8. Result…………………………………….

9.Conclusion……………………………….

**ABSTRACT**

Our dataset concentrates on chronic kidney disease. Both numerical data and categorical data are given in the dataset. 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. Firstly, we cleaned the missing values using mode imputation, mean imputation, then we did exploratory data analysis to analyze the data in better way. Then we drawn distplots, statsmodels to choose the feature which is affecting target more. Then we applied many algorithms on the given dataset and concluded that is LR better model by considering many evaluation metrics.



**INTRODUCTION**

Kidney disease can affect your body’s ability to clean our blood, filter extra water out of our blood, and help control your blood pressure. It can also affect red blood cell production and vitamin D metabolism needed for bone health.

When your kidneys are damaged, waste products and fluid can build up in your body. That can cause swelling in your ankles, nausea, weakness, poor sleep, and shortness of breath. Without treatment, the damage can get worse and your kidneys may eventually stop working. That’s serious, and it can be life-threatening.

**What Your Kidneys Do**

Healthy kidneys:

* Keep a balance of water and minerals (such as sodium, potassium, and phosphorus) in your blood
* Remove waste from your blood after digestion, muscle activity, and exposure to chemicals or medications.
* Make an active form of vitamin-D, needed for bone health and other things

**Kidney Problems**

* Kidney stones
* Chronic kidney disease

**Importing packages**

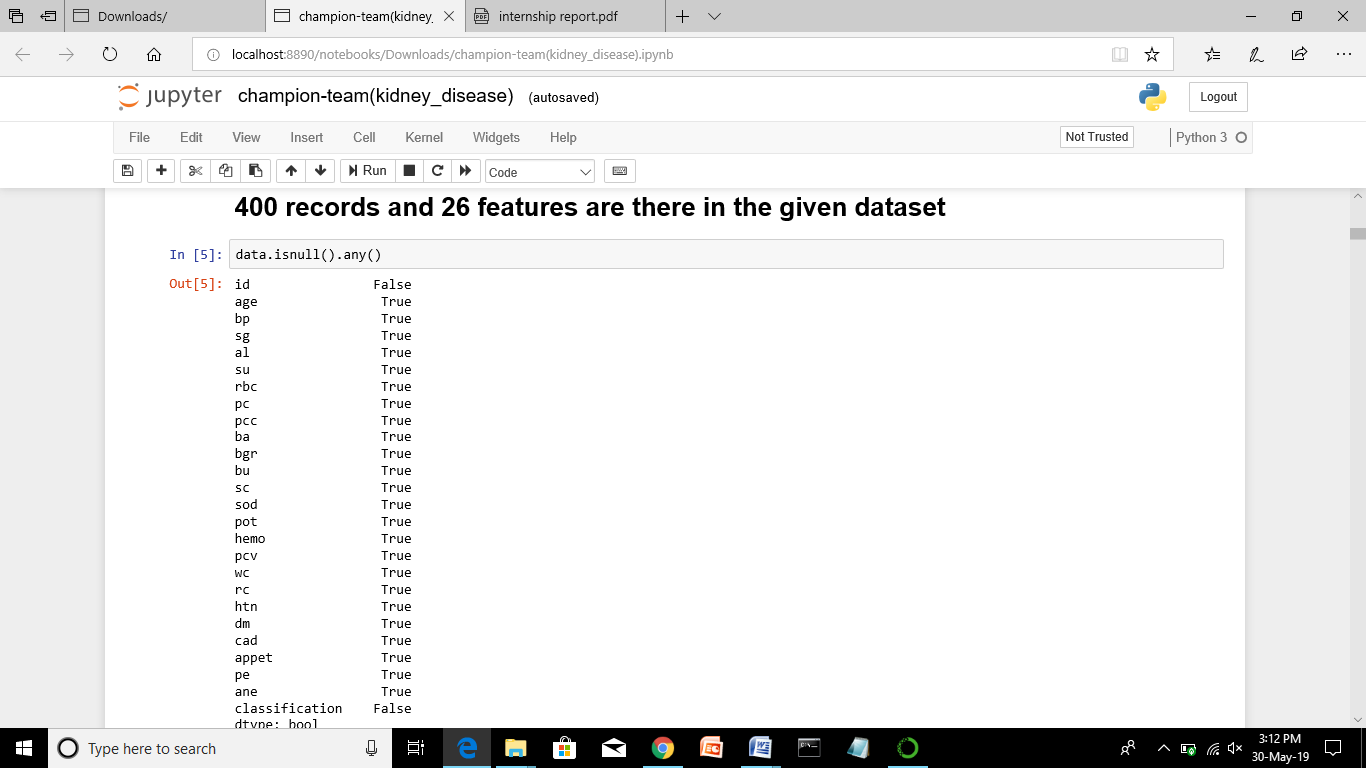
* Numpy
* Pandas
* Imputer
* Seaborn
* Matplotlib
* Sklearn(Logistic Regression, SVC, Decision Tree Classifier,

Random Forest Classifier)

**Exploratory Data Analysis**

**Data Cleaning**:-

**Figure 4.1A:-**



We found missing values in both numerical and categorical data by using the data.isnull().any() function

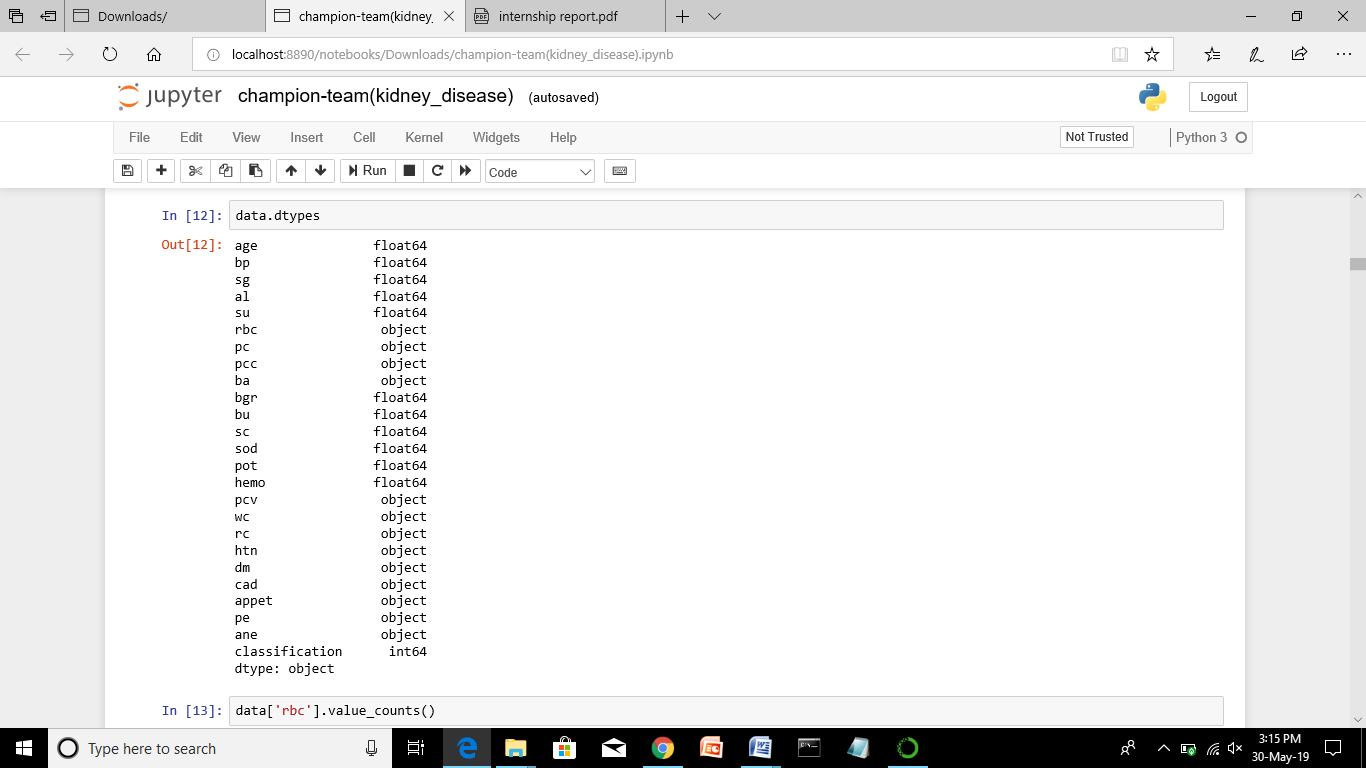
Numerical data can be cleaned using mean imputation

Categorical data can be cleaned using mode imputation

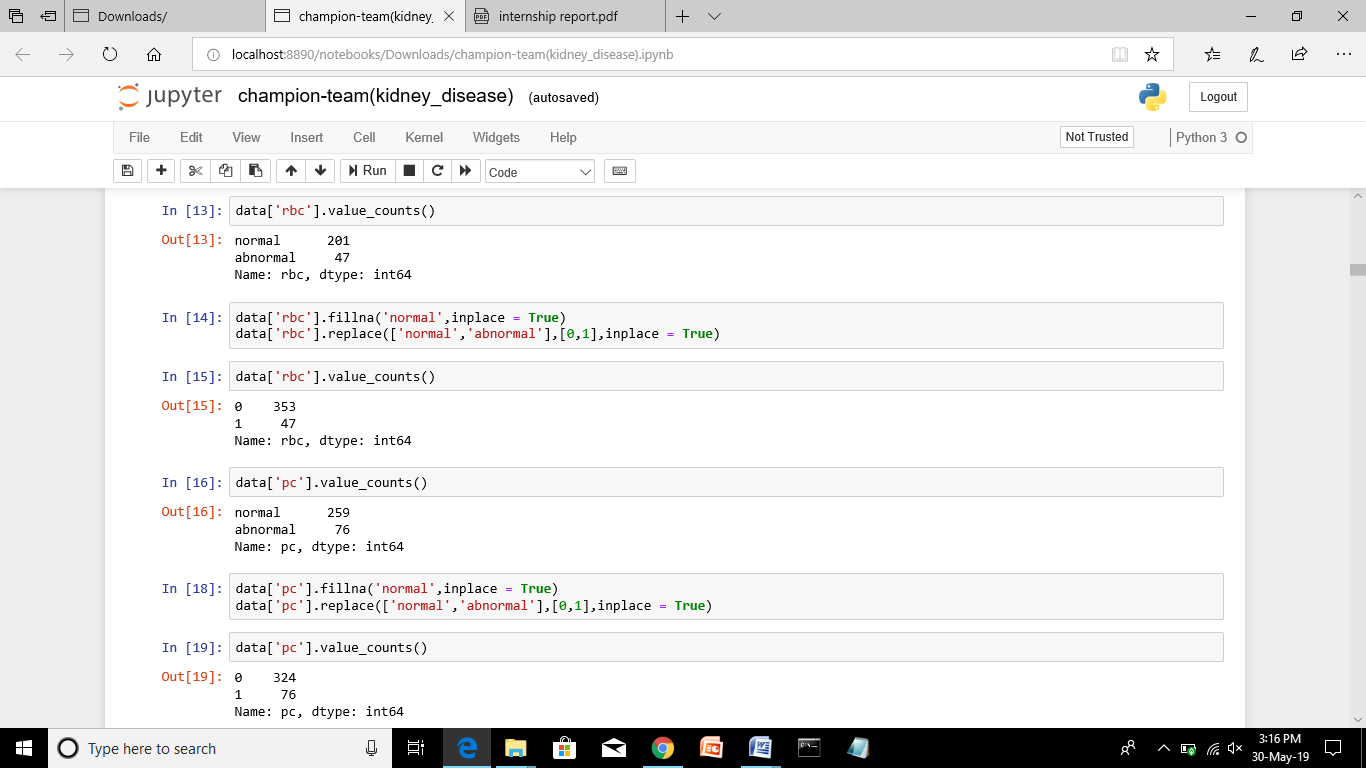
We should check the datatypes() inorder to implement imputation

**Datatype:-**

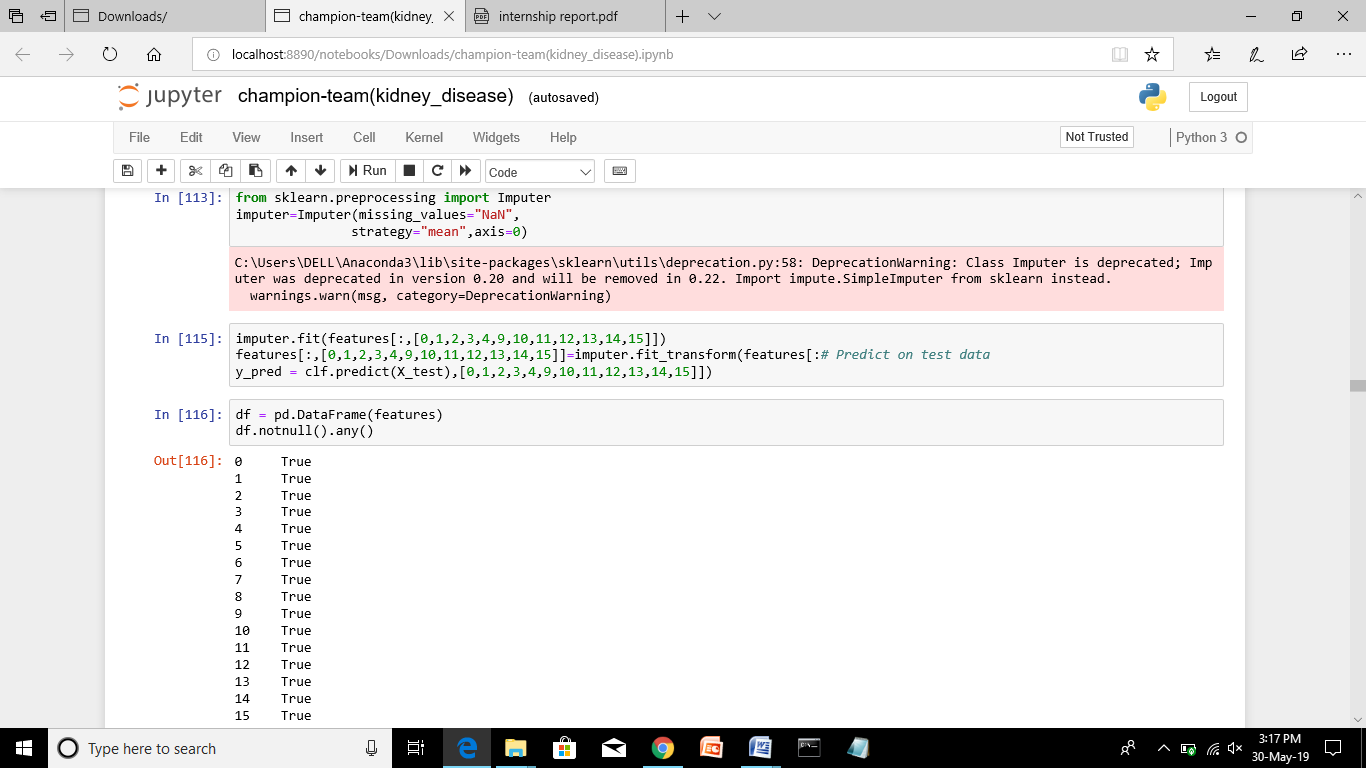
**Figure 4.1B:-**



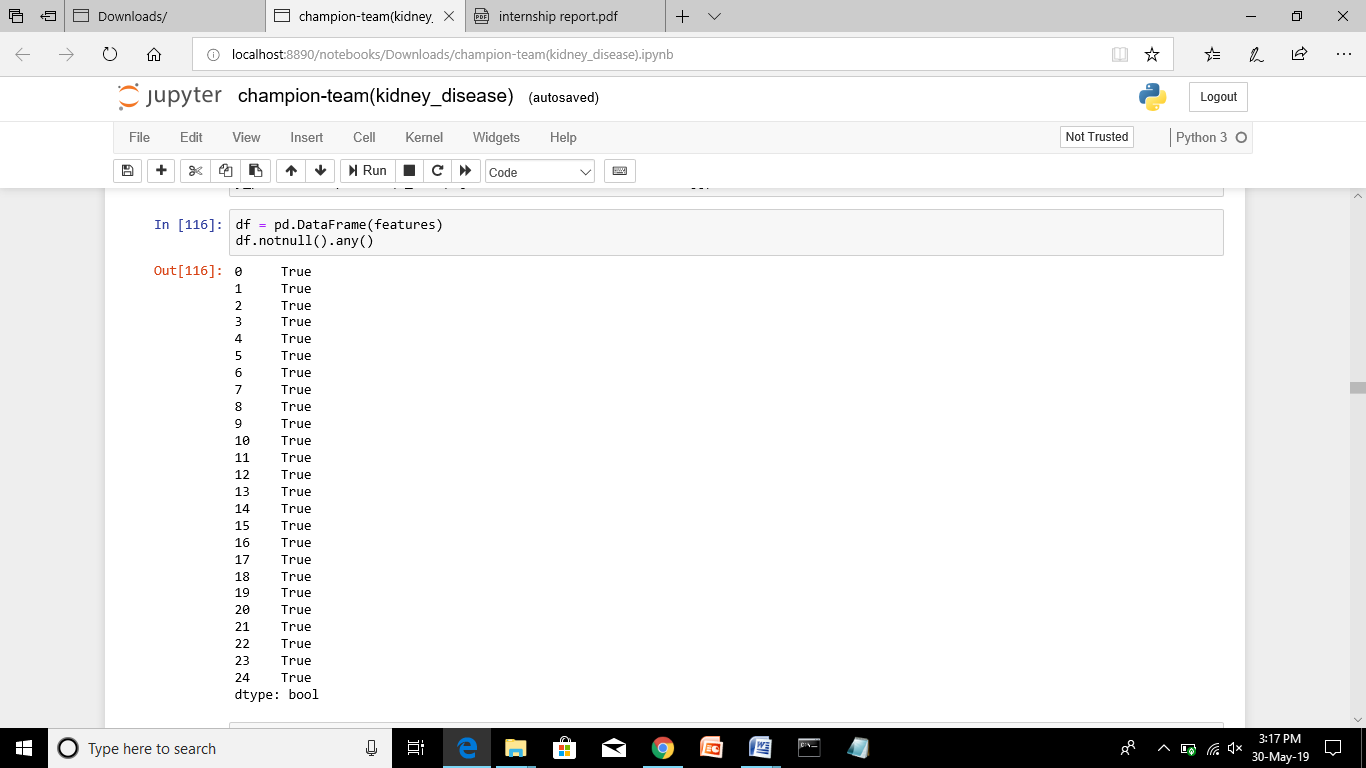
**Figure:- 4.1 C Mode imputation for categorical cleaning:-**



**Figure:- 4.1D Mean imputation for numerical data:**



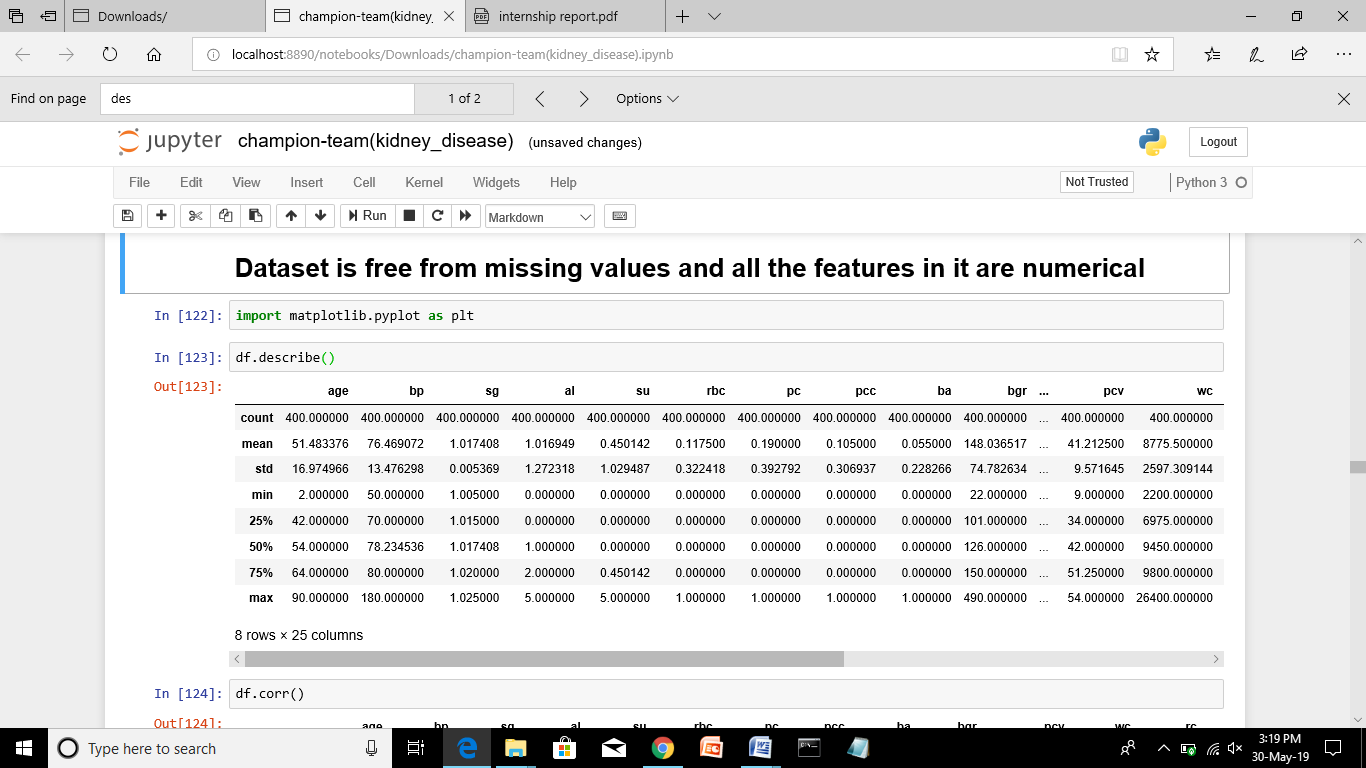
Now, we can see that data is clean from missing values.



It involves the clear understanding about data, by finding Missing data, data description, data correlation , heat maps.

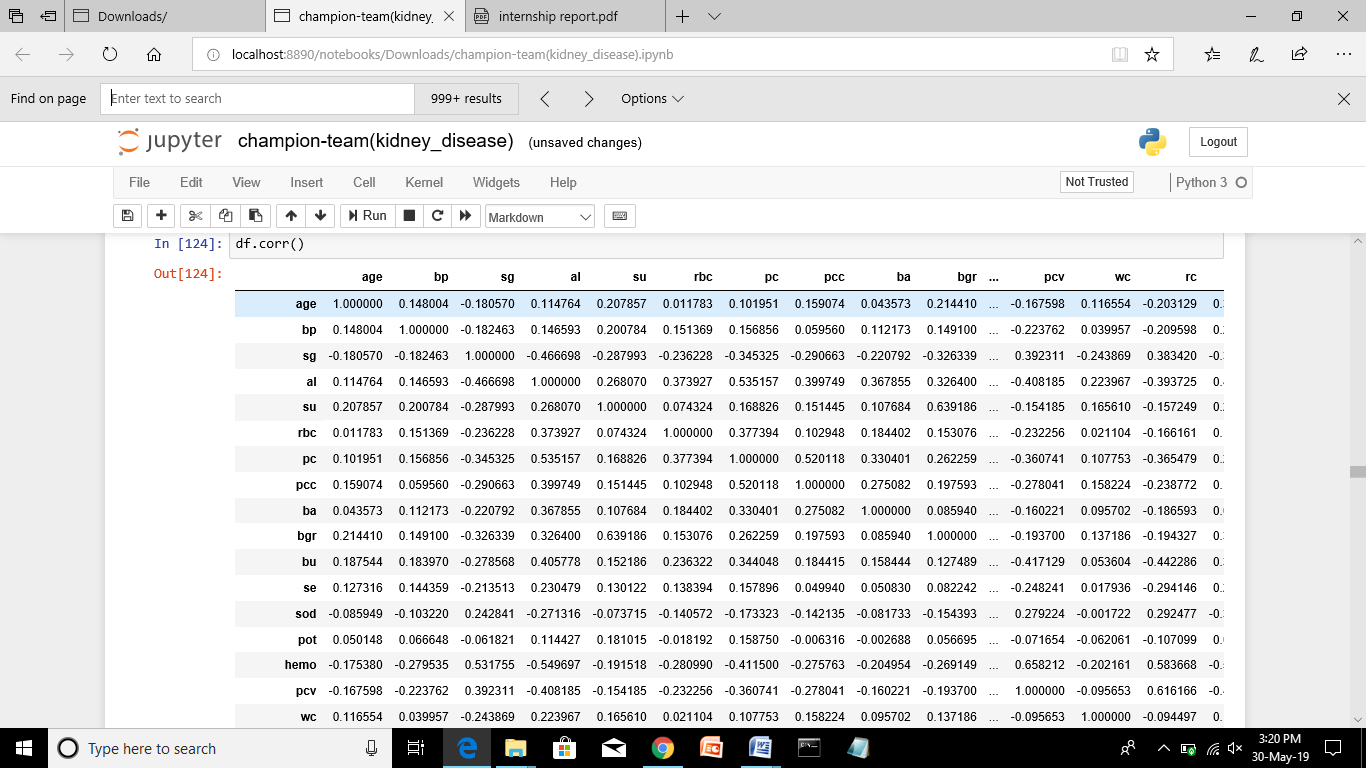
**Data Description:**

**Figure:-4.2 A:-**



**Data Correlation:-**

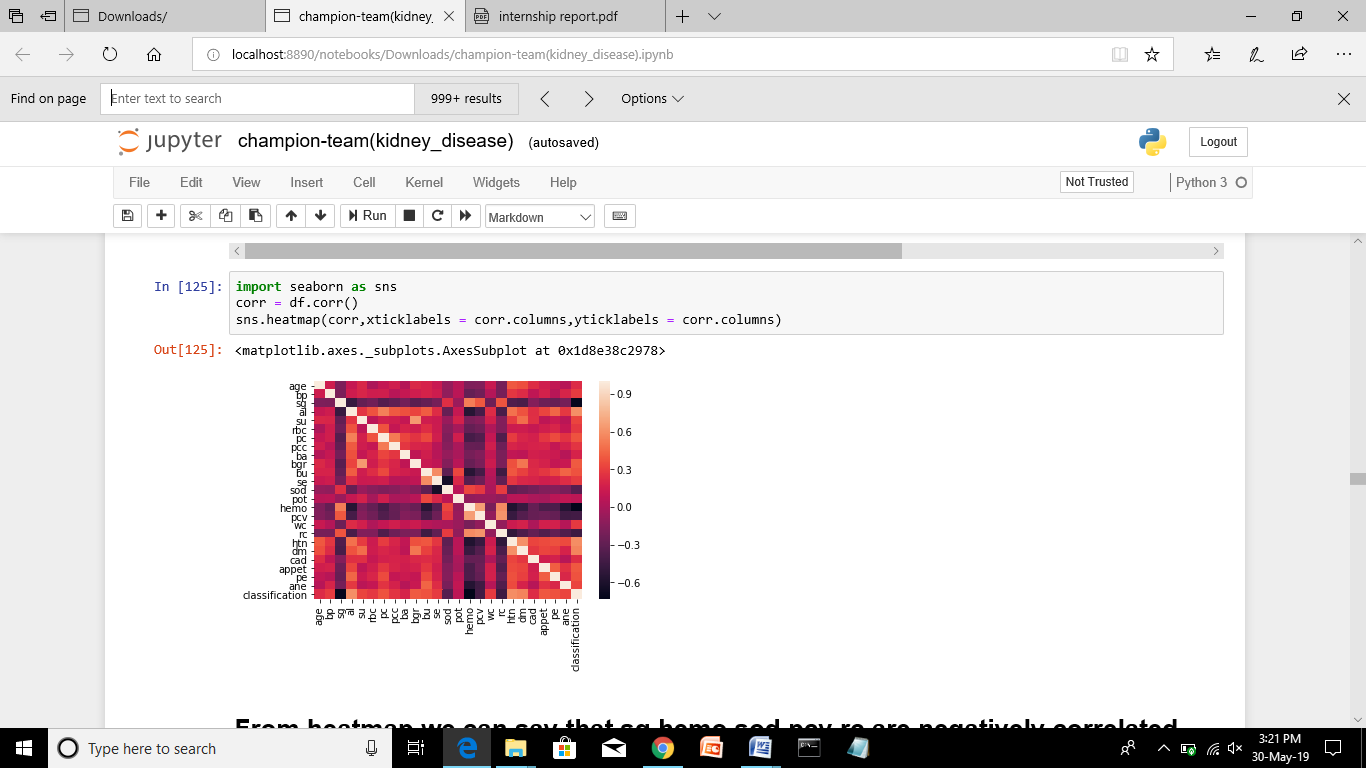
**Figure:- 4.3 A:-**



From correlation we observed that sg, sod, hemo,pcv are negatively correlated with classification

**Heat Map**

**Figure:- 4.4 A**



From Heatmap also we observed that al, dm are highly positively correlated with classification.

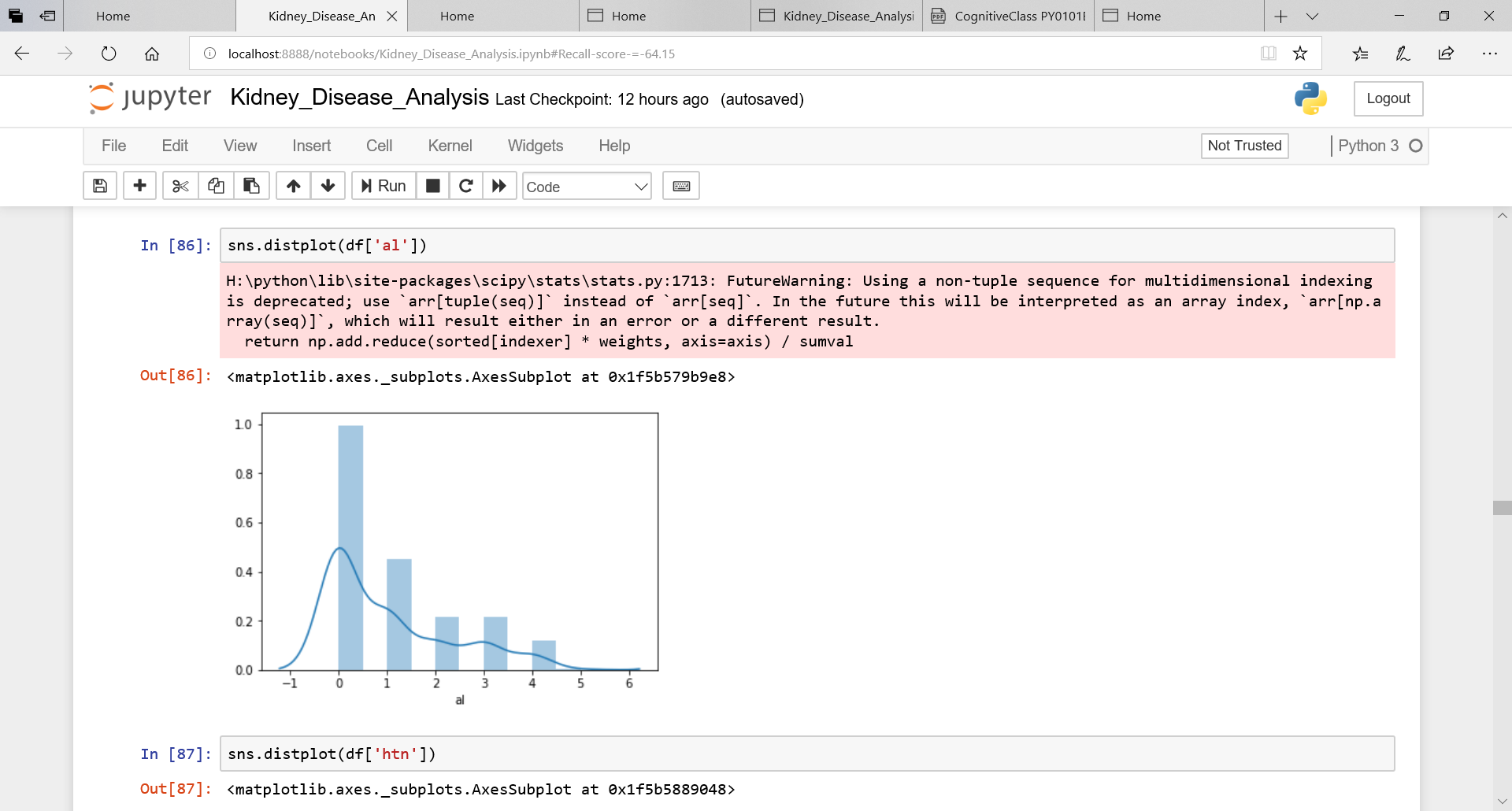
**Plots and Models**

**Distplots:**

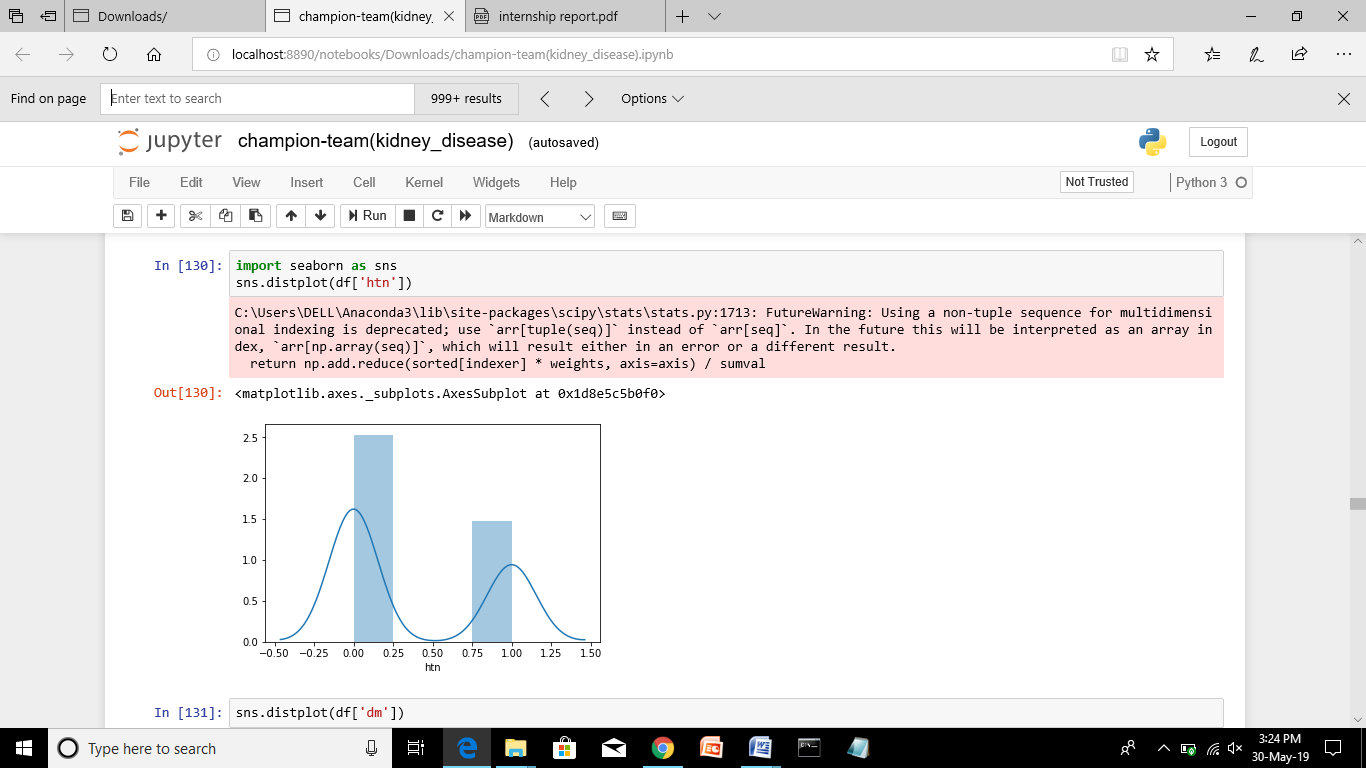
From correlation table we have considered the features al, dm

**Albumin:**

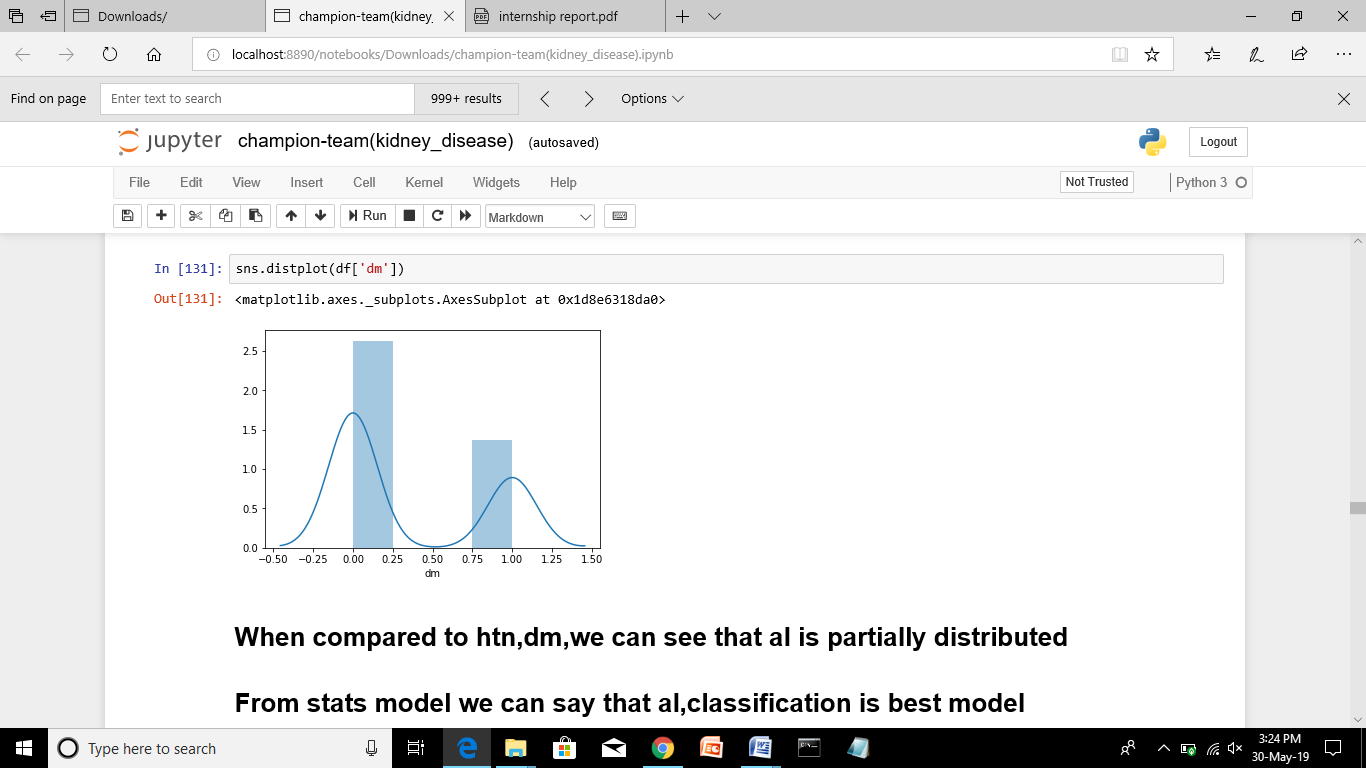
**Figure: 5.1A**



**Hyper Tension:-**



**Diabetes Mellitus:-**



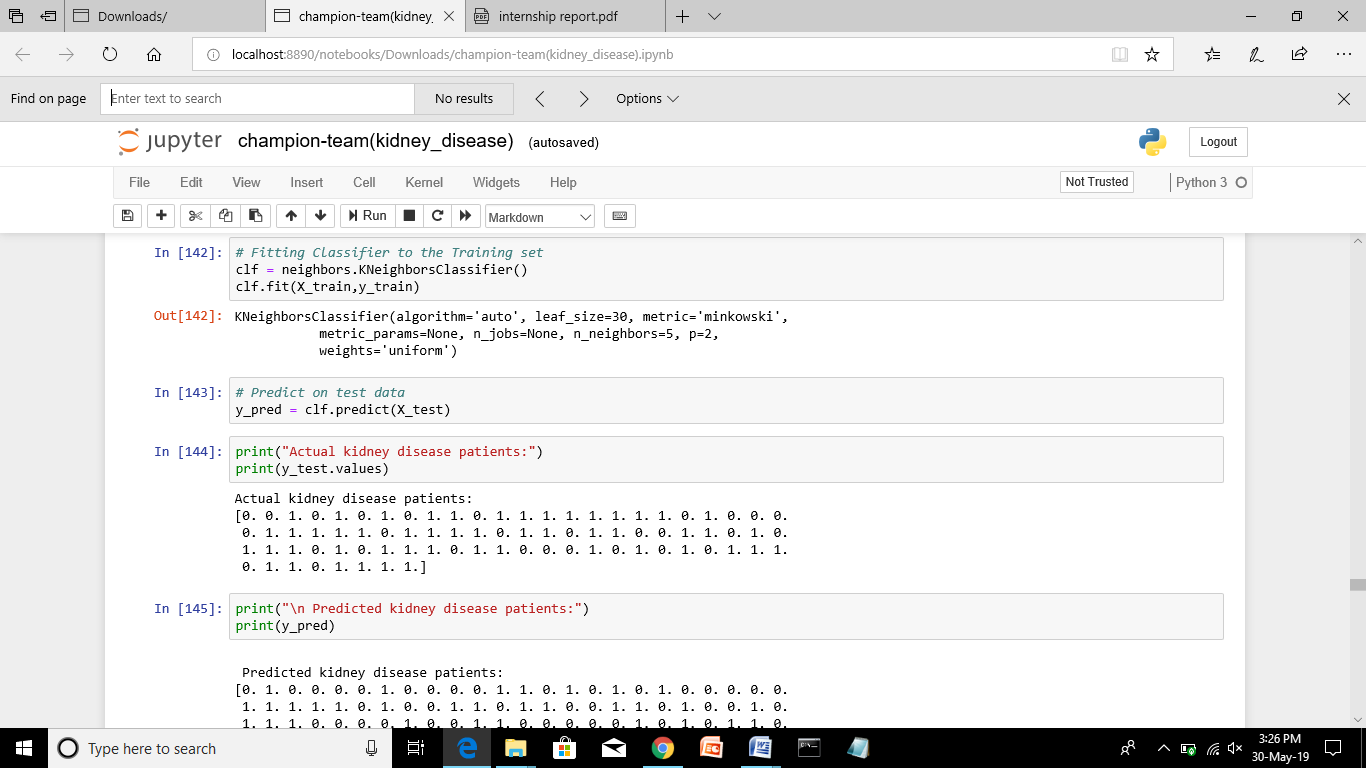
From these distplots we observed that al is partially normally distributed.

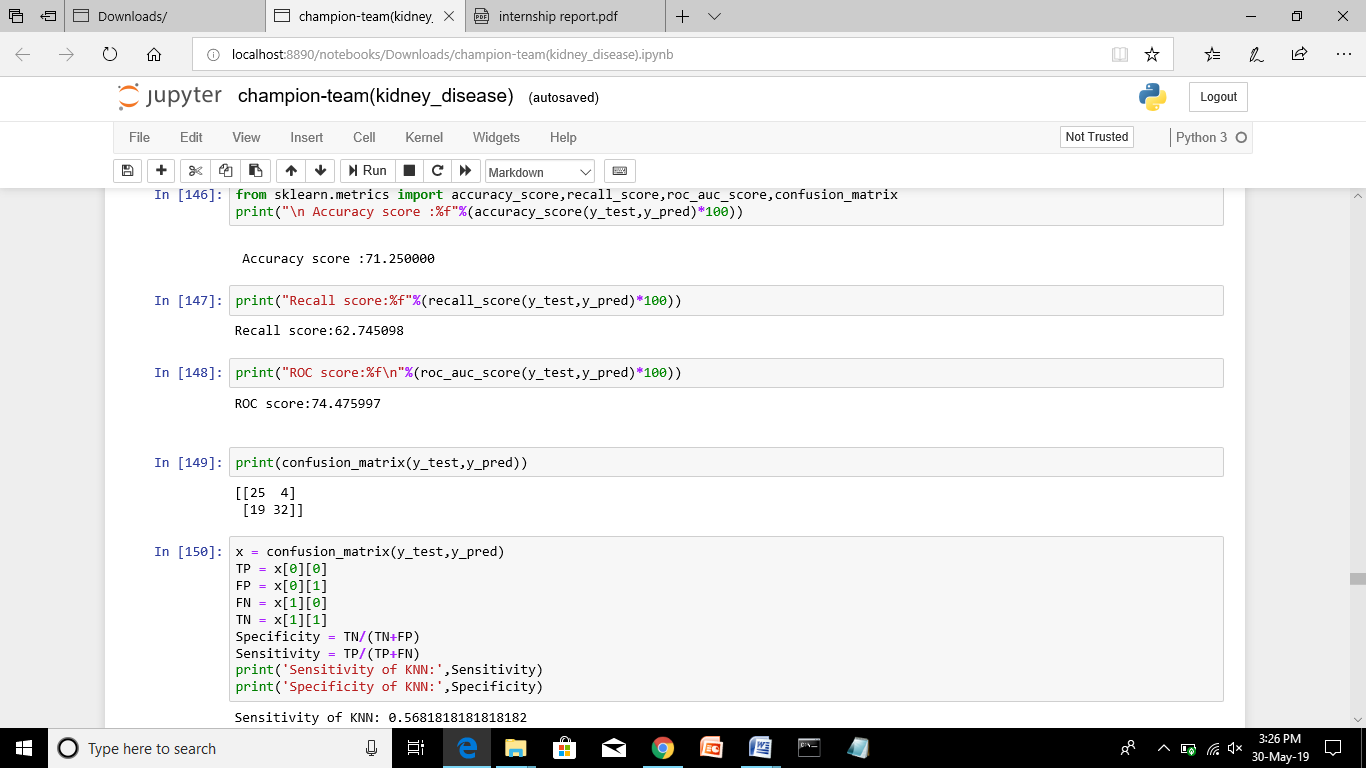
While htn, dm are non normally distributed.

**ALGORITHMS:-**

**KNN:-**

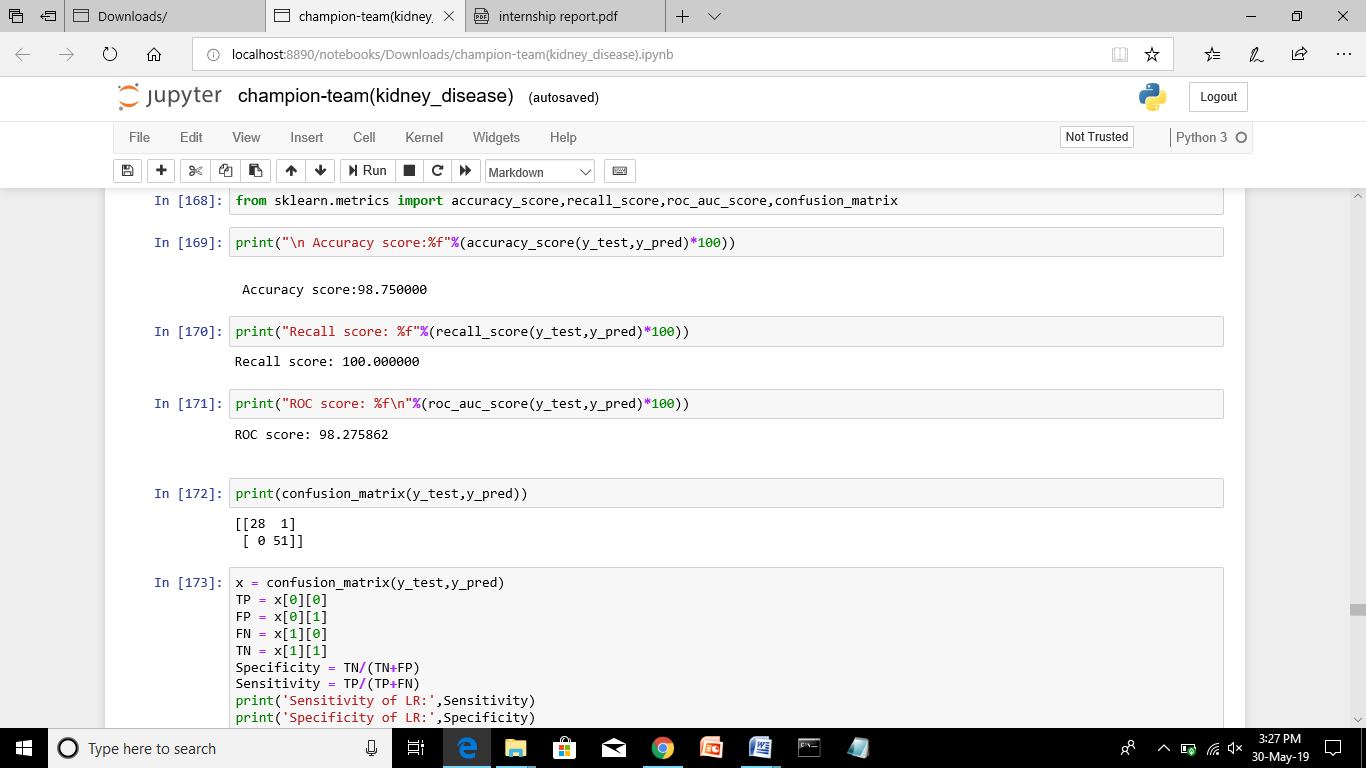
**Figure:6.1A:-**



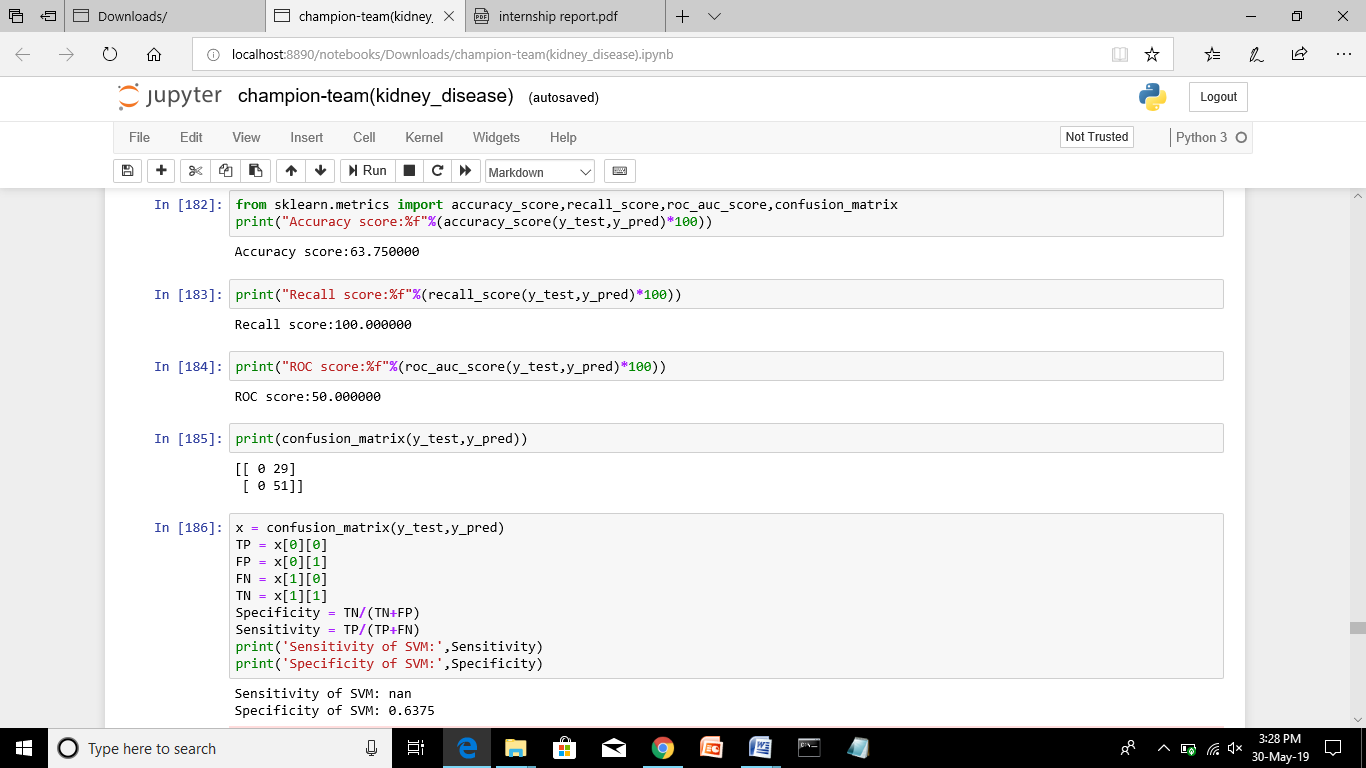


**LR:**

**Figure:6.1 B**

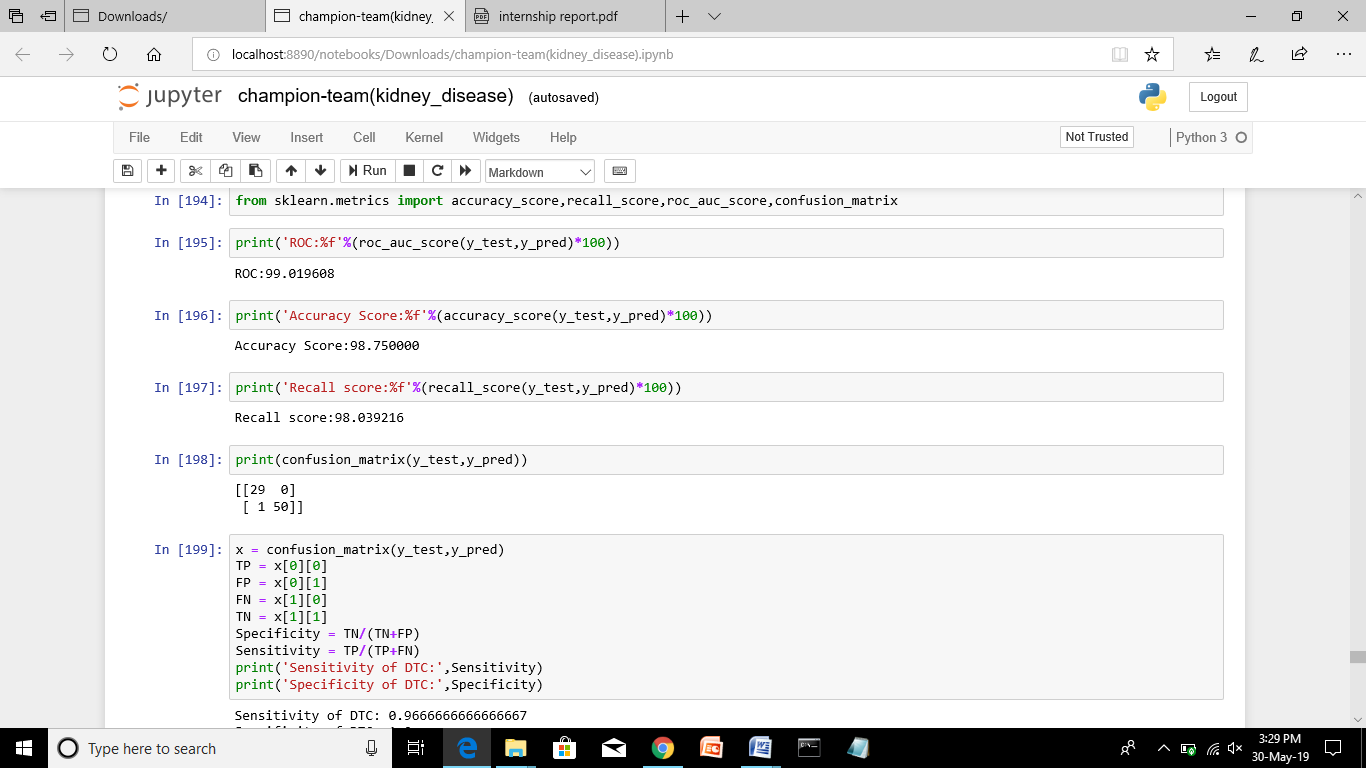
**SVM:**

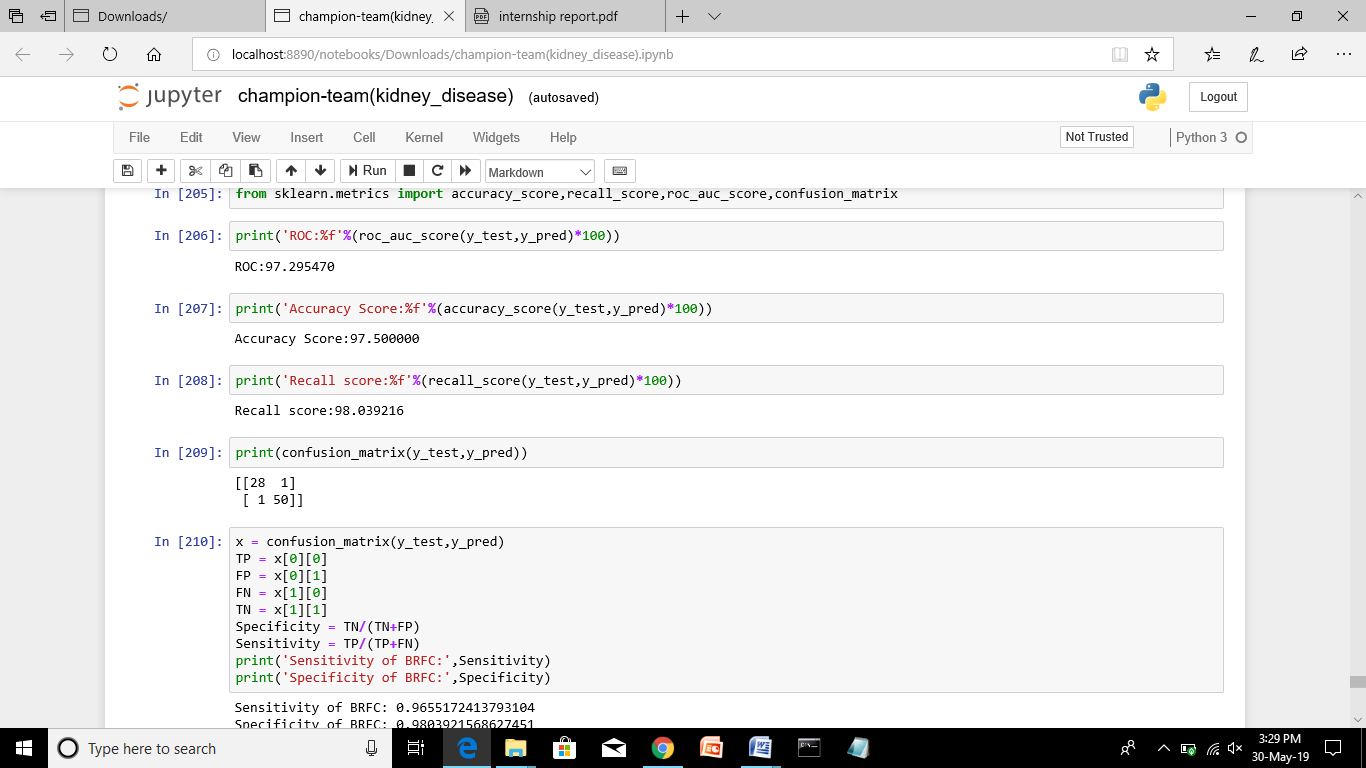
**Figure:6.1 C:-**



**DTC:-**

**Figure:6.1 D:-**



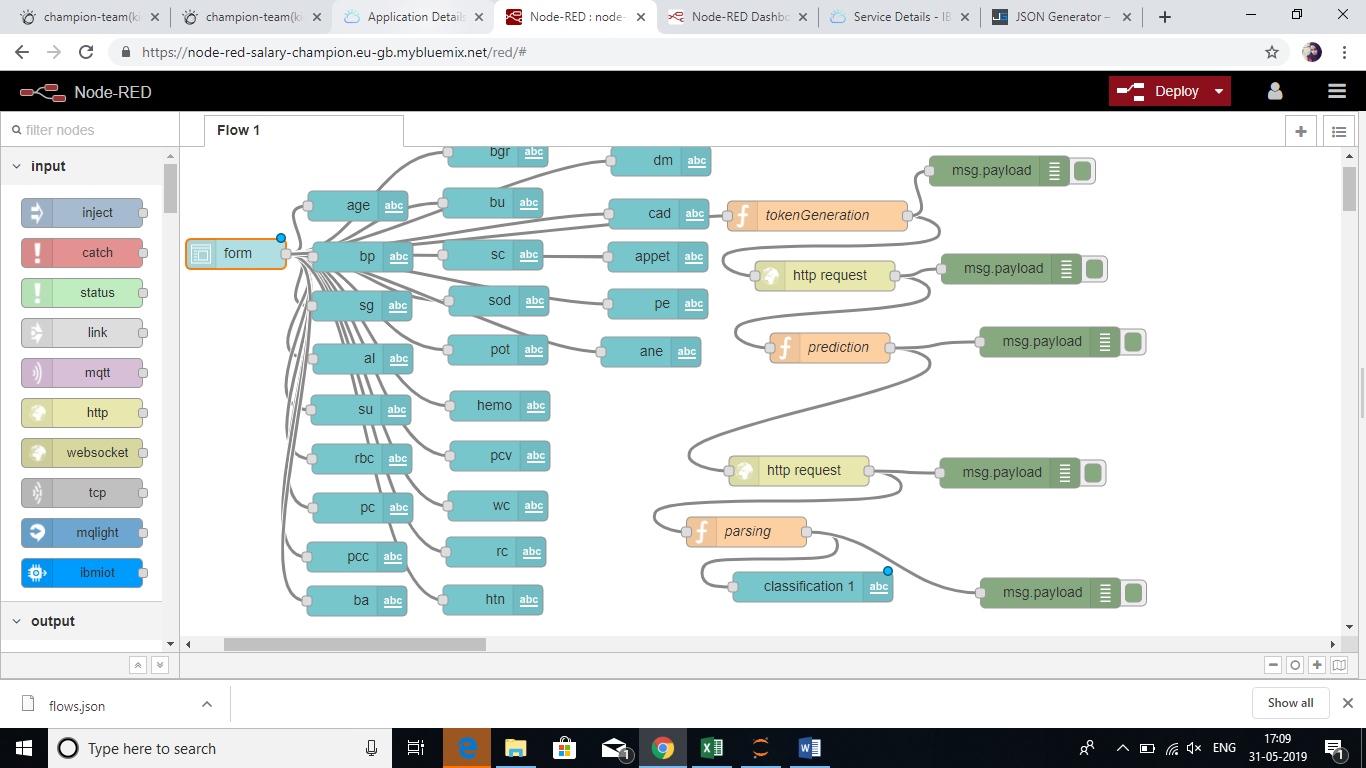
**BRFC:Figure 6.1** From these algorithms we observed that RTC is better than remaining algorithms by considering many evaluation metrics.

NODERED:

Node-RED provides a browser-based flow editor that makes it easy to wire together flows using the wide range of nodes in the palette. Flows can be then deployed to the runtime in a single-click.

JavaScript functions can be created within the editor using a rich text editor.

A built-in library allows you to save useful functions, templates or flows for re-use.



**RESULT**

These are the following observations from the above models:

**KNN:**

**Accuracy of the model** = 71.25%

**Recall score** = 62.74%

**ROC score** = 74.47%

**Confusion matrix**(TP = 25,FP = 4,FN = 19,TN = 32)

**Sensitivity of KNN** : 0.5681818181818182

**Specificity of KNN**: 0.8888888888888888

**Logistic Regression**:-

**Accuracy of the model** = 98.75%

**Recall score** = 100%

**ROC score** = 98.27%

**Confusion matrix**(TP = 28,FP = 1,FN = 0,TN = 51)

**Sensitivity of LR:** 1.0

**Specificity of LR:** 0.98

**SVC**

**Accuracy of the model =** 63.75%

**Recall score** = 100%

**ROC score** = 50%

**Confusion matrix**(TP = 0,FP = 29,FN = 0,TN = 51)

**Sensitivity**: nan

**Specificity of SVM:** 0.6375

**DTC:-**

**Accuracy of the model** = 98.75%

**Recall score** = 98.03%

**ROC score** = 99.01%

**Confusion matrix**(TP = 29,FP = 0,FN = 1,TN = 50)

**Sensitivity of DTC:** 0.97

**Specificity of DTC:** 1.0

**BCRF:-**

**Accuracy of the model** = 97.50

**Recall score** = 98.03

**ROC score** = 97.29

**Confusion matrix**(TP = 28,FP = 1,FN = 1,TN =50)

**Sensitivity of BRFC**:0.96

**Specificity of BRFC**: 0.98

**CONCLUSION**

By performing all the models, I observed that DTC, LR, BRFC are better than other models.

So I conclude that LR and DTC is best model