Python ML Internship

Project Report

Analysis of Dry Bean Dataset

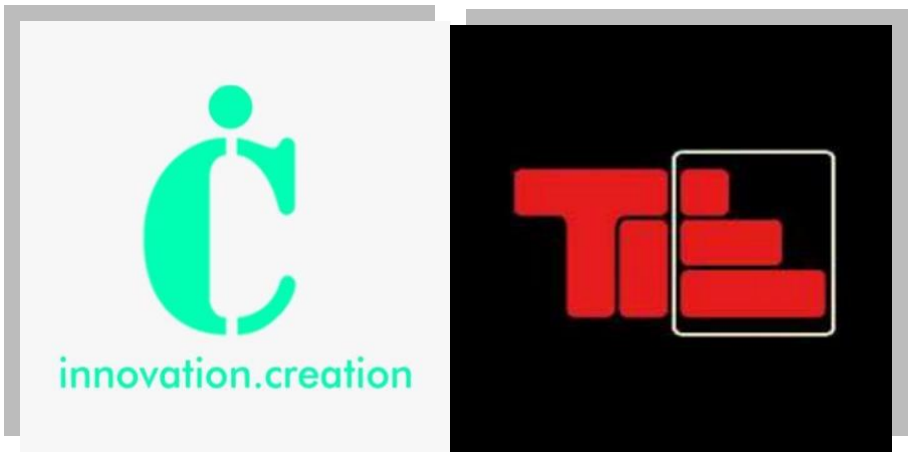
Submitted by:

Thanushree S Babu [4VM17EE72 tsbabu480@gmail.com]
Prajwal M S [1VE18EC070 gowdaprajwal009@gmail.com]
Shreya S [1VE18CS150 shreyasureshshet@gmail.com]

Online Internship Organizes By:

IC Solutions

In association with **Takeiteasy_Engineers(TIE)**



Under the guidance of

Mr.Abhishek C

Acknowledgement

Firstly I would like to express my special thanks of gratitude to **Take It Easy Engineers**(**TIE**) for arranging this internship program.Also I would really like to thank **IC Solutions** for giving the students such a golden opportunity to do **Python ML internship** at just **₹799**. Providing such a quality training at low price is really appreciable.As doing an internship is a must for all the VTU students, it was really difficult to find good internship program during the pandemic. This online internship has really helped me.

I would like to extend my gratitude to my instructor **Mr Abhishek C.** I'm really fortunate that such a good trainer was assigned to me. He has so much knowledge in this area, so all the eleven sessions of this internship program were really informative. He shared his experience in the field of ML during the sessions which was really great. He used to clear all the doubts asked by each & every student, due to which all the concepts taught by him are crystal clear.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

Hope to continue cooperation with all of you in the future.

Sincerely,
Thanushree S Babu
Prajwal M S
Shreya S

Place: Bangalore

Date: 12/04/2021

Abstract

The current project is to predict the Accuracy of the Dry Bean Dataset using ML. An Analysis of the Dry Bean Dataset requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes is examined for the reliable and accurate prediction. In this project, we were asked to experiment with a real world dataset, and to explore how machine learning algorithms can be used to find the patterns in the data. We were expected to gain experience using a common data-mining and machine learning library, and were expected to submit a report about the dataset and the algorithm used. After performing the required tasks on the dataset of Dry bean Data, here lies my final report. To build a model for predicting the Accuracy of the Dataset, three machine learning techniques (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine) have been applied. The data used for the prediction is 'Dry_Bean_Dataset_.csv'.

About the company

IC Solutions(ICS) is a digital service provider that aims to provide software, designing and marketing solutions to individuals and businesses. ICS believes that service and quality is the key to success.

They provide all kinds of technological and designing solutions from Billing Software to Web Designs or any custom demand that you may have. Experience the service like none other!

Development - They develop responsive, functional and super fast websites. They keep User Experience in mind while creating websites. A website should load quickly and should be accessible even on a small view-port and slow internet connection.

Mobile Application - They offer a wide range of professional Android, iOS & Hybrid app development services for global clients, from a start-up to a large enterprise.

Design - They offer professional Graphic design, Brochure design & Logo design. They are experts in crafting visual content to convey the right message to the customers.

Consultancy - They provide expert advice on the client's design and development requirement.

Videos - They create a polished professional video that impresses the audience..

INDEX

Introduction

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

An ML algorithm is supposed to perform task and gain experience with the passage of time. The measure which tells whether ML algorithm is performing as per expectation or not is its performance (P). P is basically a quantitative metric that tells how a model is performing the task, T, using its experience, E. There are many metrics that help to understand the ML performance, such as accuracy score, r2_score, confusion matrix, precision, recall, sensitivity etc.

From this internship program I learned the basics of Artificial Intelligence (AI), Machine Learning using Python, Data Analysis & Data Visualization using different libraries, Training & Testing the models using ML algorithms like Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, Random Forest & K Nearest Neighbors.

Using the knowledge gained by this internship I have completed a ML project which involved Exploratory Data Analysis, Training & testing the model using three different algorithms.

Problem Statement

To predict the Accuracy of Dry Bean based on the given data set. Using these data set we have to train a Machine Learning model to find Accuracy and R2_score of the Dataset.

Objective

1. Data Analysis is done to analyse the given data set & summarize their main characteristics.
2. To predict the Accuracy of the dataset, we need to apply Regression algorithm. After training & testing the model the r2_score has to be evaluated for all the three algorithms.

System Requirements

Hardware Specifications (Minimum Requirement):-

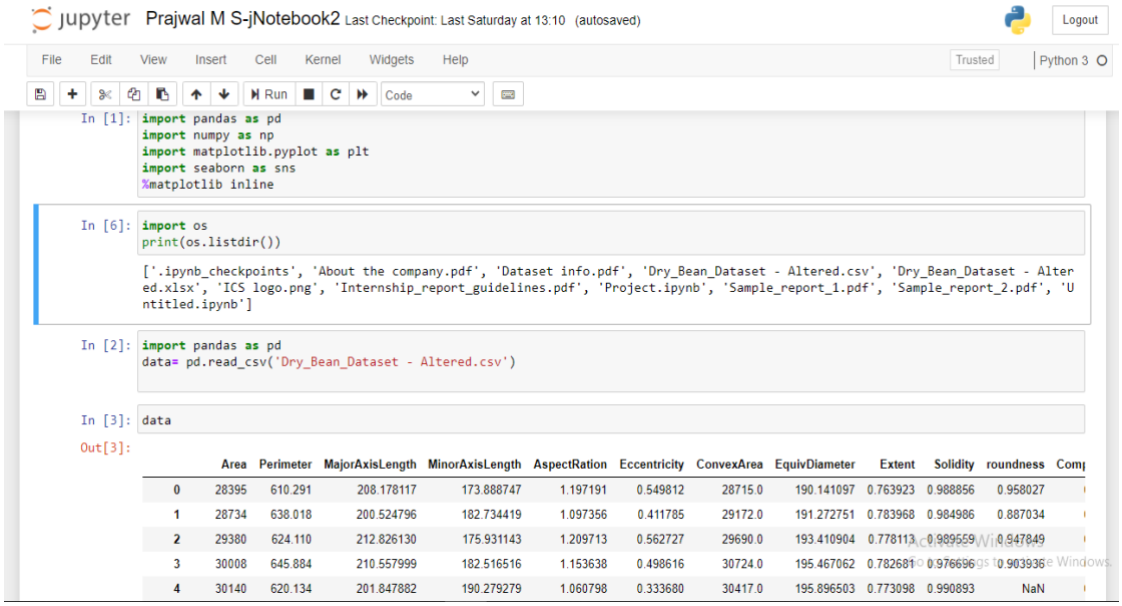
- RAM: 4 GB
- CPU: Processor above Intel Corei3 8th Gen
- OS: Windows 10/Mac OS

Software Requirements:-

- Jupyter Notebook
- Pandas
- NumPy
- Scikit-learn

Exploratory Data Analysis

1. Reading the data set:



jupyter Prajwal M S-jNotebook2 Last Checkpoint: Last Saturday at 13:10 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [6]:
```
import os
print(os.listdir())
```

```
['.ipynb_checkpoints', 'About the company.pdf', 'Dataset info.pdf', 'Dry_Bean_Dataset - Altered.csv', 'Dry_Bean_Dataset - Altered.xlsx', 'ICS logo.png', 'Internship_report_guidelines.pdf', 'Project.ipynb', 'Sample_report_1.pdf', 'Sample_report_2.pdf', 'Untitled.ipynb']
```

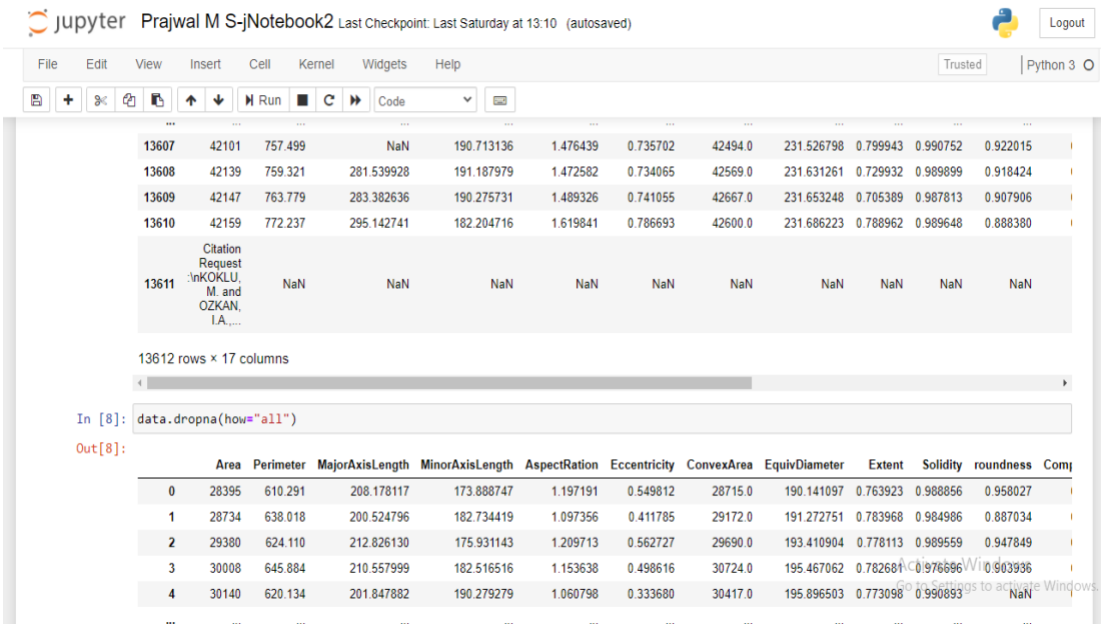
In [2]:
```
import pandas as pd
data= pd.read_csv('Dry_Bean_Dataset - Altered.csv')
```

In [3]:
```
data
```

Out[3]:

	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	roundness	Comp
0	28395	610.291	208.178117	173.888747	1.197191	0.549812	28715.0	190.141097	0.763923	0.988856	0.958027	
1	28734	638.018	200.524796	182.734419	1.097356	0.411785	29172.0	191.272751	0.783968	0.984986	0.887034	
2	29380	624.110	212.826130	175.931143	1.209713	0.562727	29690.0	193.410904	0.778113	0.989559	0.947849	
3	30008	645.884	210.557999	182.516516	1.153638	0.498616	30724.0	195.467062	0.782681	0.976696	0.903936	
4	30140	620.134	201.847882	190.279279	1.060798	0.333680	30417.0	195.896503	0.773098	0.990893	NaN	

2. Cleaning the data set:



jupyter Prajwal M S-jNotebook2 Last Checkpoint: Last Saturday at 13:10 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

13607	42101	757.499	NaN	190.713136	1.476439	0.735702	42494.0	231.526798	0.799943	0.990752	0.922015	
13608	42139	759.321	281.539928	191.187979	1.472582	0.734065	42569.0	231.631261	0.729932	0.989899	0.918424	
13609	42147	763.779	283.382636	190.275731	1.489326	0.741055	42667.0	231.653248	0.705389	0.987813	0.907906	
13610	42159	772.237	295.142741	182.204716	1.619841	0.786693	42600.0	231.686223	0.788962	0.989648	0.888380	
13611	Citation Request :\nKOKLU, M. and OZKAN, I.A.,...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

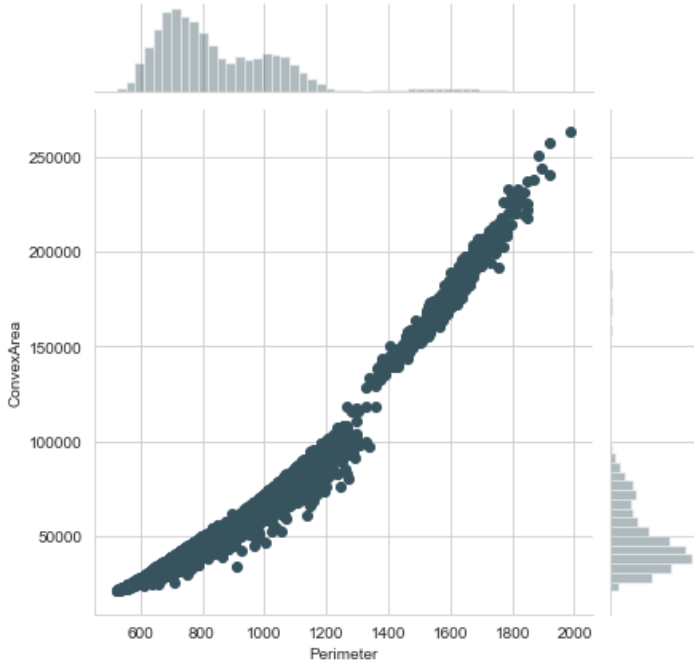
13612 rows × 17 columns

In [8]:
```
data.dropna(how="all")
```

Out[8]:

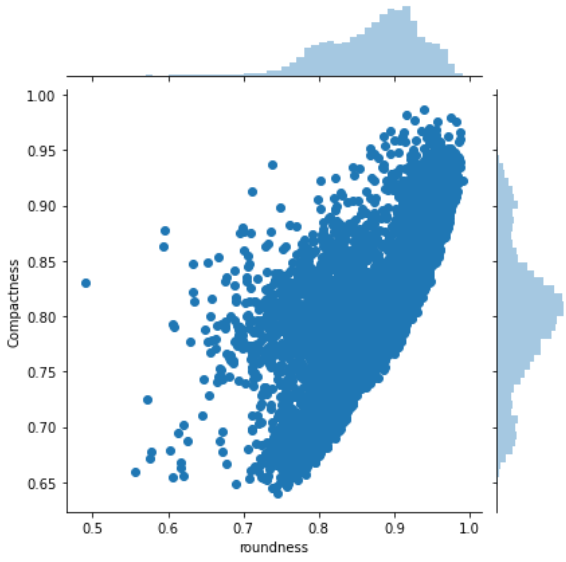
	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	roundness	Comp
0	28395	610.291	208.178117	173.888747	1.197191	0.549812	28715.0	190.141097	0.763923	0.988856	0.958027	
1	28734	638.018	200.524796	182.734419	1.097356	0.411785	29172.0	191.272751	0.783968	0.984986	0.887034	
2	29380	624.110	212.826130	175.931143	1.209713	0.562727	29690.0	193.410904	0.778113	0.989559	0.947849	
3	30008	645.884	210.557999	182.516516	1.153638	0.498616	30724.0	195.467062	0.782681	0.976696	0.903936	
4	30140	620.134	201.847882	190.279279	1.060798	0.333680	30417.0	195.896503	0.773098	0.990893	NaN	

3. Data Analysis:

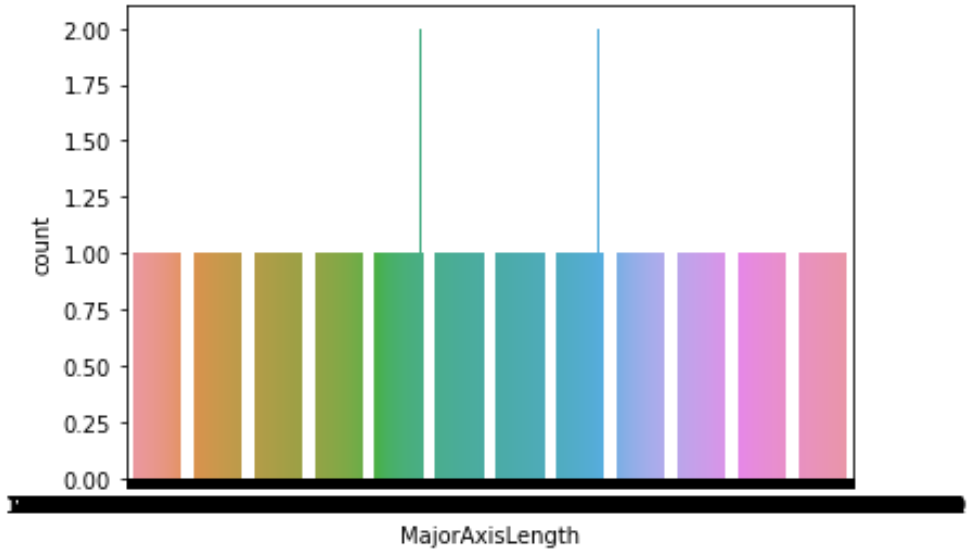


i) From the above graph we can conclude that ConvexArea Increses with the Increases in the Perimeter.

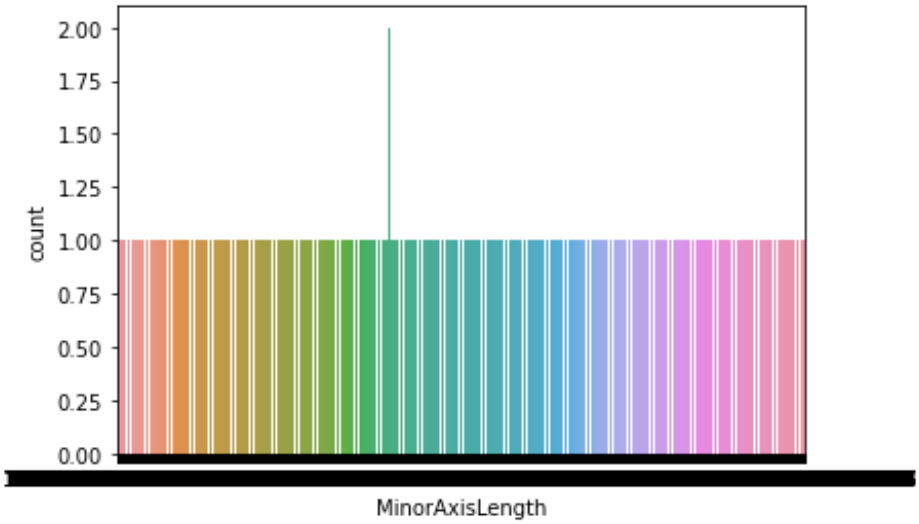
ii) There are very few Data Available above 1800 Perimeter.



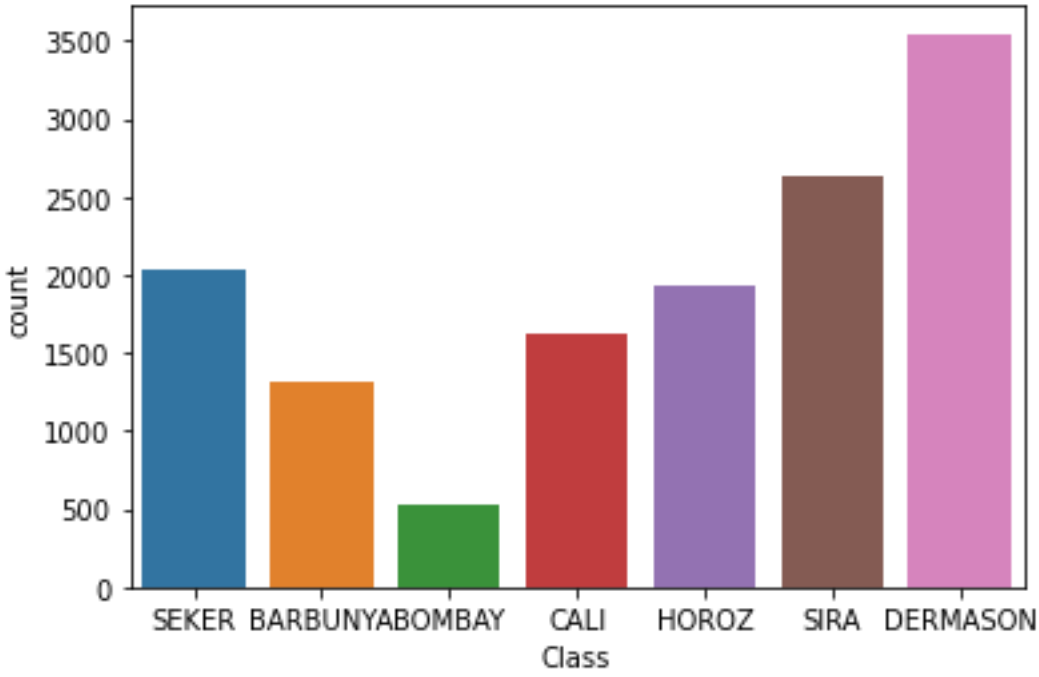
i).The above graph is a Jointplot of Roundness Vs Compactness.
ii).Most of the Dataset lies between the 0.7-0.8(roundness) and 0.90-0.75(compactness).



i) From the above graph we can conclude that the MajorAxisLength of the Data set is 1.00.

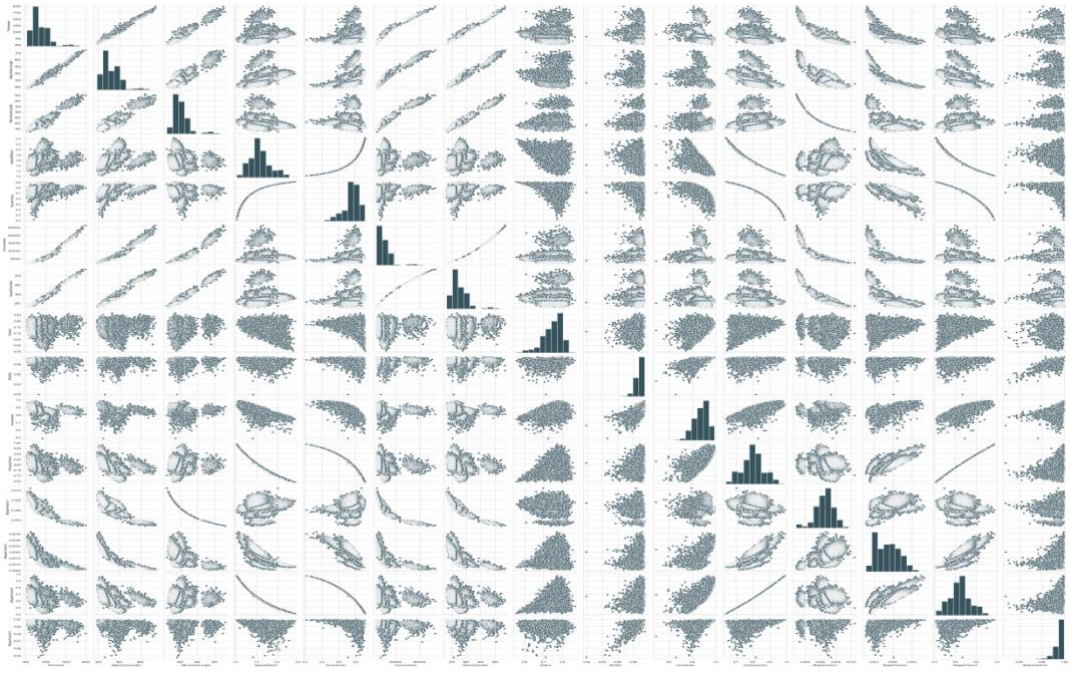


i).From the above graph we can conclude that the MinorAxisLengh of the Data set is 1.00.

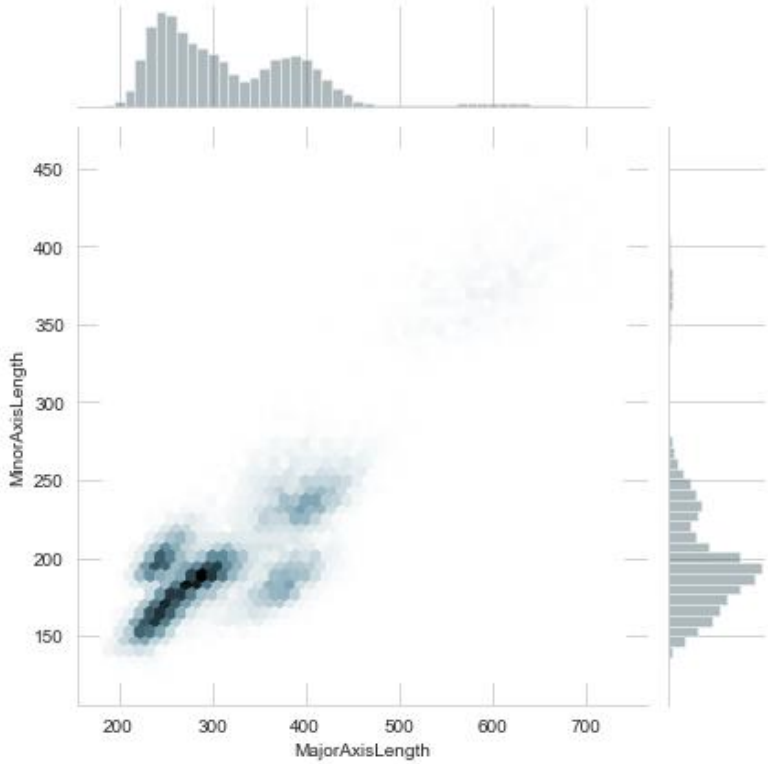


i) The Most of the Dry Bean Lies under the Dermason Class. There are 3500 Dermason Class of Data.

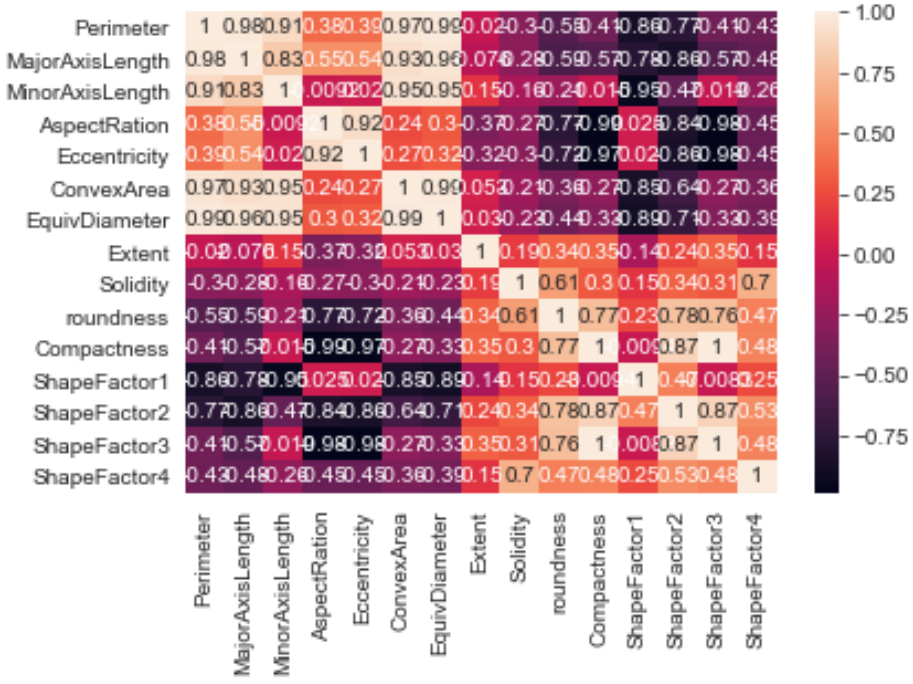
ii) The Bombay Class is the least one. There are 500 Bomabay Class of Data.



i). Above Graph is the pairplot of the Dry Bean Dataset, it describe the multiple pairwise bivariate distribution in a dataset.



i). The above graph is the Jointplot of MinorAxisLength Vs MajorAxisLength, it shows that there are maximum of MajorAxisLength data at 200-300 and there are maximum of Minor axislength data at150-200.



i) The above graph is the heatmap of the Dry Bean Dataset,it shows the relation between the two variables in the dataset, one plotted on each axis

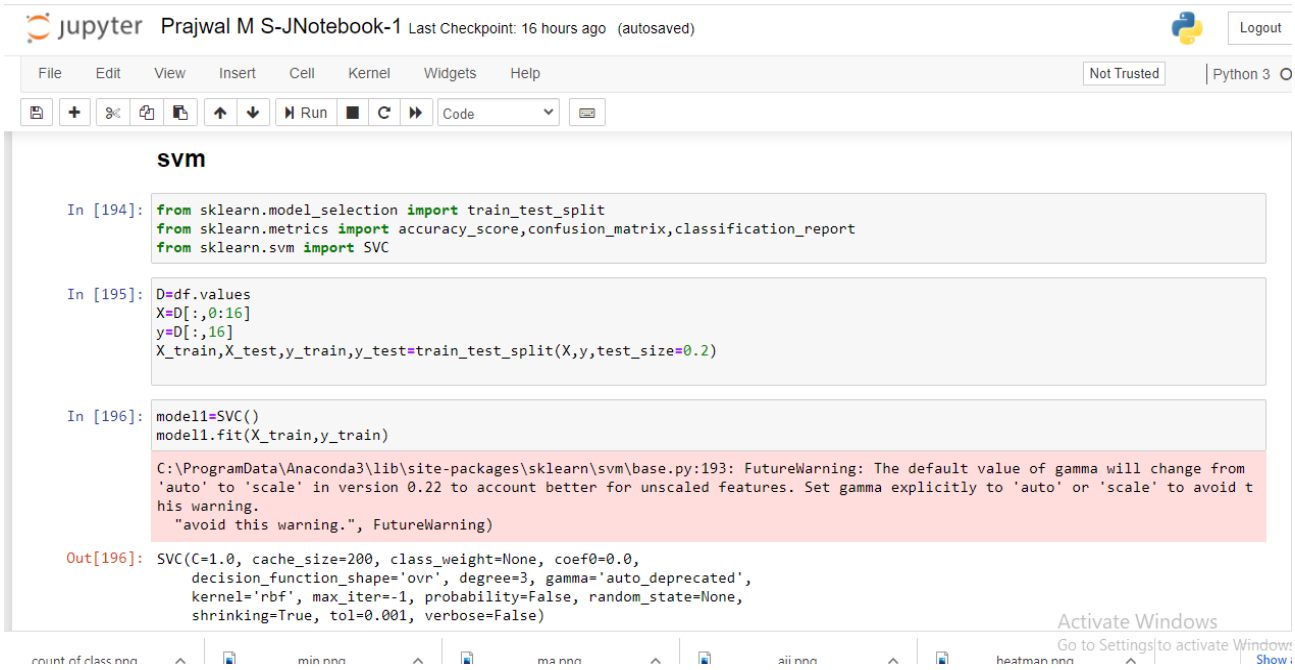
Machine Learning Models

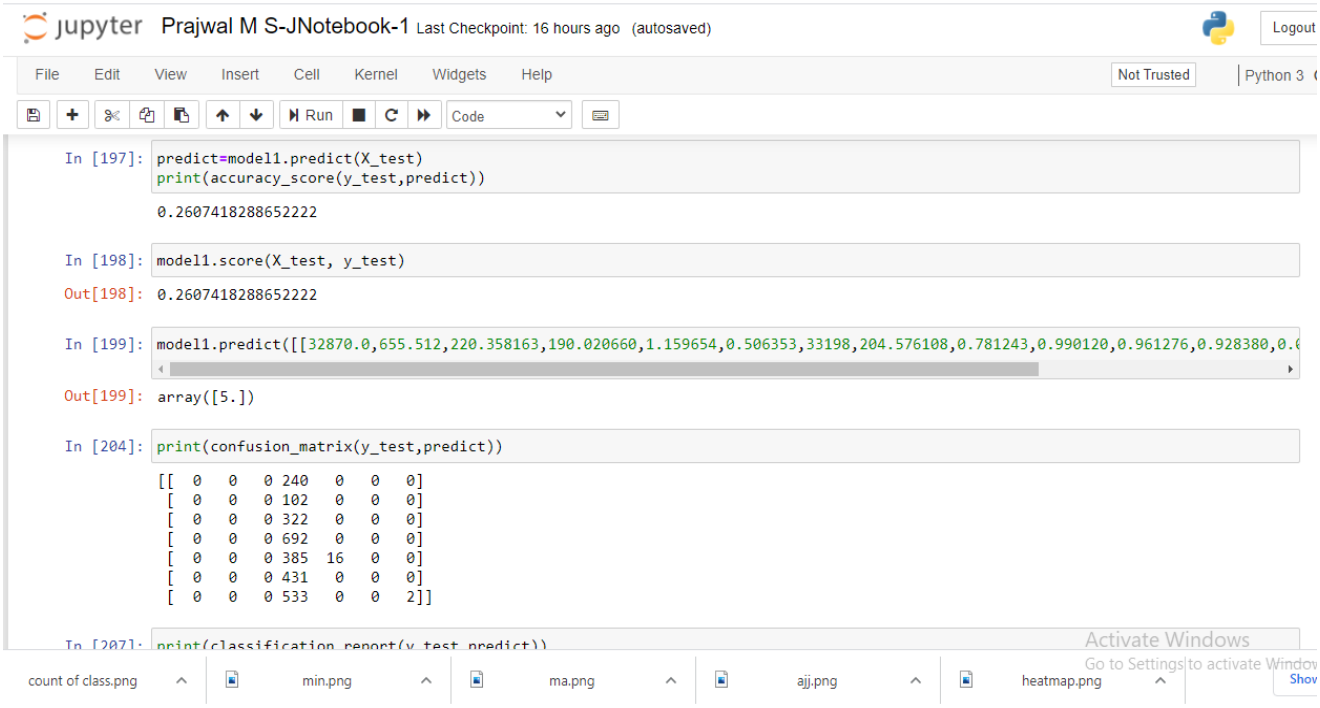
We need to use Regression algorithms on the given data set in order predict the Accuracy of the Dry Bean.

There are many Regression algorithms, like

A. Linear Regression
B. Lasso Regression
C. Support Vector Regression
D. Decision Tree Regression
E. Logistic Regression.

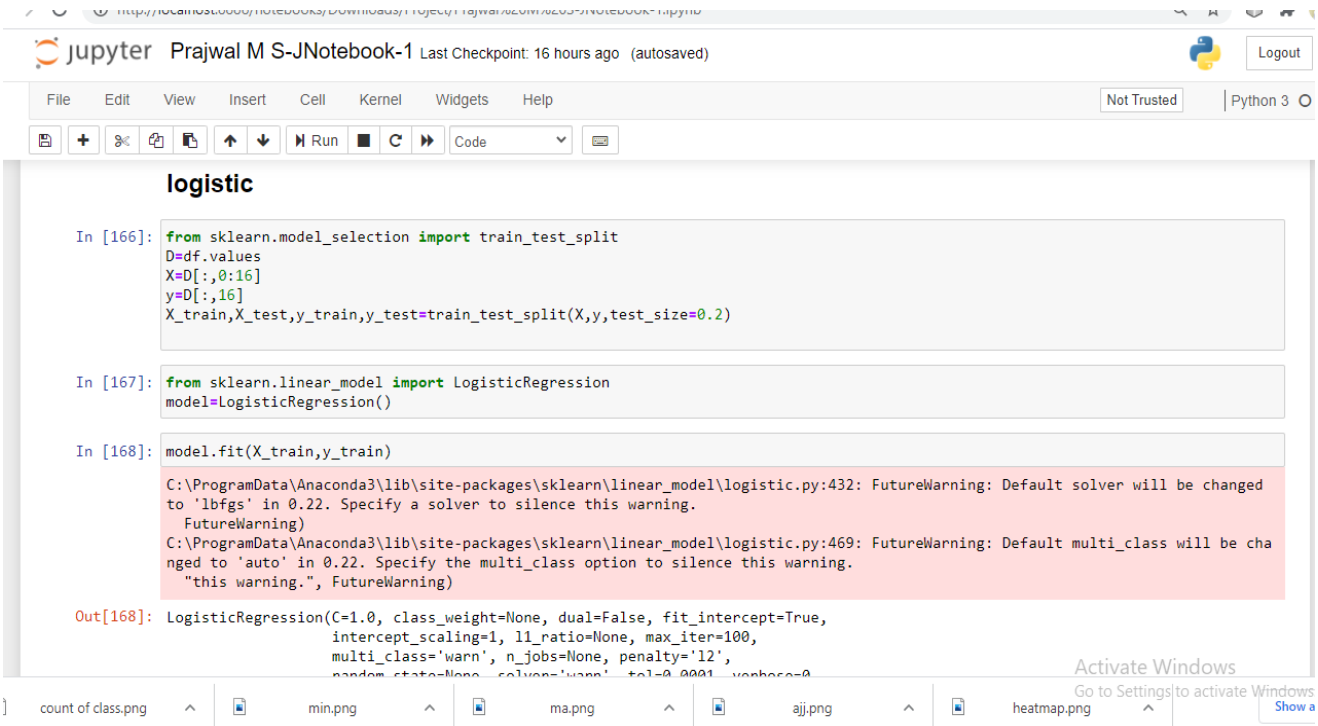
1) Support Vector Machine (SVM)

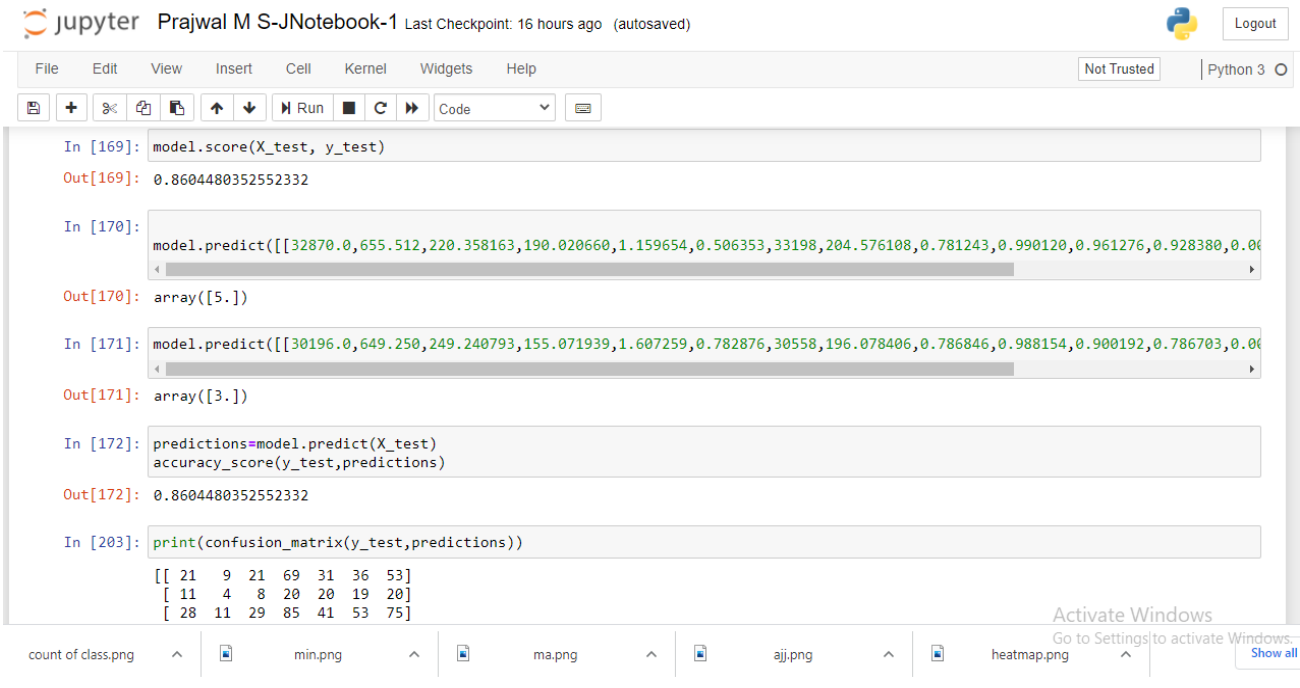




The r2_score of SVM model is **0.26074**

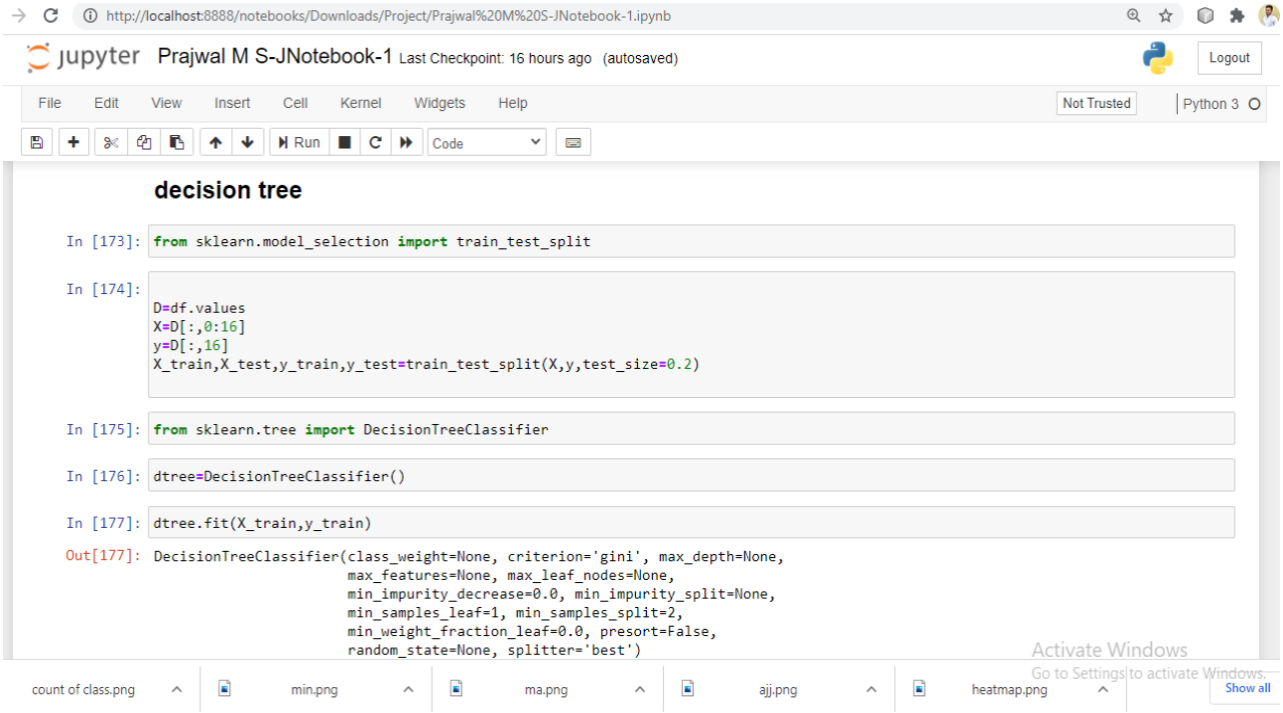
2) Logistic Regression

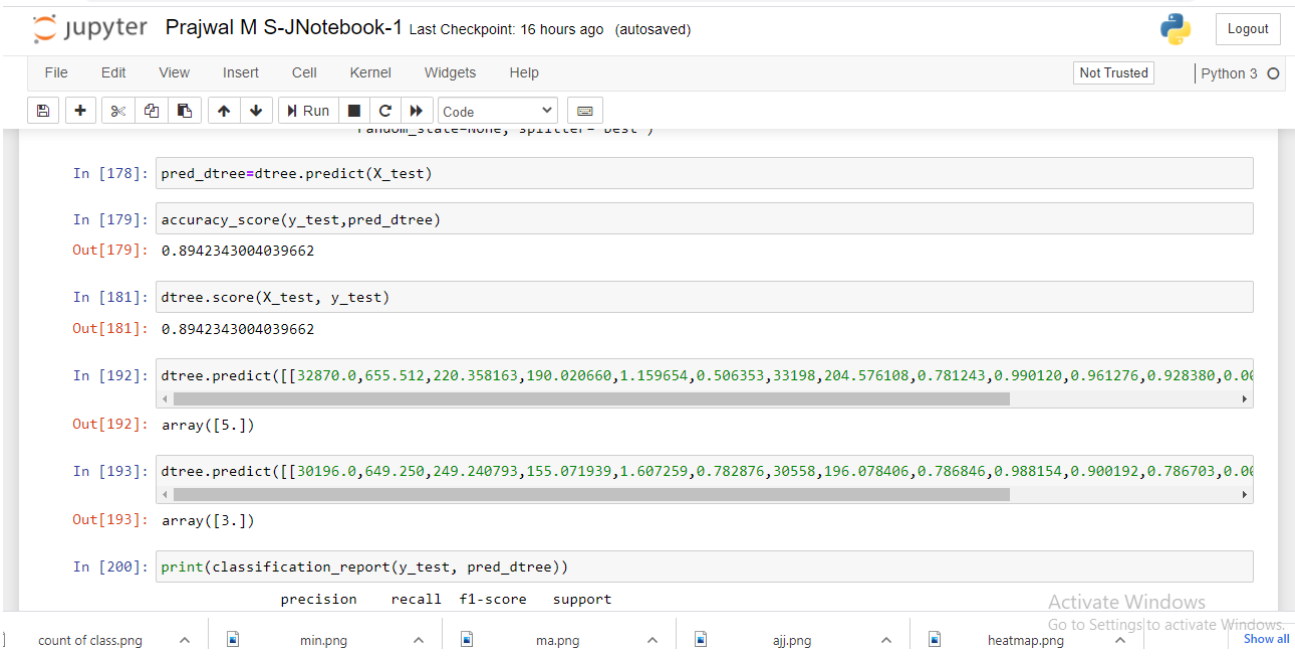




The r2_score of Logistic Regression model is **0.86044**

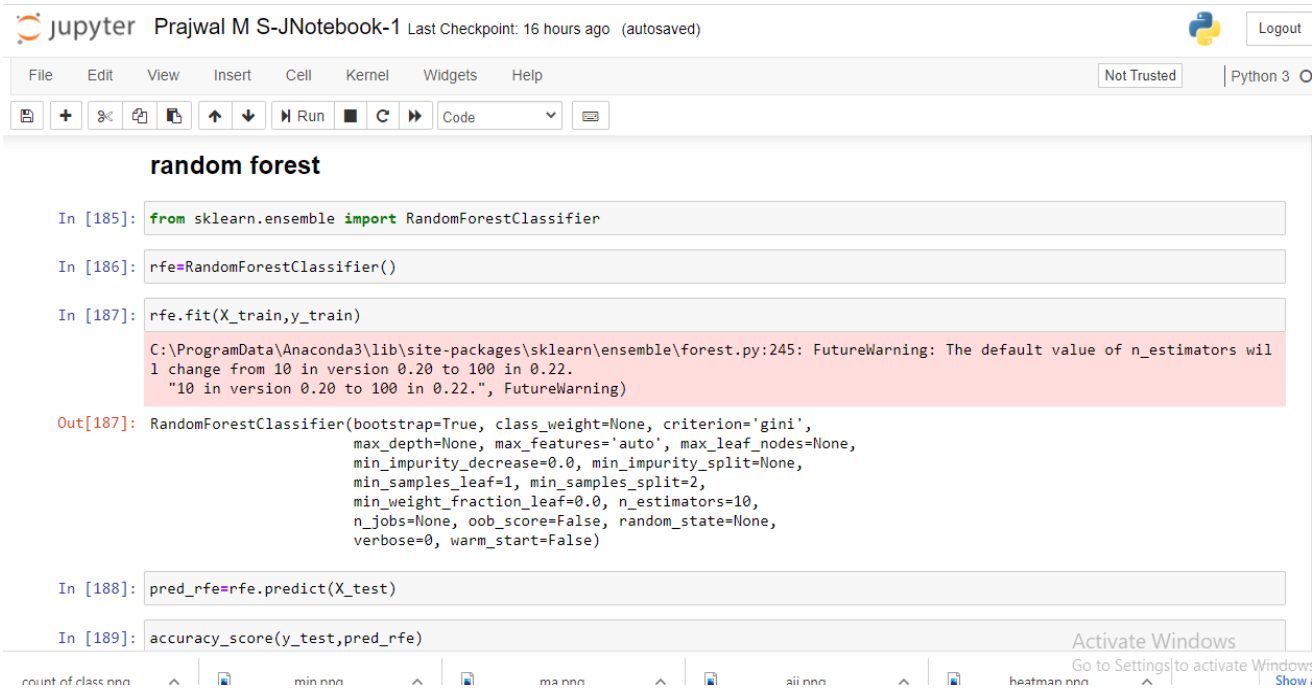
3) Decision Tree





The r2_score of Decision Tree model is **0.894234**

4) Random Forest



The r2_score of Random Forest is **0.915901**

ML Model Chart

Serial Number	Algorithm Name	R2_score
1	**Support vector Machine**	**0.260742**
2	**Logistic Regression**	**0.86044**
3	**Decision Tree**	**0.894234**
4	**Random Forest**	**0.91590**

Hurdles

a) I was getting a negative r2_score for SVM model. Then I realized that if I change the parameter, then the r2_score will improve. After doing so I got better r2_score.

b) While performing Grid search I had given the Kernel value as Logistic & it was taking forever to run that particular line. Logistic gives the best fit, but it takes too much time to run. So I changed the Kernel to its default value, then there was no issue.

Conclusion

Analysis of Dry Bean Dataset was a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data.

Data cleaning is one of the processes that increases prediction performance.

On the whole, this internship was a useful experience. I have gained new knowledge, skills and met many new people. I achieved several of my learning goals.

The internship was also good to find out what my strengths and weaknesses are. This helped me to define what skills and knowledge I have to improve in the coming time.

Bibliography

1. Seaborn Jupyter Notebook given by the tutor.
2. Support Vector Machines Jupyter Notebook given by the tutor.
3. Linear Regression with Sklearn Jupyter Notebook given by the tutor.