

**AVOCADO PROJECT**



Submitted by:

CHETHANA M

Batch ID : 1843

DataTrained

**INTRODUCTION**

**Avocado is a fruit consumed by people heavily in the United States and now a days it is being consumed most of the countries due to its health benefits.**

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados.

Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags.

The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

**Goals :**

* Predicting Average price using Regressors
* Predicting region using Classifiers

Our main aim today is to make a model which can give us a good prediction on the Average price and Region of the Avocado based on other variables. We are going to use Linear Regression, Support vector regressor, Decision Tree Regressor, K-Neighbors Regressor, Ridge and Lasso, Random forest classifier, DecissionTreeClassifier, KNeighborsClassifier to build the different models for this dataset and see which model gives us a good accuracy.

**Steps Involved in this Project :**

1. Understanding the Problem

2. Data Analysis

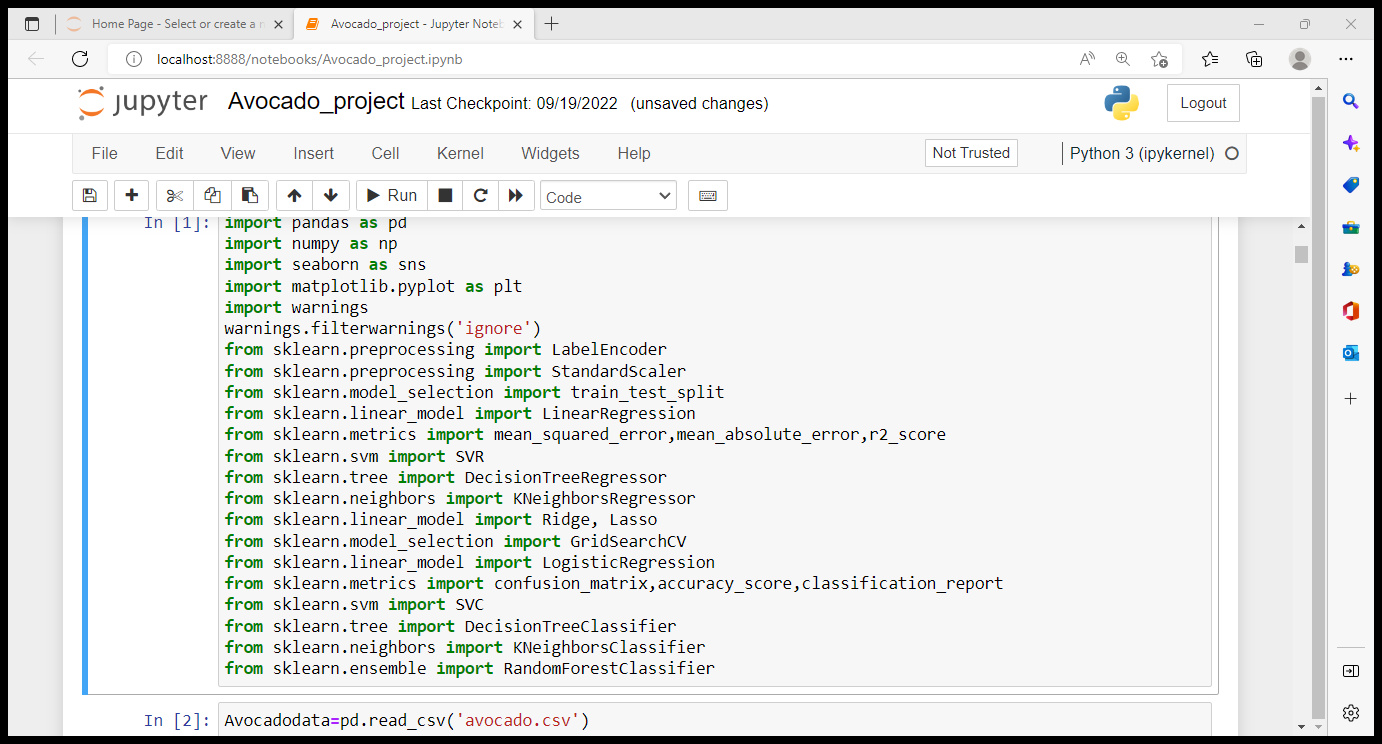
3. EDA Concluding Remark

4. Pre-Processing

5. Building Machine Learning Models and finalizing the best model

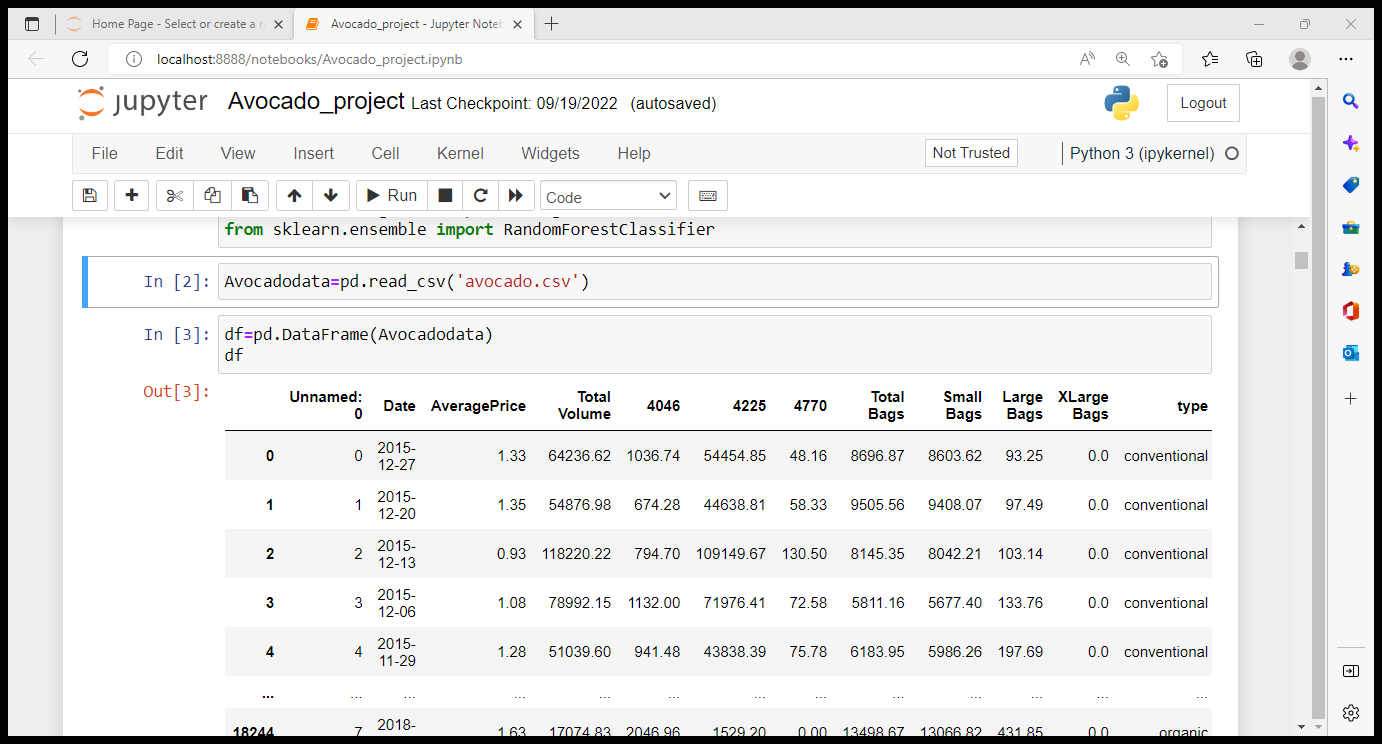
6. Concluding Remarks

**Loading the Libraries :**

****

**EDA and Visualization**

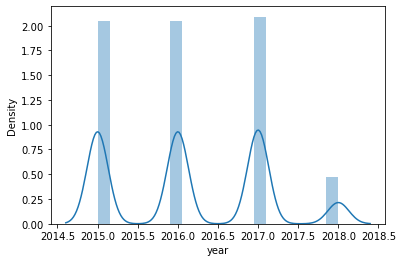
**Loading the data :**



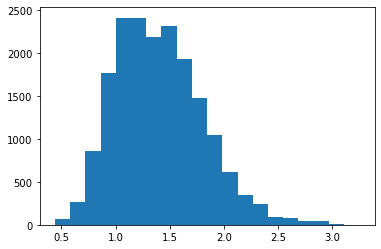
**Checking for features of dataset :**

The data set has 18,249 rows and 14 variables. The variables in the data set and their dtypes are,

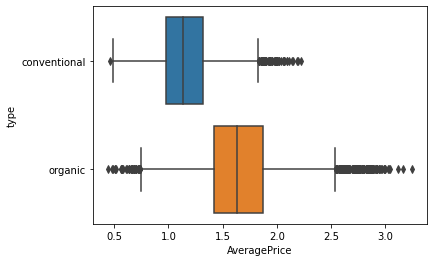
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Dtype | Description | Missing values | Outliers | Skew |
| 1.Unnamed | int64 | Which is just an index given in dataset, we can drop it in further steps | - | - | - |
| 2.Date | object | The date of the observation | No | Yes | No |
| 3.AveragePrice | float64 | the average price of a single avocado | No | Yes | No |
| 4.Total Volume | float64 | Total number of avocados sold | No | Yes | Yes |
| 5.4046 | float64 | Total number of avocados with PLU 4046 sold | No | Yes | Yes |
| 6.4225 | float64 | Total number of avocados with PLU 4225 sold | No | Yes | Yes |
| 7.4770 | float64 | Total number of avocados with PLU 4770 sold | No | Yes | Yes |
| 8.Total Bags | float64 | Total No of bags sold | No | Yes | Yes |
| 9.Small Bags | float64 | Total No of small bags sold | No | Yes | Yes |
| 10.Large Bags | float64 | Total No of Large bags sold | No | Yes | Yes |
| 11.XLarge Bags | float64 | Total No of XLarge bags sold | No | Yes | Yes |
| 12.Type | object | Type of Avocados | No | No | No |
| 13.Year | int64 | The year in which Avocado was sold | No | No | No |
| 14.Region | object | Region the Avocado originated from | No | No | No |



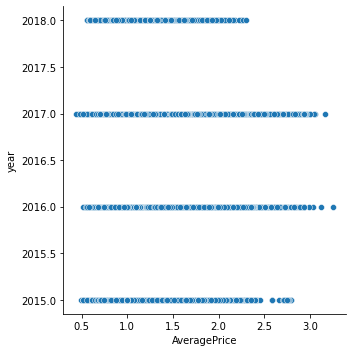
Graph tells that the number of avocados sold in the year 2015, 2016, 2017 are almost equal in number and 2018 records are lesser in number.



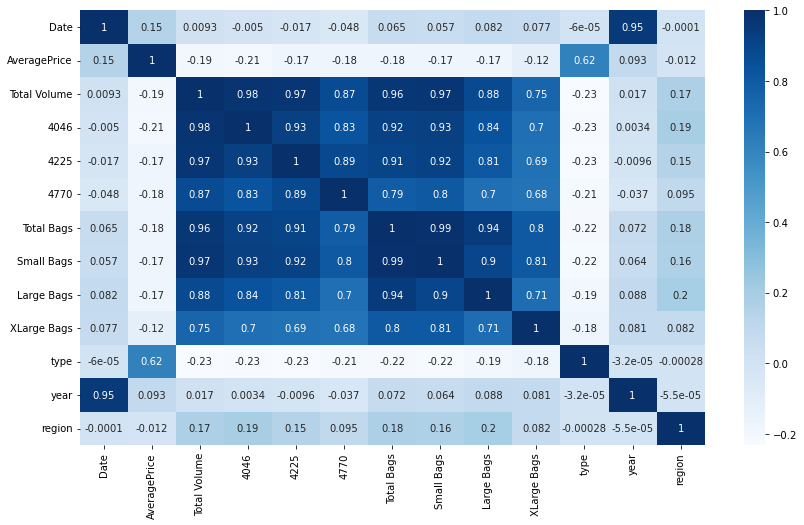
We can observe that maximum price of avocado ranges from 1 to 1.6



From the above plot we can observe that the price is more for organic avocado and less for conventional avocado.



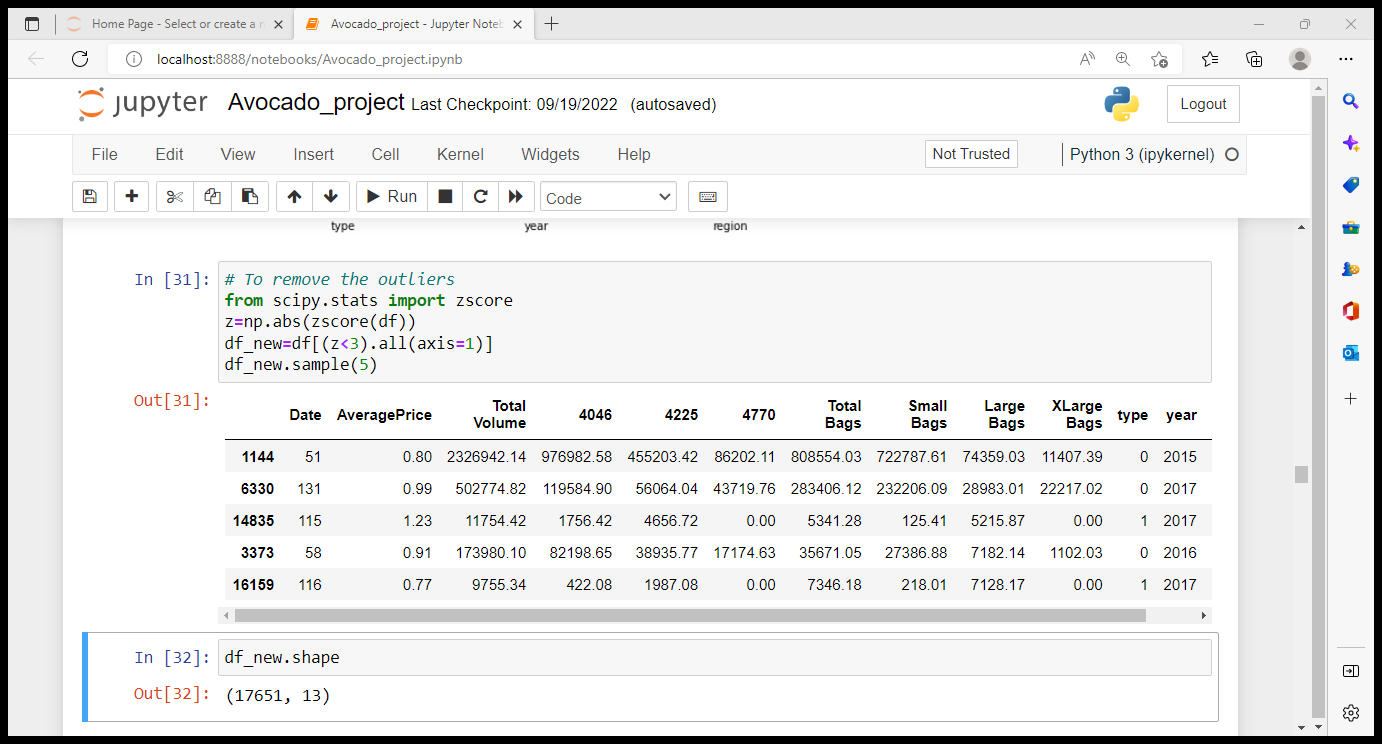
We can observe that the AveragePrice reaches highest value in the year 2017 and 2016 than in 2015 and 2018.



Observations:

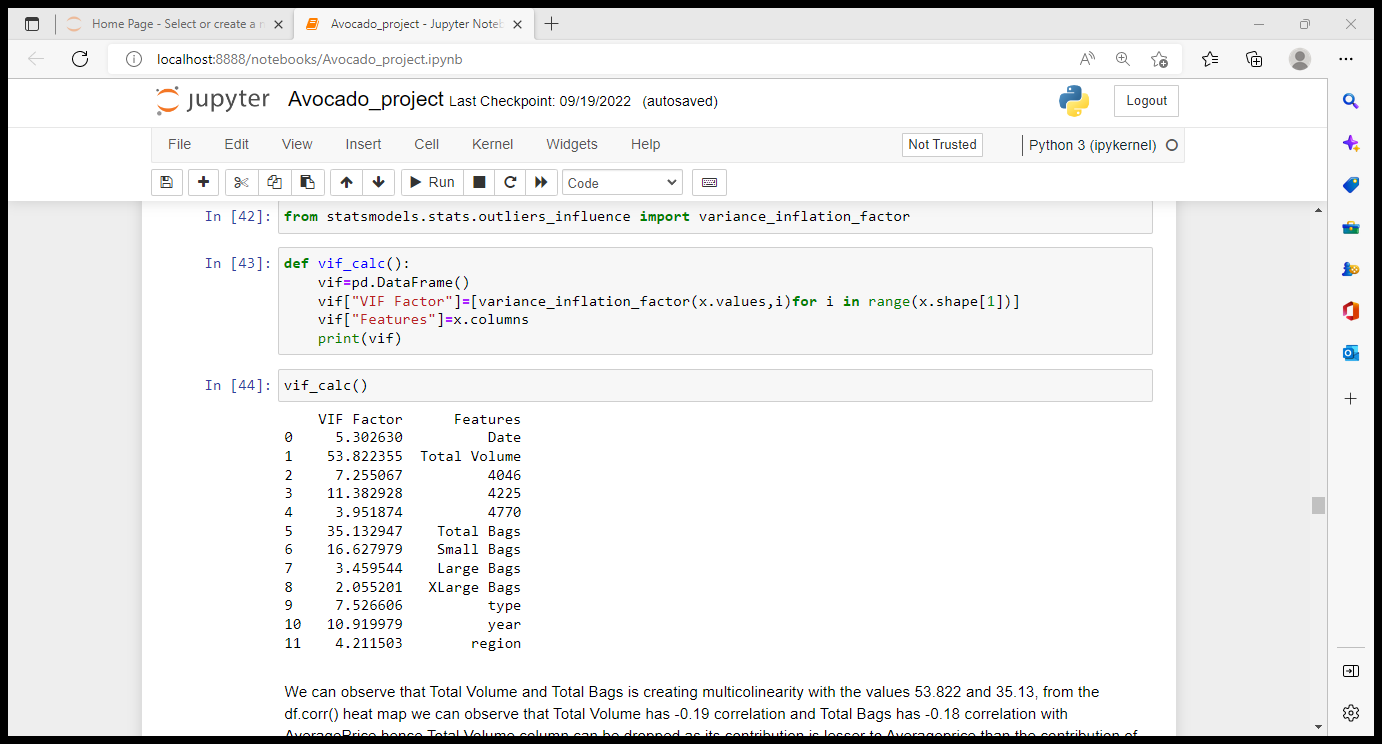
* AveragePrice is high correlation only with type and has negative correlation with other variables except region.
* Total bags has a good correlation with small bags and total volume.
* Total volume and 4046 are highly correlated.
* Region has positive correlation with XLarge Bags, Small Bags, Large Bags, Total Bags, 4046, 4225, 4770 and Total Volume.

As we discussed before now we can encode object type data to integer data using label encoder remove Outliers using Z-score method and remove skewness using 'yeo-johnson' method. Once after doing all the pre-processing the size of data we have is 17651 rows and 13 columns the data look like below.



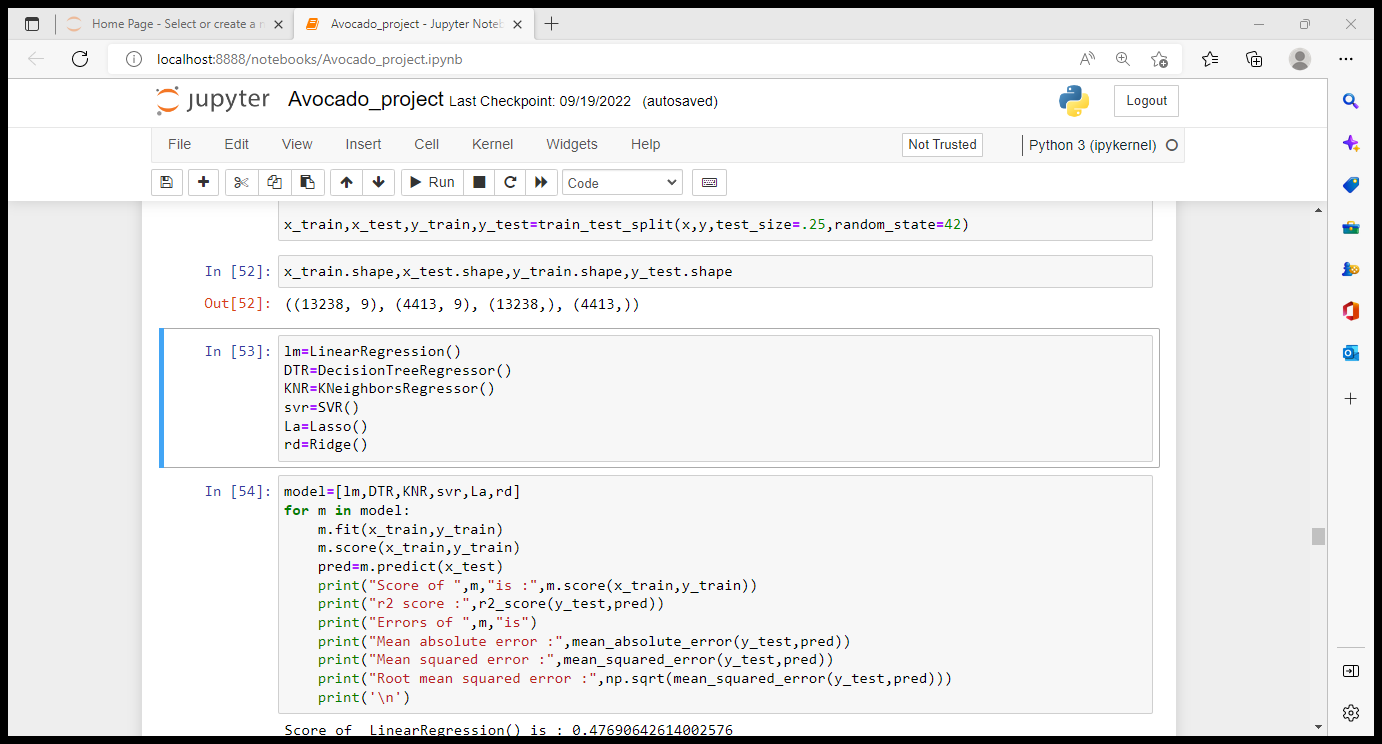
# Avocado AveragePrice Prediction using regressors

We can now split the input and target variables and check for multicollinearity present in the data



As we can observe that Total Volume and Total Bags are creating multicollinearity in the data we can drop of Total volume due to its lesser contribution to target variable. Again Total Bags is creating multicollinearity with Small Bags hence Total bags can be dropped due to its lesser contribution to target variable than that of small bags. Now the column Year is causing multicollinearity with Date, As the date has nothing to do with deciding price we can drop off date variable, Hence the multicollinearity is removed from the dataset.

**Splitting the data for training and testing the models to predict AveragePrice :**



The model Score and r2 score od the above built models are as follows,

Score of LinearRegression() is : 0.47690642614002576

r2 score : 0.4690412634055062

Score of DecisionTreeRegressor() is : 1.0

r2 score : 0.6872103909929783

Score of KNeighborsRegressor() is : 0.8941862860604416

r2 score : 0.8290833823087831

Score of SVR() is : 0.0397605353460464

r2 score : 0.040570823280558166

Score of Lasso() is : 0.0

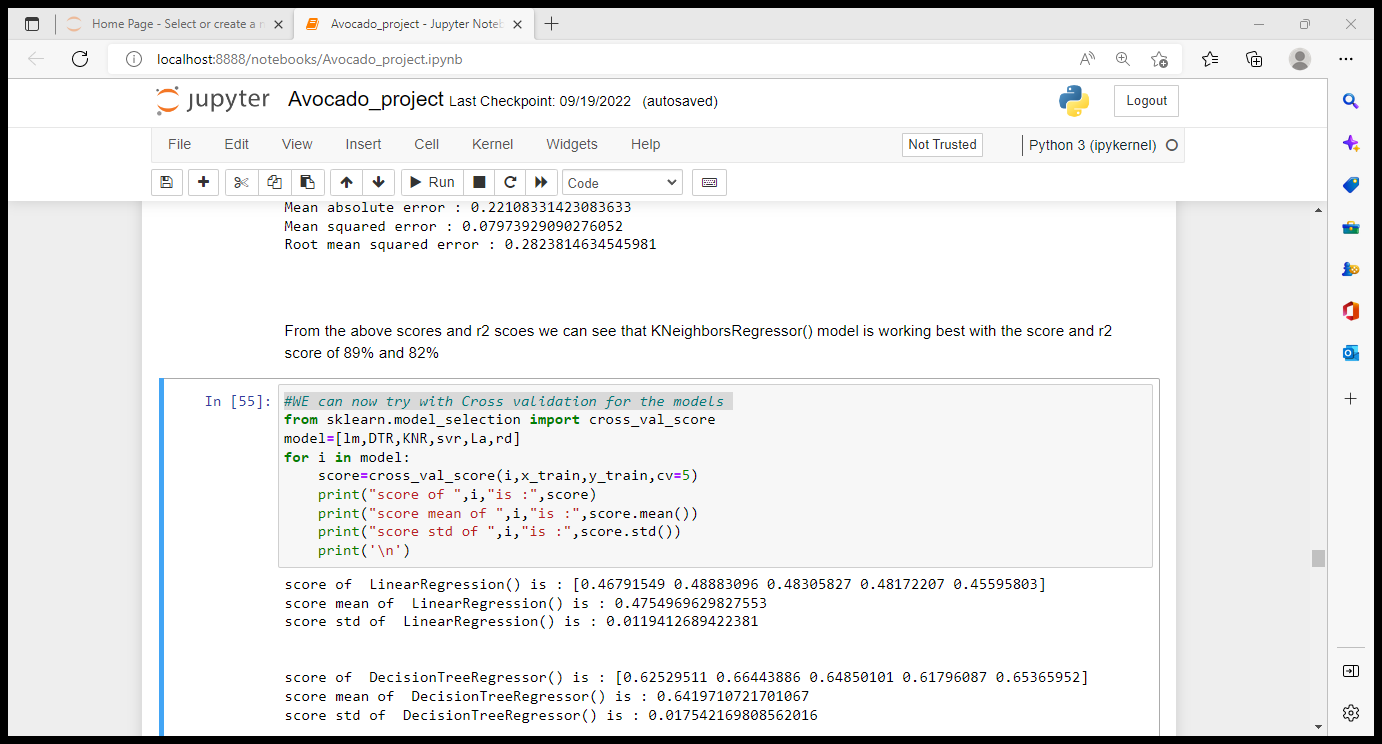
r2 score : -5.600800817040508e-05

Score of Ridge() is : 0.4769063251976897

r2 score : 0.4690435598876084

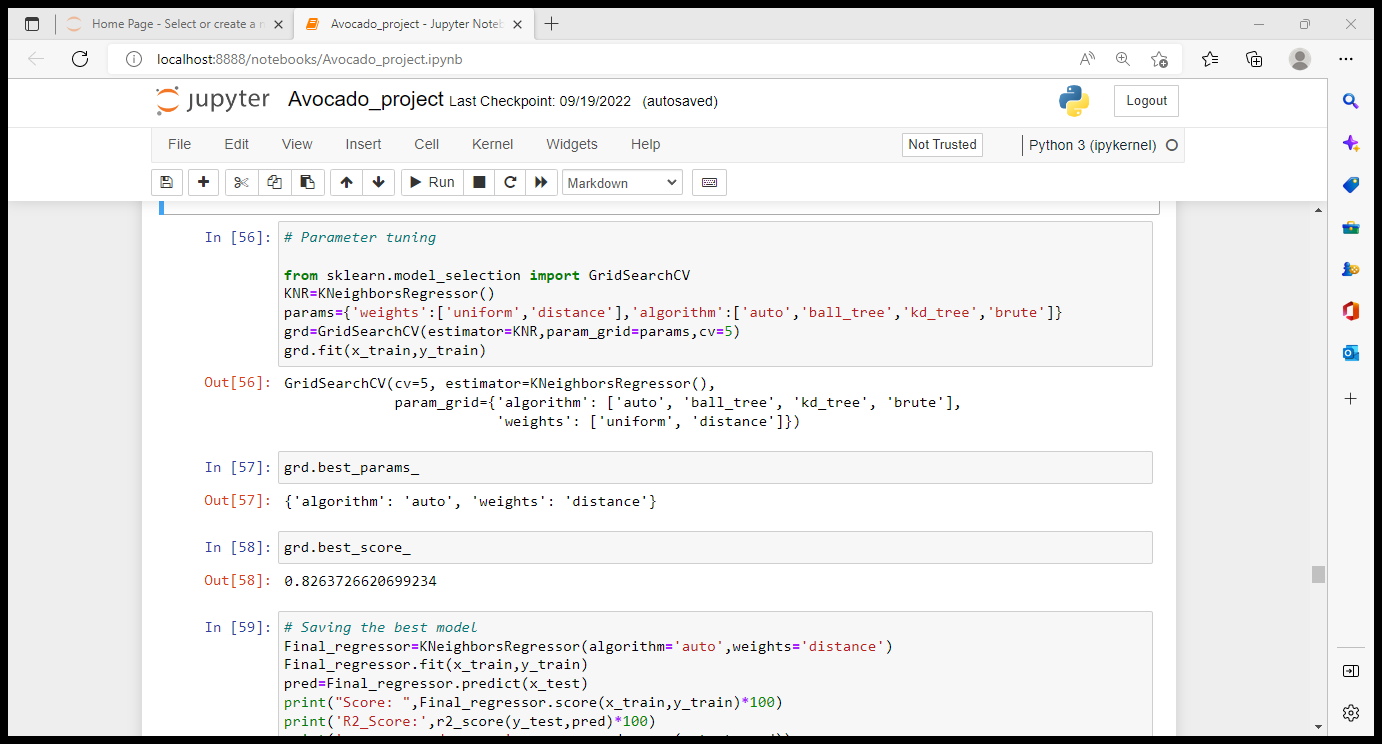
From the above scores and r2 scores we can see that KNeighborsRegressor() model is working best with the score and r2 score of 89% and 82%.

We can now try with Cross validation for the models

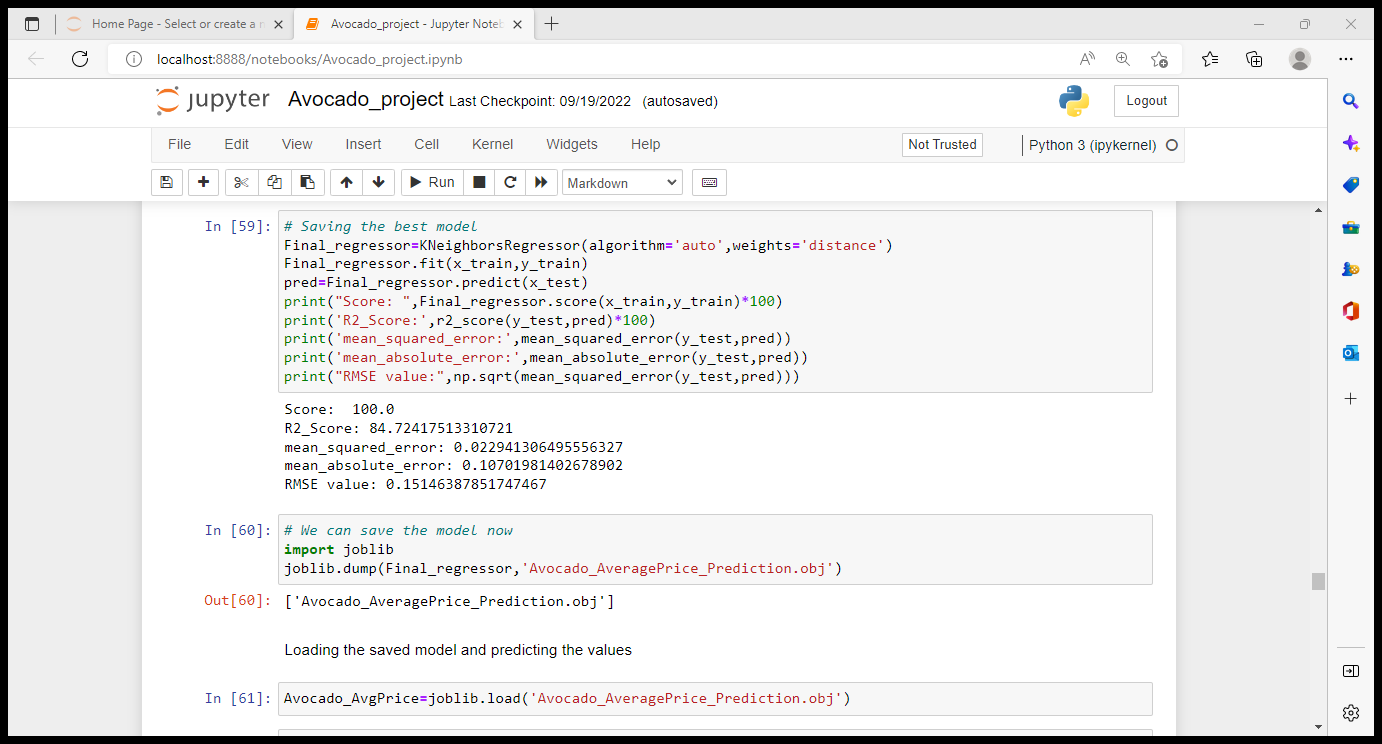


Even with CV we can see that KNeighborsRegressor() model is performing well with mean score of 80%.

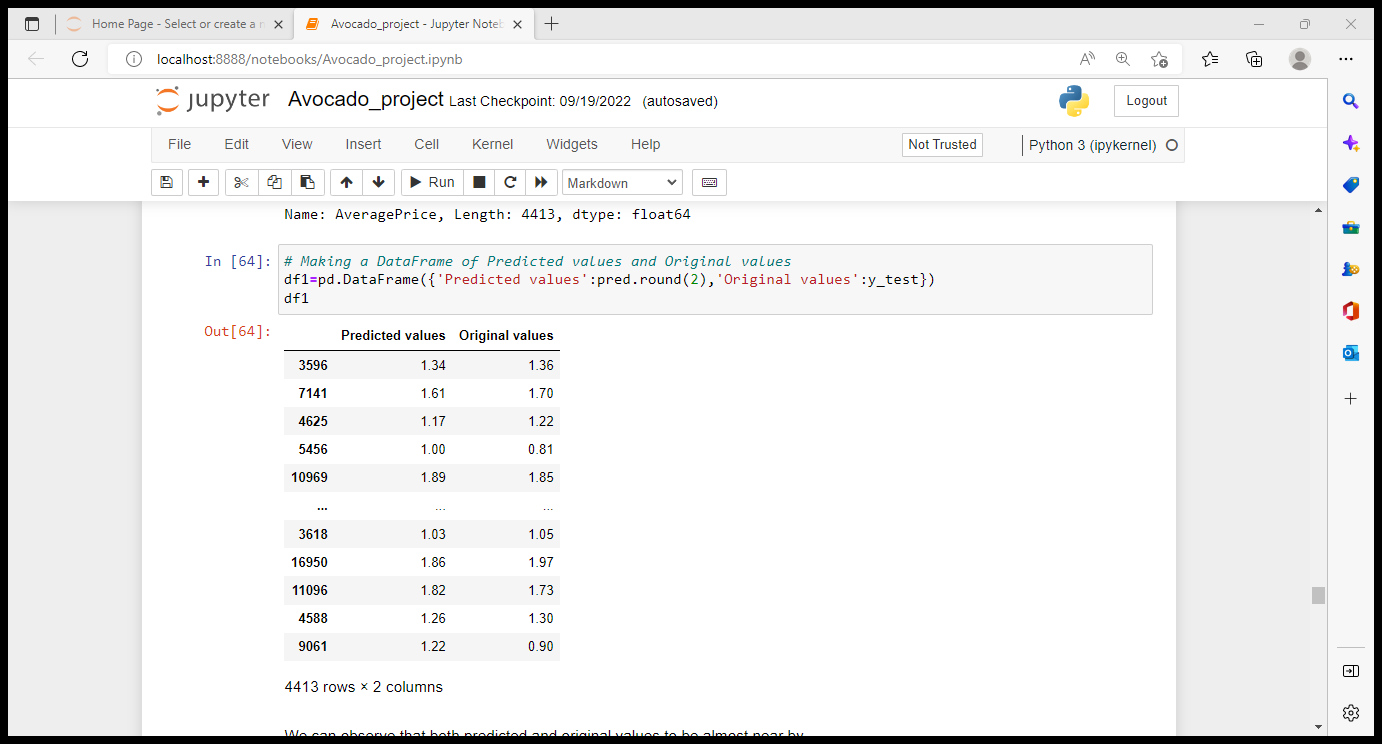
Parameter Tuning :



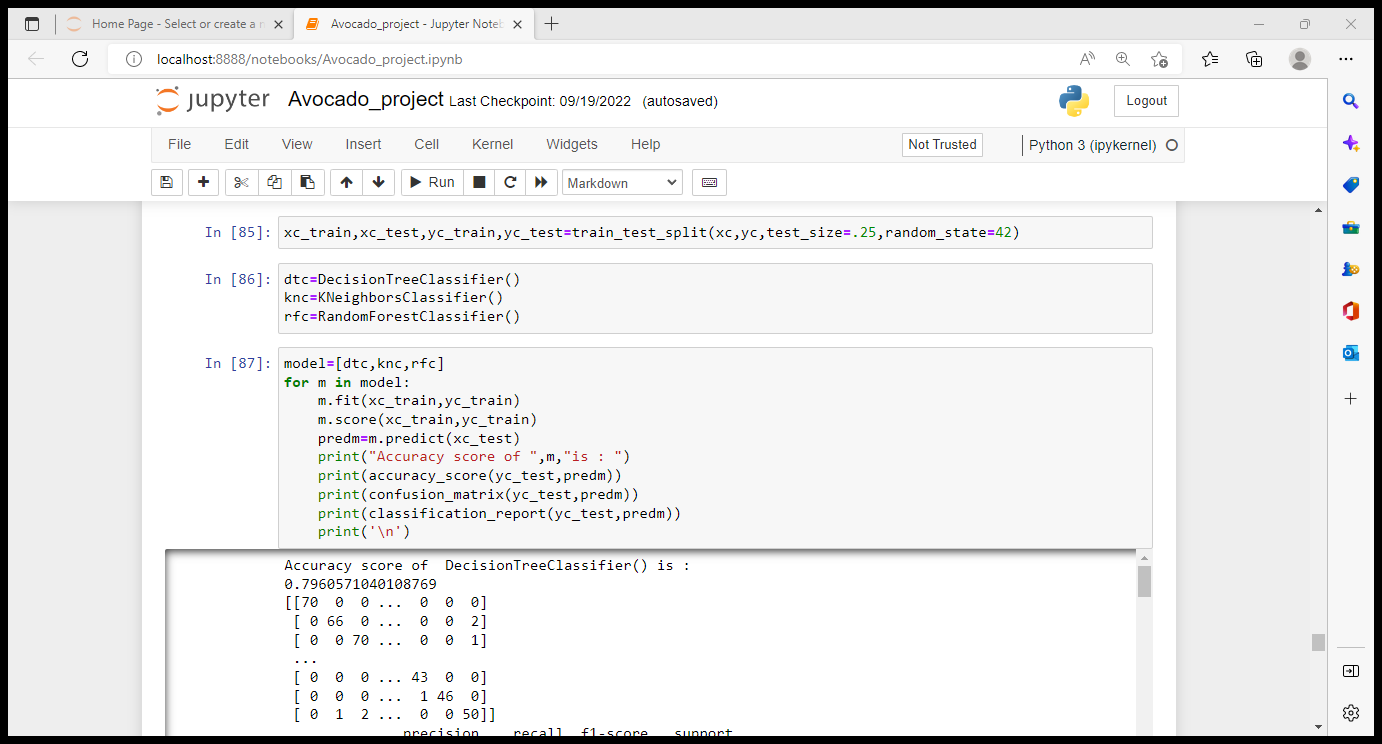
Saving the best model :



Checking for predicted outputs :

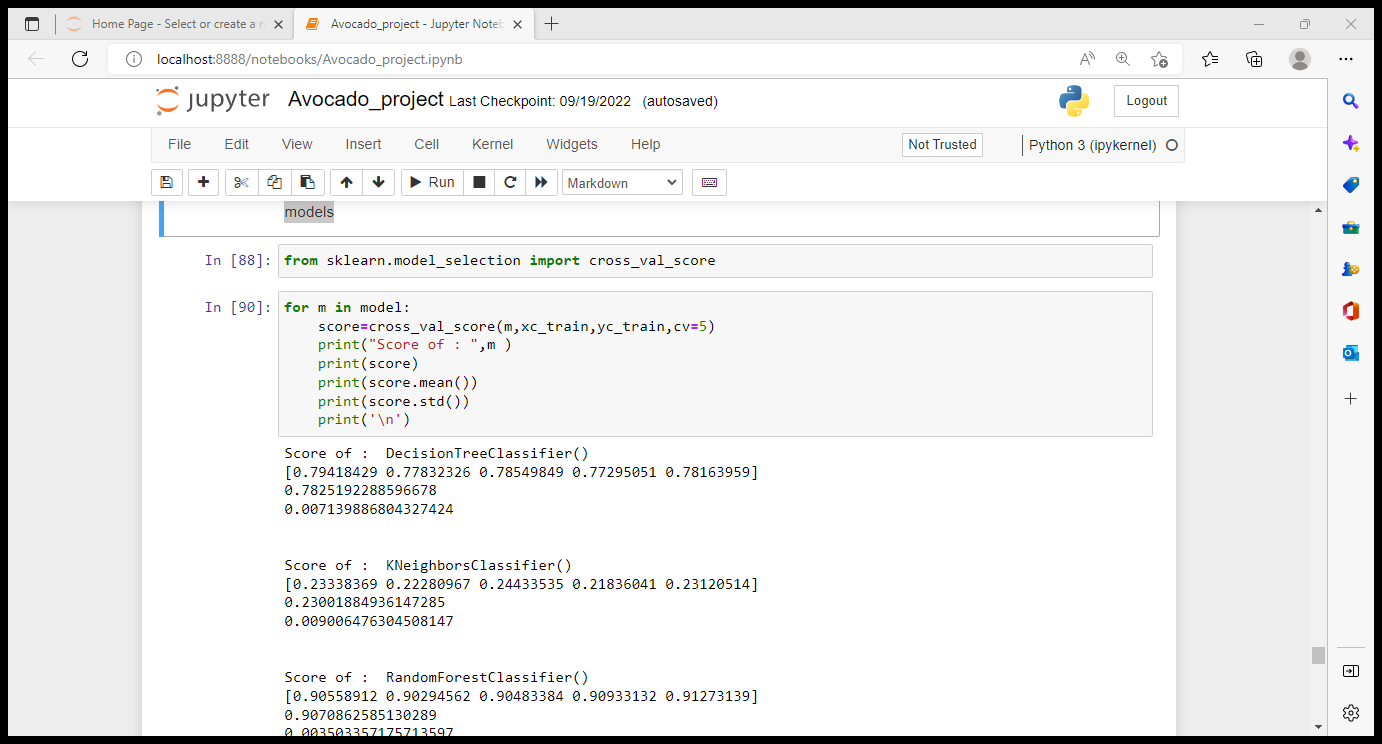


# Avocado Region prediction using classifiers



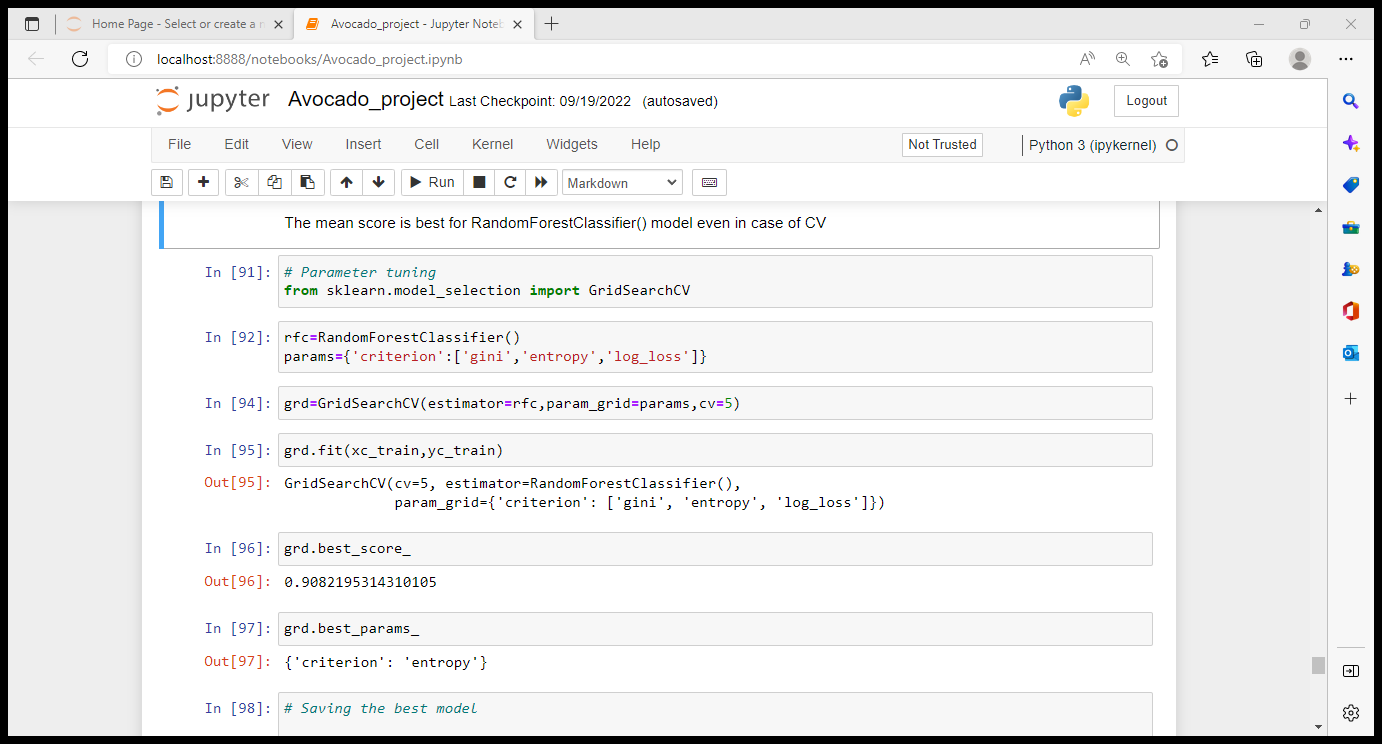
We can consider RandomForestClassifier() to be the best model as its accuracy score is 90% and higher compared to other models.

We can now check for the performance of above models with cross validation.

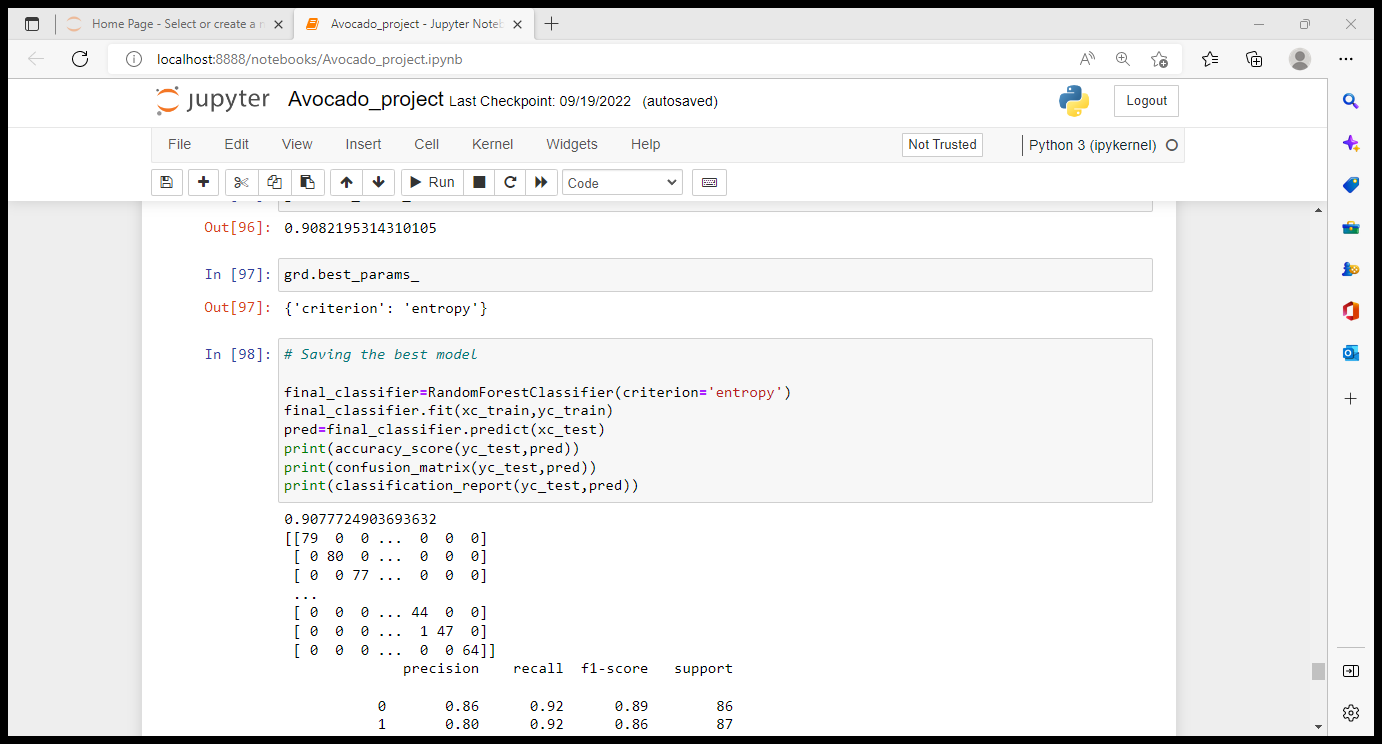


The mean score is best for RandomForestClassifier() model with 90% mean score even in case of Cross validation.

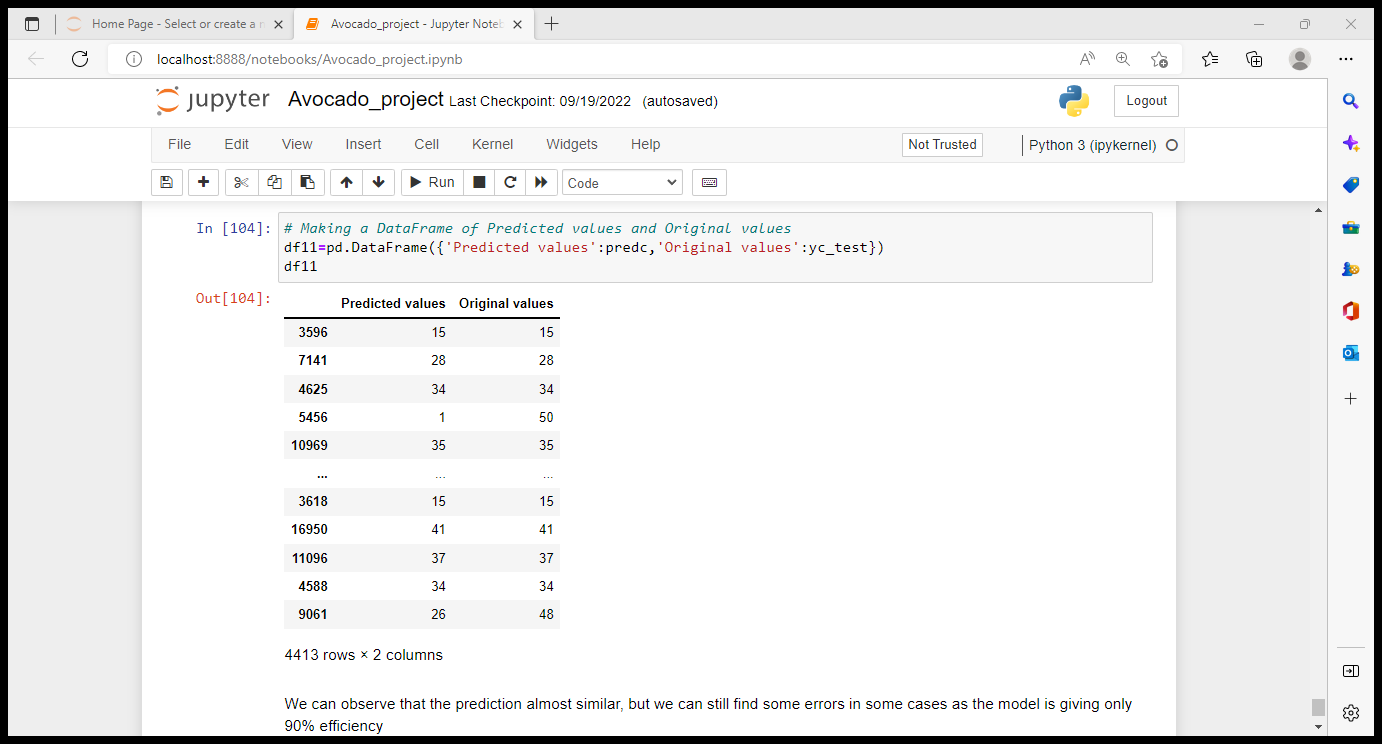
Parameter tuning for RandomForestClassiofier :



Saving the best model to predict region :



Checking the original and Predicted values :



**Conclusion**

Hence we have built different models for predicting both Average price and Region of the Avocado. From the above built models we can observe that in case of Average price prediction we chose KNeighbors Regressor as the best model as its score is 100% with algorithm = auto and weights = distance and r2 score is 84.72%.

In case of predicting region of the Avocado we found RandomForest classifier with criterion = entropy to be the best model as it is giving a score of 90.77%, which is best compared to other models. In the predicted values of region we can observe some errors as the model is giving only 90.7% accuracy and rest 9.3% less accuracy is responsible for the occurrence of the error.