In this case study, I am going to classify given samples of wines as ‘Good’ or ‘Bad’ quality wines based on the provided physiochemical information using supervised machine learning.

Machine learning is subset of ‘Artificial Intelligence’ and ‘Computer Science’ we try to mimic human intelligence through computer modelling which can learn and improve on its own over time. At basic level machine learning uses algorithms to find patterns and then applies the patterns moving forward.

Typically, there are 3 branches of machine learning:

* Supervised learning:  
  Where we train machine learning model on historic data and use the trained model to make predictions on new data.
* Un-supervised learning:  
  Where the model is used to find trends and patterns in the data without prior training on historic data.
* Reinforcement learning:

Where the model learns through trial and error through the interaction with the environment.

**Problem Statement:**

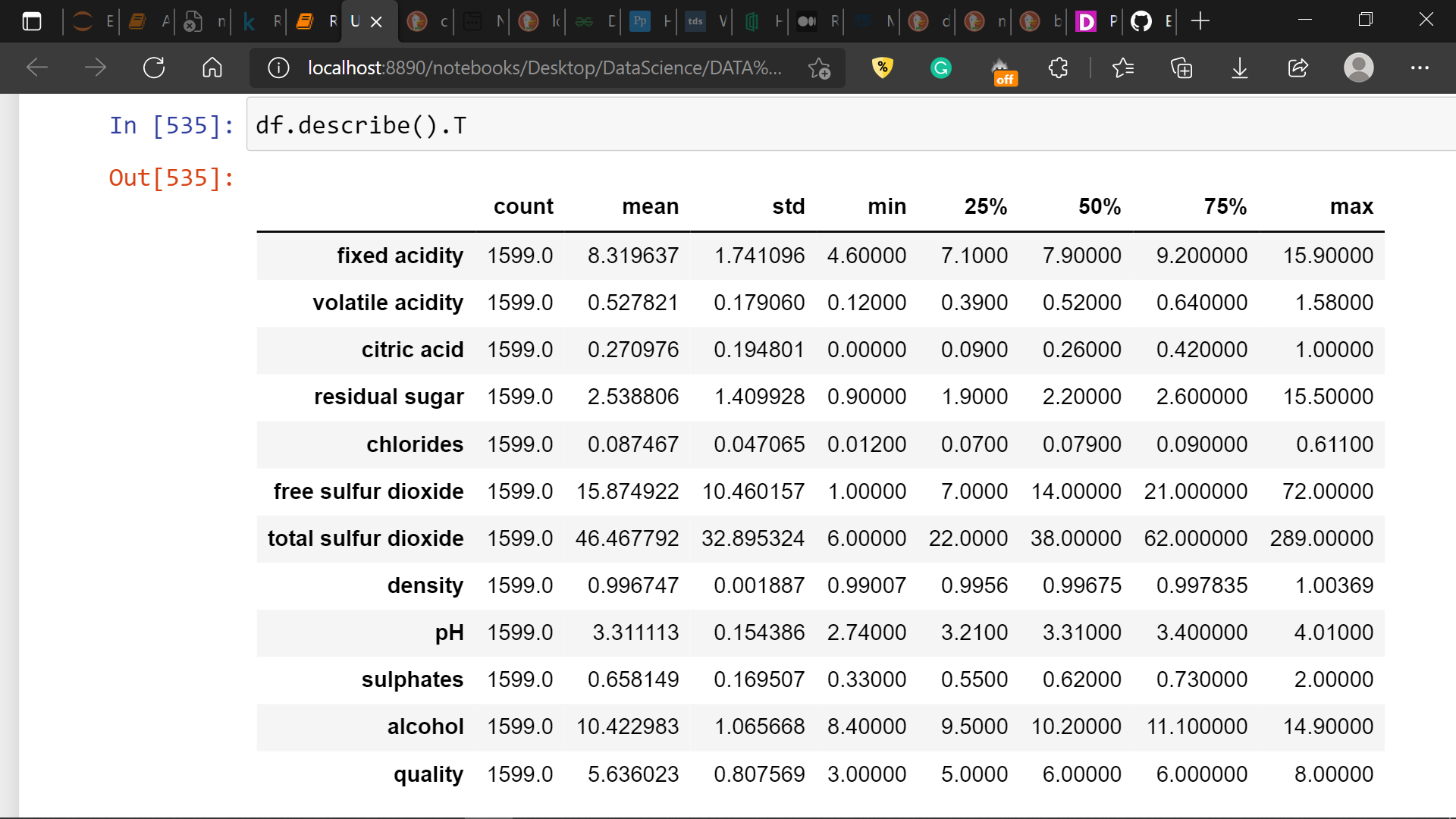
The dataset which is considered here describes various physiochemical attributes of red and white variants of the Portuguese "Vinho Verde" wine along with their quality score which ranges between 0 to 10. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

We will make this problem a binary classification task by using a ‘quality’ threshold of 7. Wines with a quality rating of 7 or better will be considered as ‘Good’ and quality rating less than 7 as ‘Bad’.

**Preliminary data analysis:**

|  |  |
| --- | --- |
|  |  |
|  |

The dataset has total of 1599 observations and 11 input/independent variables. All input variables store continuous numeric data & does not have any missing values. ‘quality’ is our dependent variable and is categorical in nature.

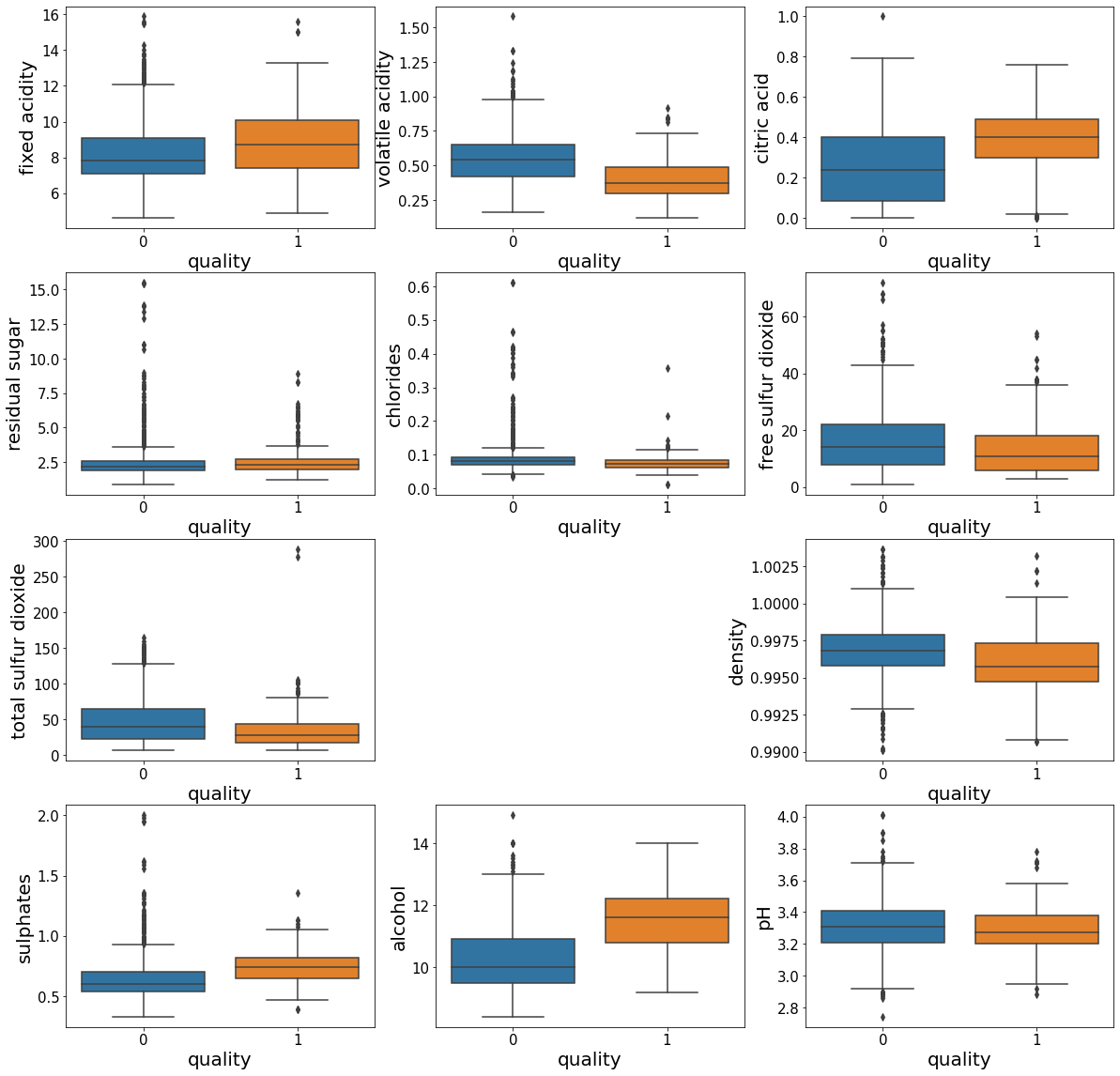


All the input features except ‘volatile acidity’, ‘citric acid’, ‘pH’ and ‘density’ seem to be skewed. This will be confirmed with the further analysis.

**Exploratory Data Analysis:**

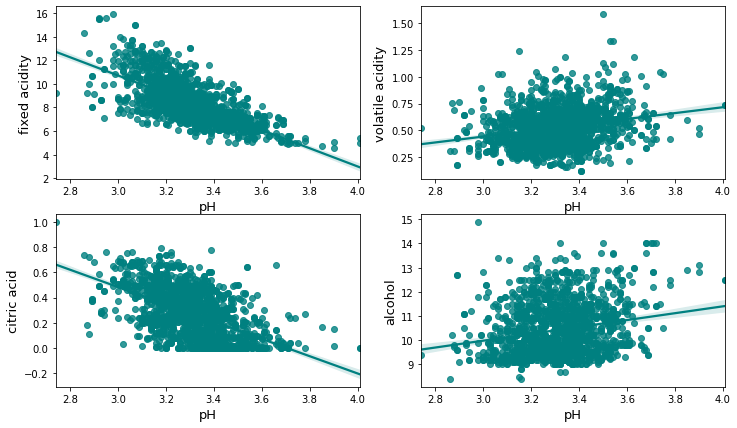
|  |  |
| --- | --- |
|  |  |

As mentioned earlier, here I convert the dependent variable ‘quality’ in a binary variable using 7 as a threshold. After this, ‘quality’ has 217 entries for ‘Good’/(1) class and 1382 entries for ‘Bad’/(0) class. The ‘Good’ class forms only 13.6% of our dependent variable which makes this dataset a imbalanced dataset.



Here I have box plot with ‘quality’ as a hue to check the relationship between independent variables and the depedent variable. Following obsevations can be made from the above boxplots:

* ‘Good’ wines tend to have higher values of ‘fixed acidity’, ‘citric acid’, ‘sulphates’ and ‘alcohol’   
  and on the other hand
* higher values of ‘volatile acidity’, ‘free sulphur dioxide’, ‘total sulphur dioxide’ and ‘density’ are related to ‘Bad’ wine quality.
* ‘residual sugar’, ‘chlorides’ and ‘pH’ seem to have very little to no effect on wine quality.



In above regression plot ‘pH’ shows a clear linear relationship with ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’ and ‘alcohol’. This is also obvious because these physiochemical properties depend on pH values of the given liquid. ‘fixed acidity’ and ‘citric acid’ content show strong negative relationship while the ‘volatile acidity’ and ‘alcohol’ content show moderate positive relationship with the ‘pH’.

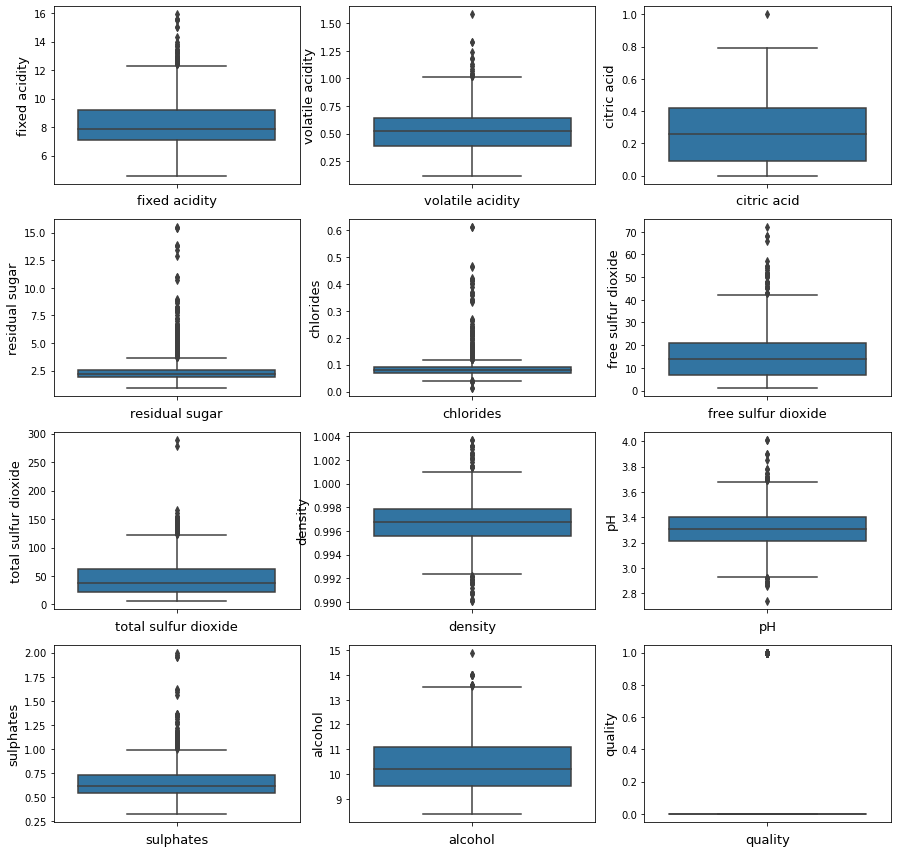
**Checking the skewness in independent variables:**

We will use distribution plots and ‘.skew()’ method to check and quantify the skewness in features.

|  |  |
| --- | --- |
|  |  |

Here, we will consider +/-0.5 as a threshold for skewness. Features which have values beyond +/-5 will be considered as skewed. Using this following observations can be made:

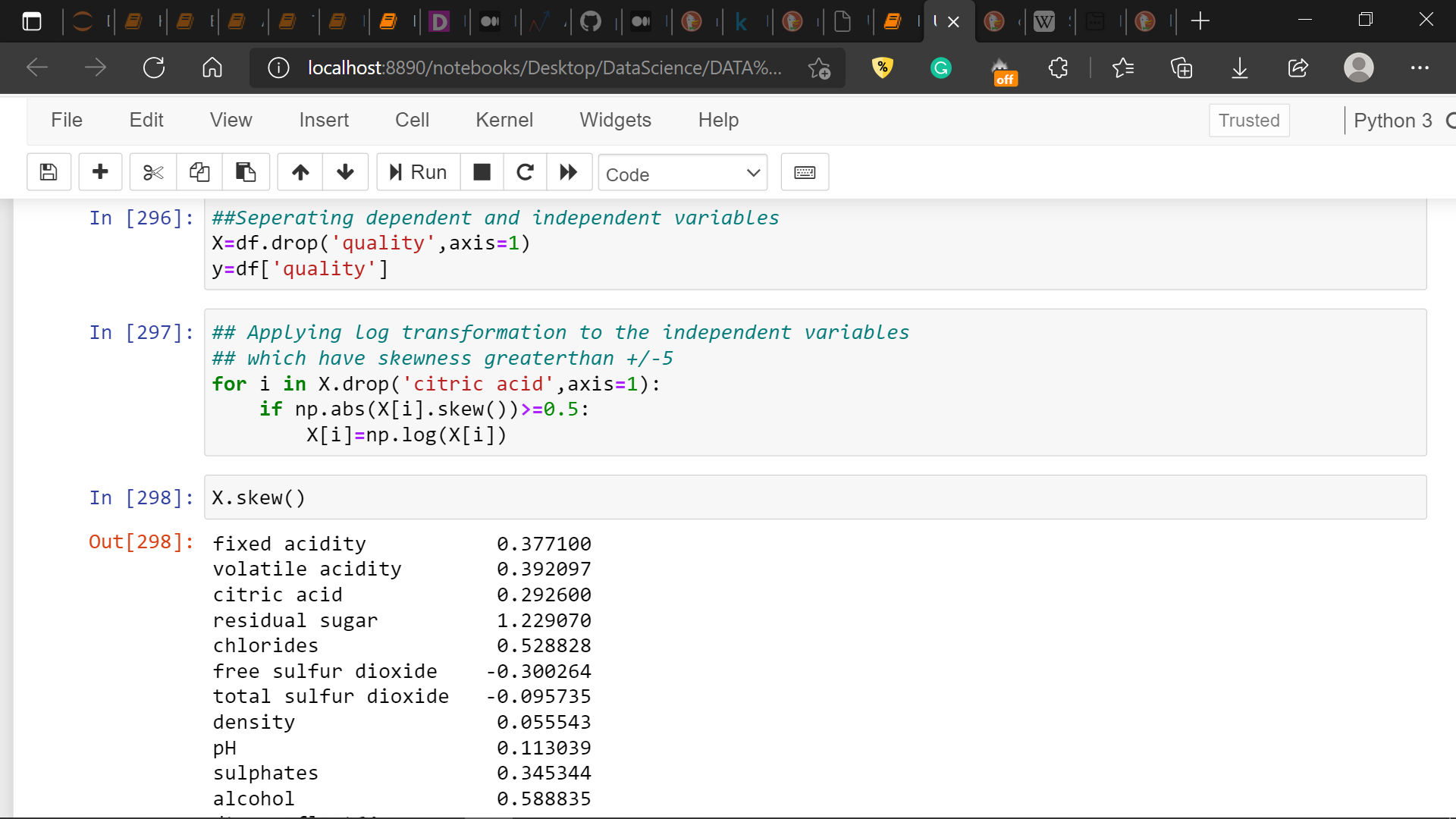
* ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘sulphates’ are highly positively skewed.
* ‘fixed acidity’, ‘volatile acidity’ and ‘alcohol’ have moderate positive skewness.
* ‘citric acid’, ‘density’, ‘pH’ show very low skewness hence can be considered as normally distributed.



All the dependent variables show the presence of outliers. We will use Z-score method with threshold 3 to remove outliers from the dataset.

|  |  |
| --- | --- |
|  |  |

With this, we have dropped 141 observations which means we are losing about 8.81% of the total data. After dropping the outliers using z-score threshold of 3, the skewness in skewed features is reduced but is still beyond the acceptable limit.

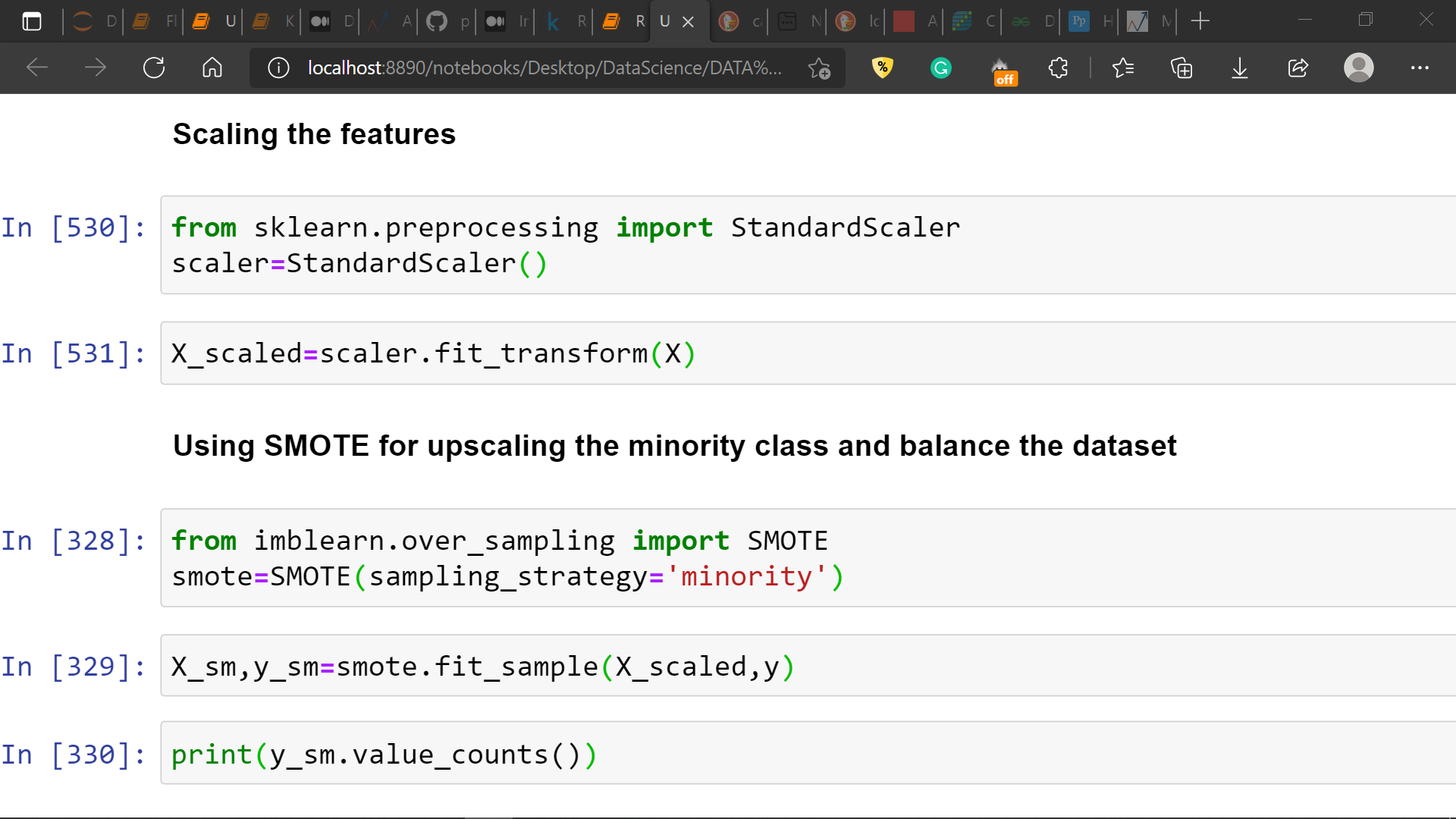


Here, I have separated the independent variables and the dependent variable and have stored them in X and y dataframes respectively. Further after this, I have applied logarithmic transformation to the independent variables which have skewness values beyond +/-5 threshold, in order to further minimize the skewness. We will ignore the skewness which persist in ‘residual sugar’, ‘chlorides’ and ‘alcohol’ after the above treatment.

**Feature Selection using SelectKBest with ANOVA f-values:**

|  |  |
| --- | --- |
|  |  |

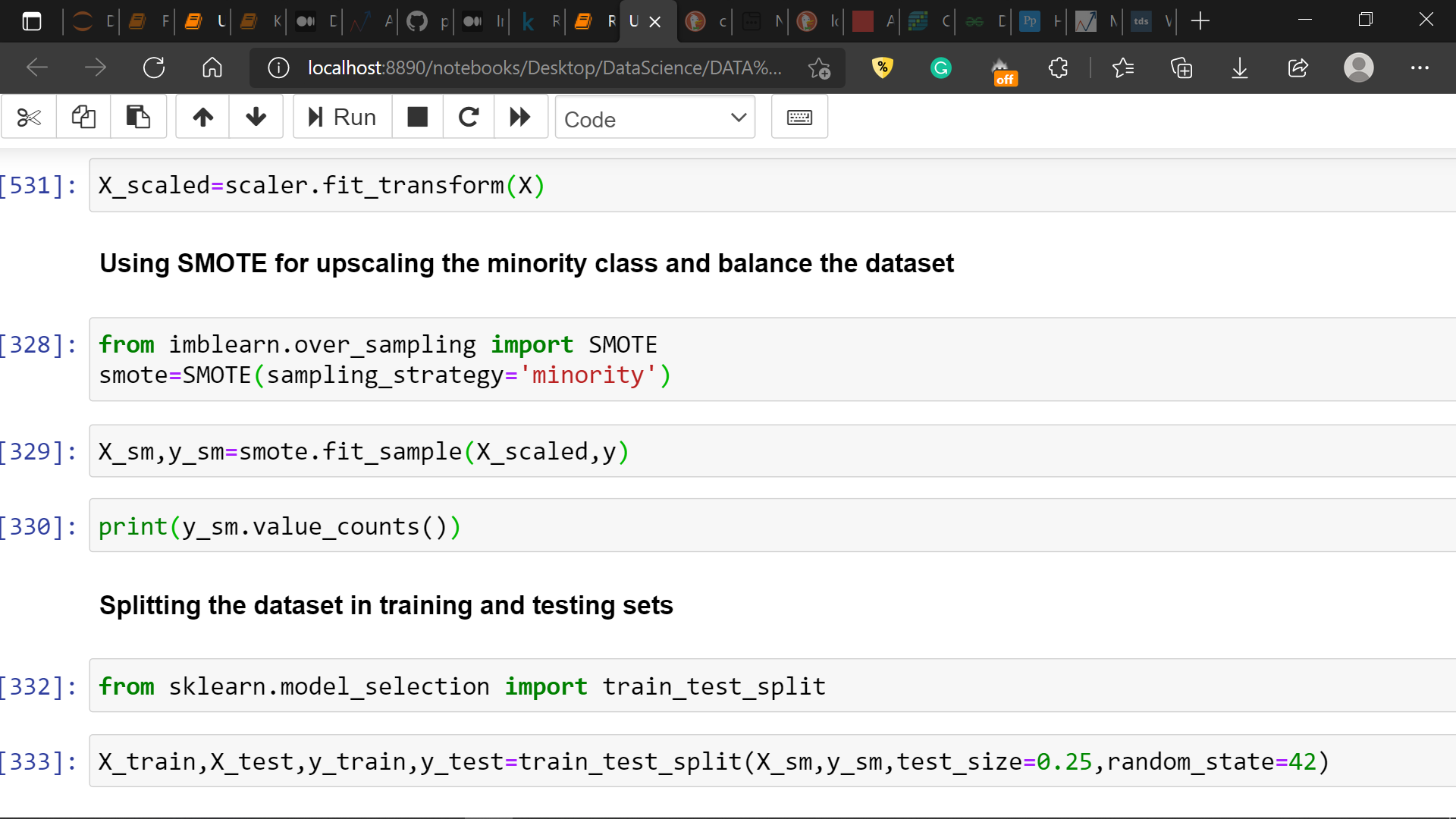
Here, I have used SelectKBest and f\_classif from sklearn’s ‘feature\_selection’ liabrary. ‘f\_classify’ score function uses ANOVA f-values to quantify the relation between continuous features and catagorical dependent variable. As we can see in the above table, ‘pH’, ‘free sulfur dioxide’, ‘residual sugar’ have very low score. As we have very limited features in this dataset I will not drop any features here to retain whatever little information they posses.



Here, I have done feature scaling using ‘StandardScaler’ from Sklearn’s preprocesing liabrary. Usually in real word datasets features have very diferent magnitudes, range and units than each other. Hence, we perform feature scaling on such datasets so that Machine learning algorithms can interpret these features on same scale.

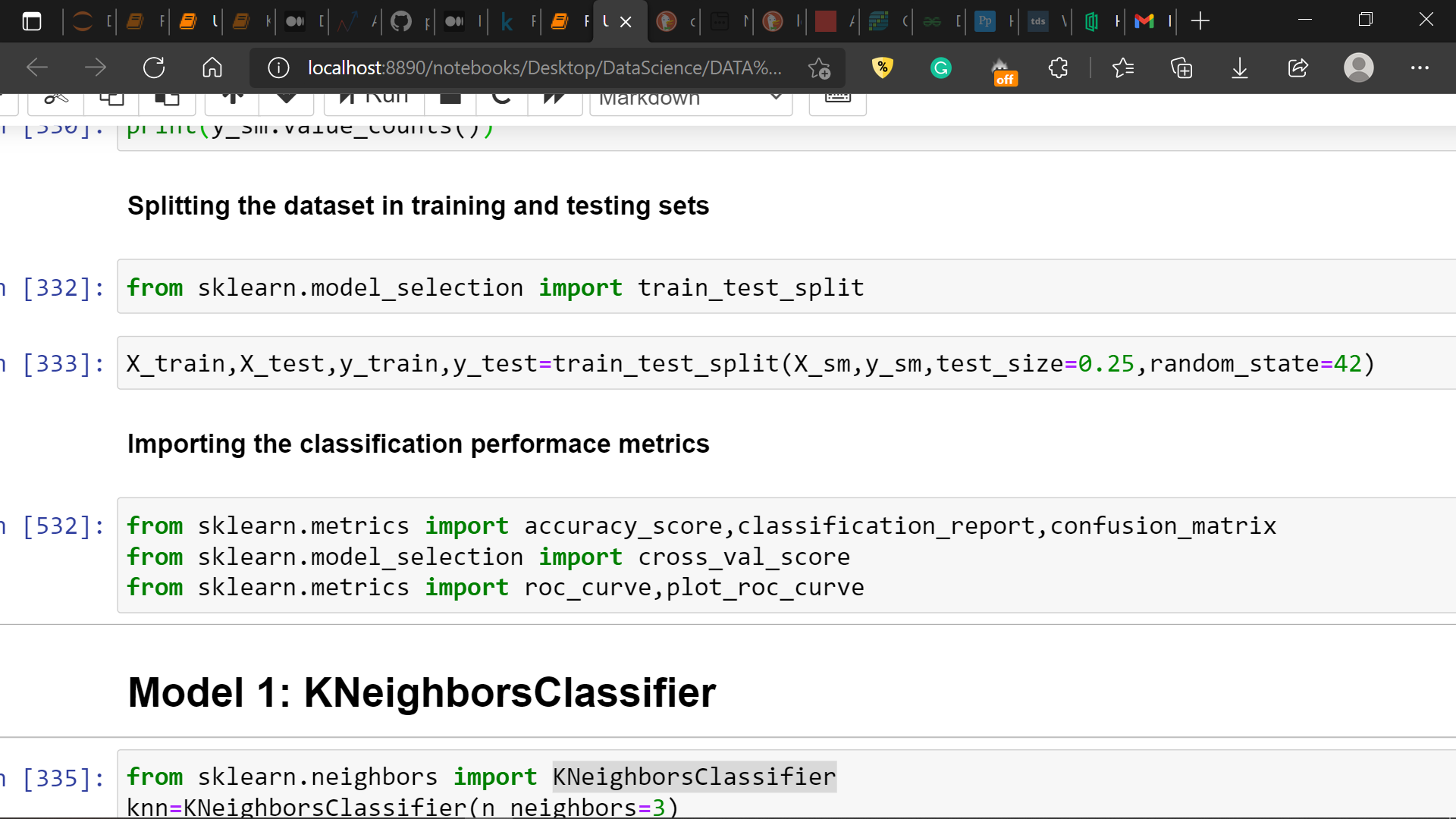
As seen during the preliminary analysis, this dataset is ‘Imbalanced’ i.e. the number of observations for class ‘0’ are way more in number than the class ‘1’. This means that the dataset is biased towards the class ‘0’.

If the dataset is biased towards one class, an algorithm trained on the same data will be biased towards that particular class. In such cases, the machine learning algorithm/ model will learn more from the biased examples as opposed to the examples in the minority class. Which may lead us to a scenario where the algorithm/model assumes that any data you feed it belongs to the majority class. This scenario makes a model seem naïve in its predictions which mean that even if such model achieves high accuracy score it may not be able to predict the minority class properly. To avoid such issues, we use dataset balancing techniques like oversampling, undersampling etc.



Here, I have used SMOTE to ‘oversample’ the minority class in this dataset which is ‘1’ (Good quality) in order to balance the frequency of both the classes in our dependent variable. Further to this I have split the dataset in to training and testing sets of features and the dependent variable.

Post this we enter the modelling stage where we will apply machine learning algorithms/models to the prepared data and evaluate them based on various performance metrics.

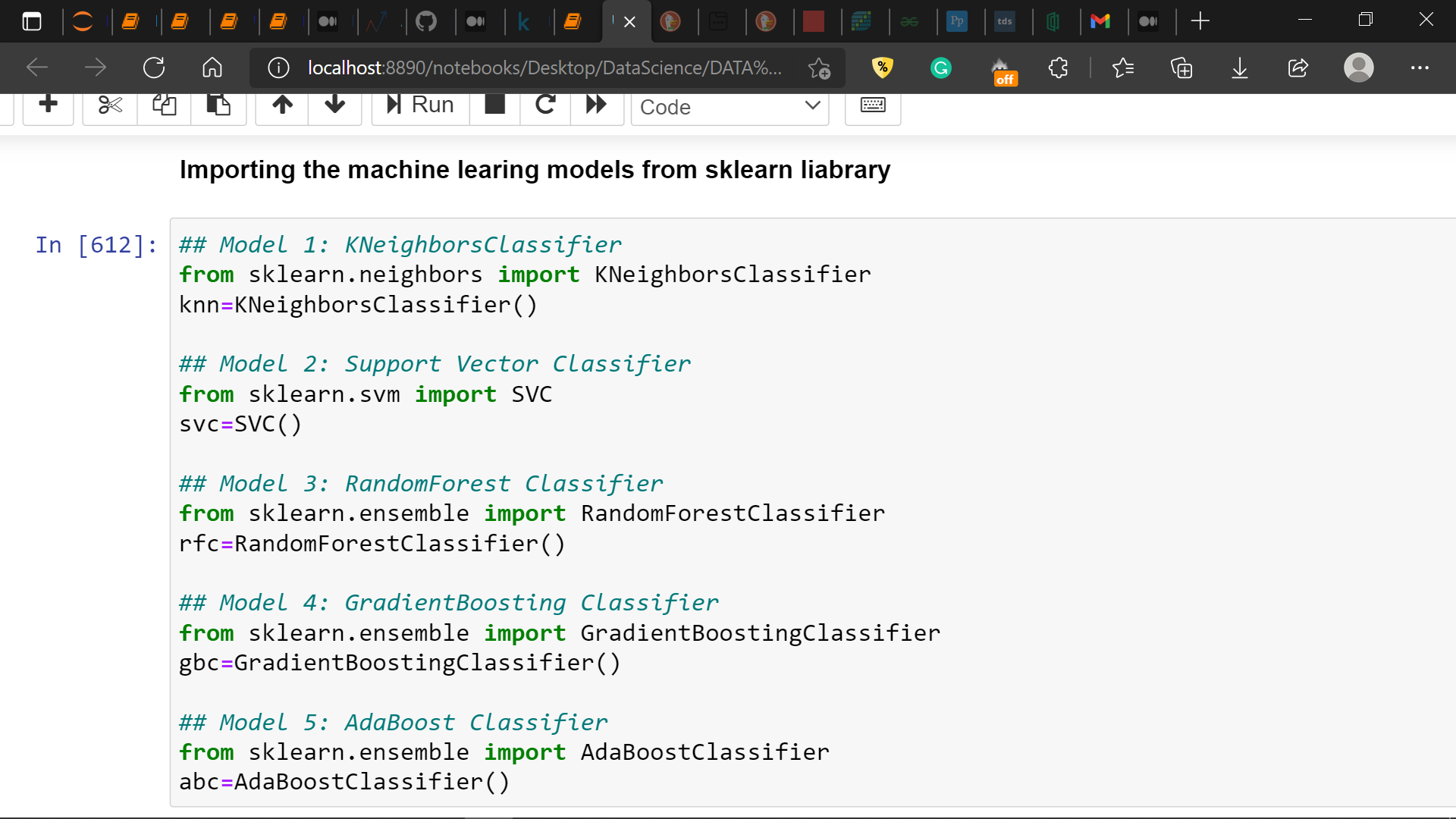


In this case I will be using, accuracy score, cross validation score, f1\_score, confusion matrix, classification report and the roc\_auc curve to check the model performance.

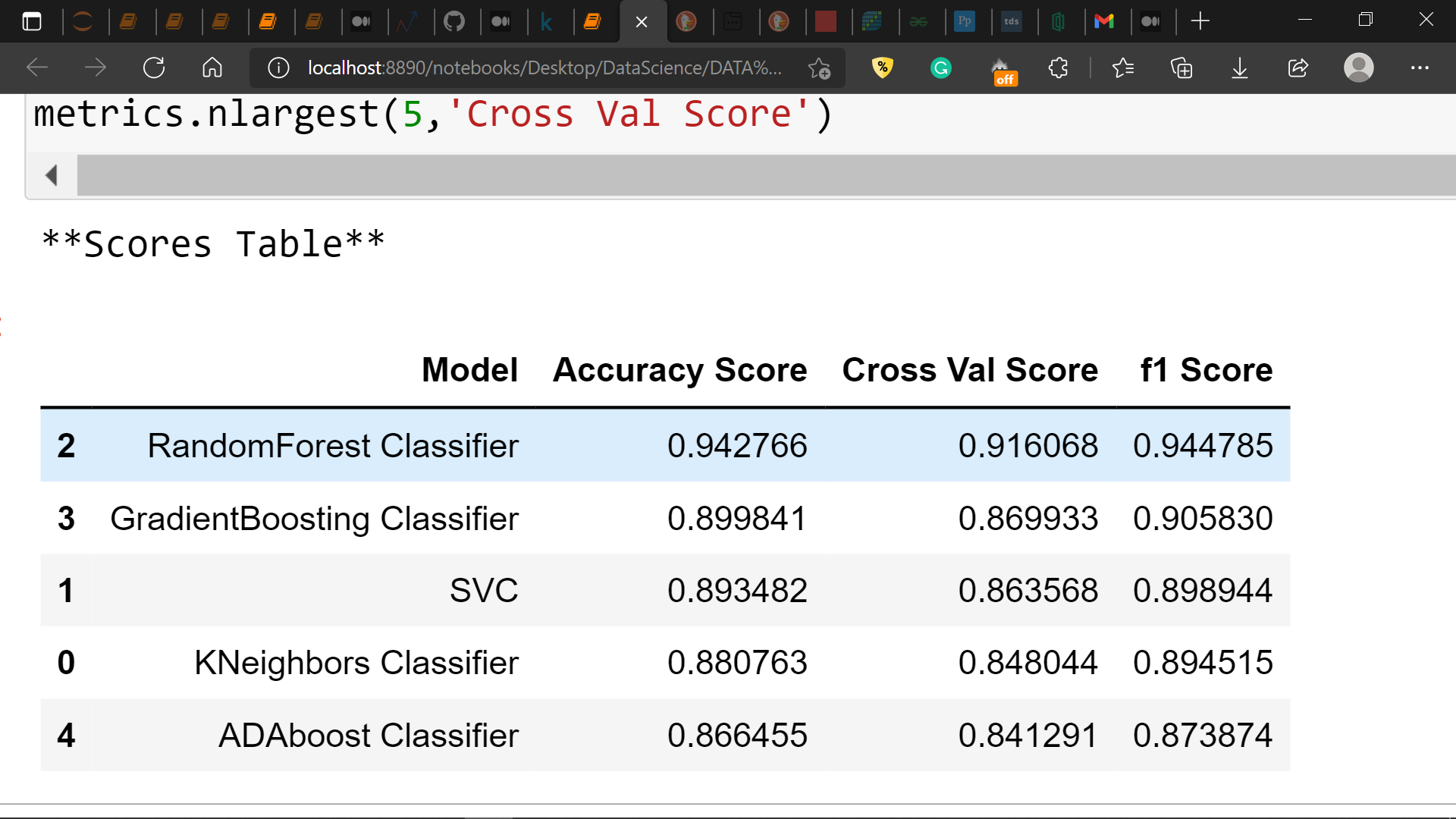
**Applying Machine Learning models:**

For this dataset I will be using KNeighbors, Support Vector, Random Forest, Gradient Boosting and Ada Boost classifiers.

Based on the performance metrics comparison of these models, the model that performs best for this dataset will be selected.



**Following table stores the values of ‘Accuracy Score’, ‘Cross Validation Score’ and ‘f1 Score’ for all the applied machine learning models.**



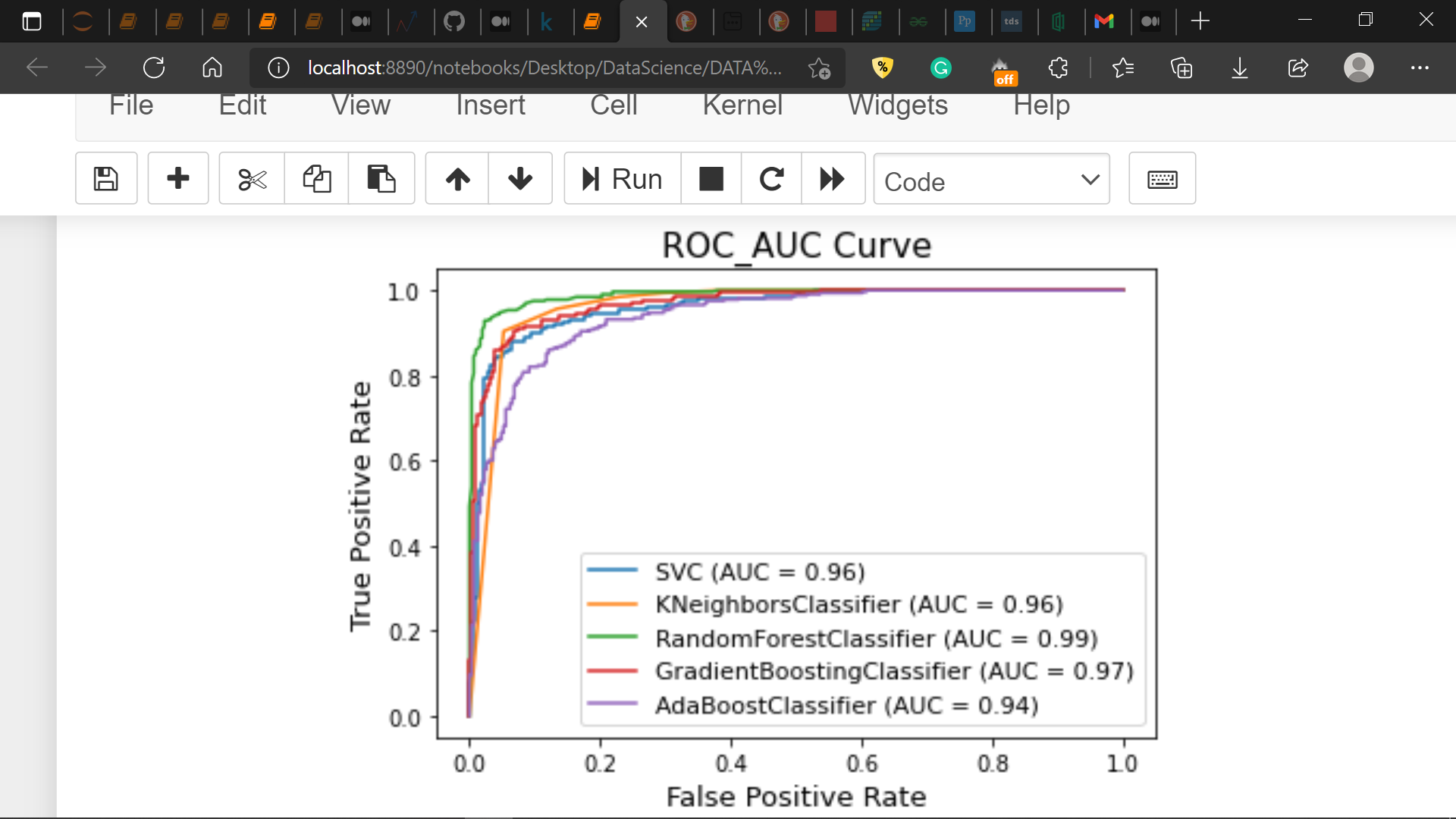
**The above scores table clearly shows that the ‘Random Forest Classifier’ model outperforms all the other models for the given dataset by achieving accuracy score of 94%, Cross validation Score of 92% and f1 score of 94%.**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

The above table stores confusion matrices for the all five machine learning models which are used here and shows that the Random forest classifier does a better job in correctly predicting both the classes.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Also the above classification reports for all models show that Random Forest Classifier achieves eaqually good ‘precision’ and ‘recall’ scores for both the classes where as the other models fails to do so.



Here, I have plotted the roc\_auc curve for all the machine learning models and again the Random Forest Classifier best AUC (Area Under the Curve) score of 0.99.

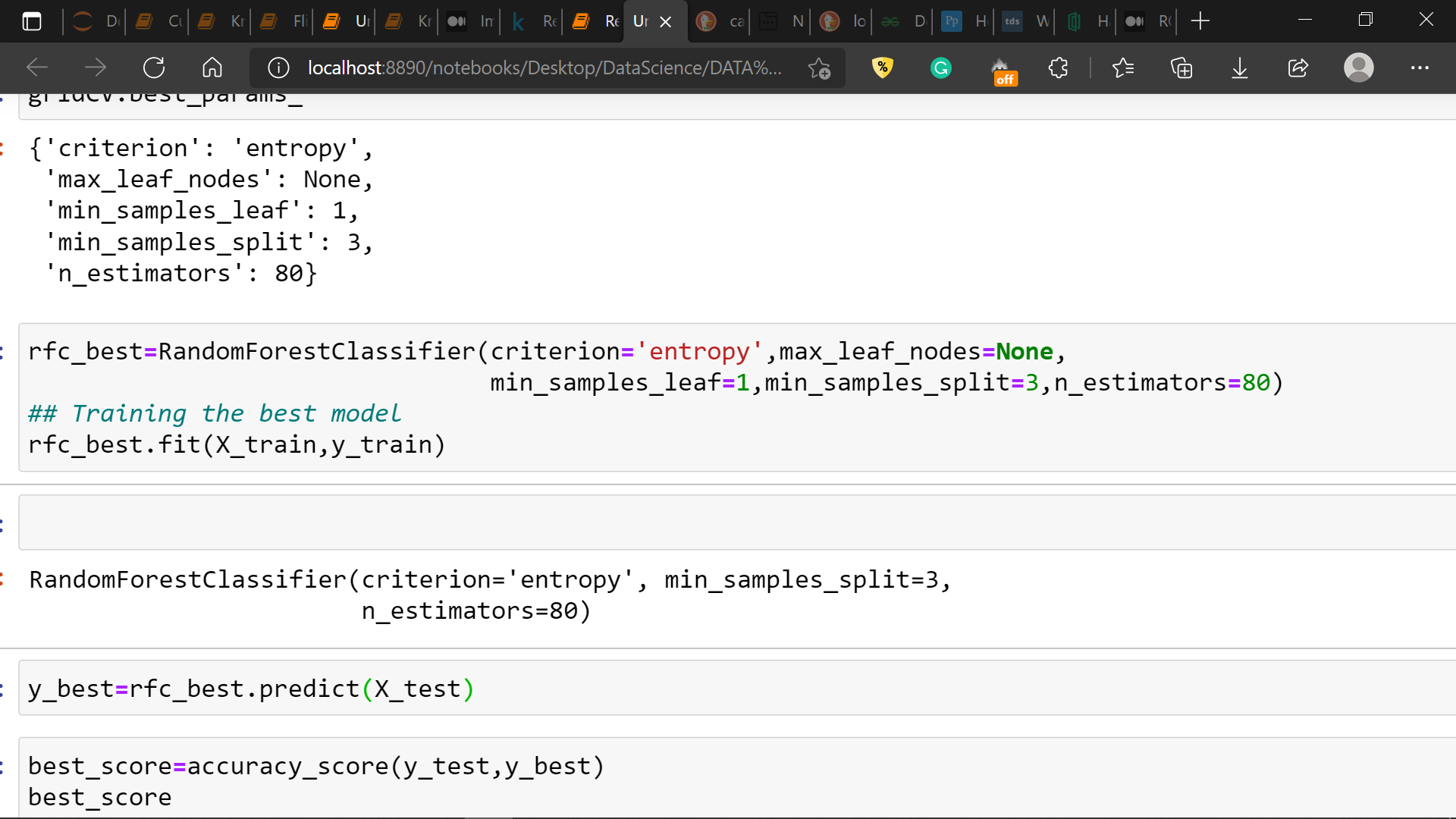
**Selecting the best performing model:**

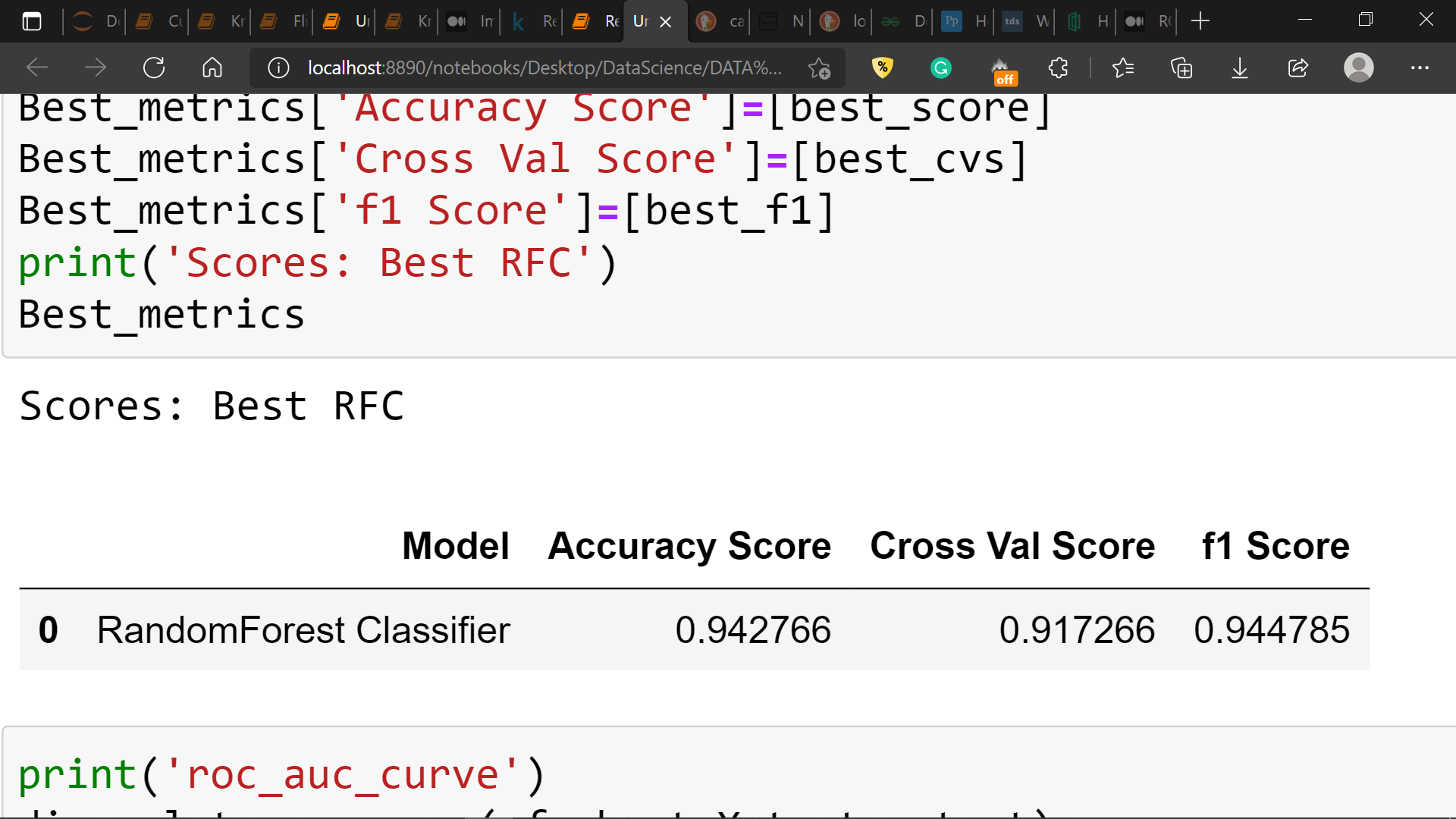
Based on all the above metric comparisons, Random Forest Classifier comes out as our best performing model by outperforming other models in all the performance metrics.

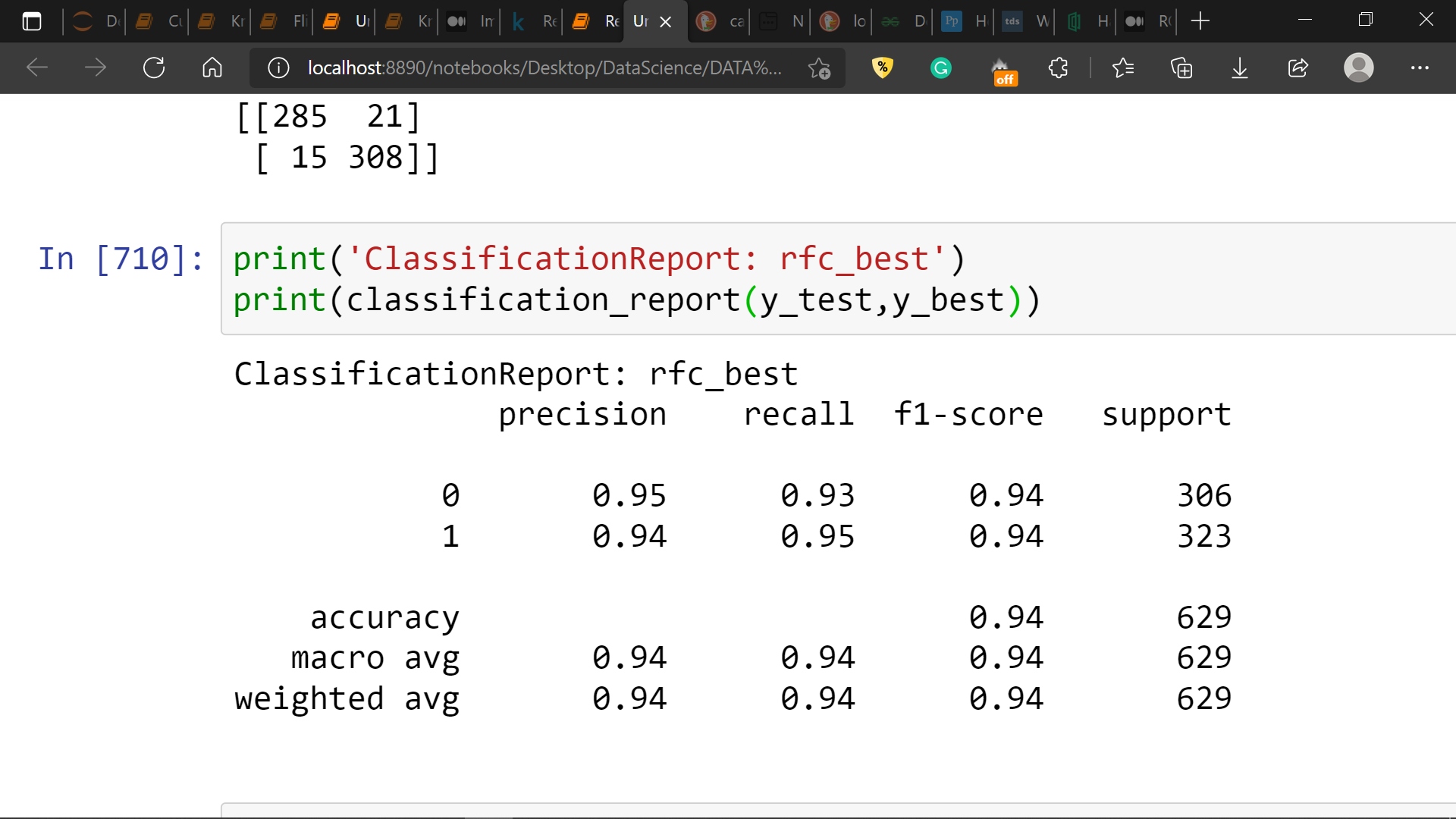
**Tuning Hyper parameters of the best performing model using Grid Search Cross Validation:**



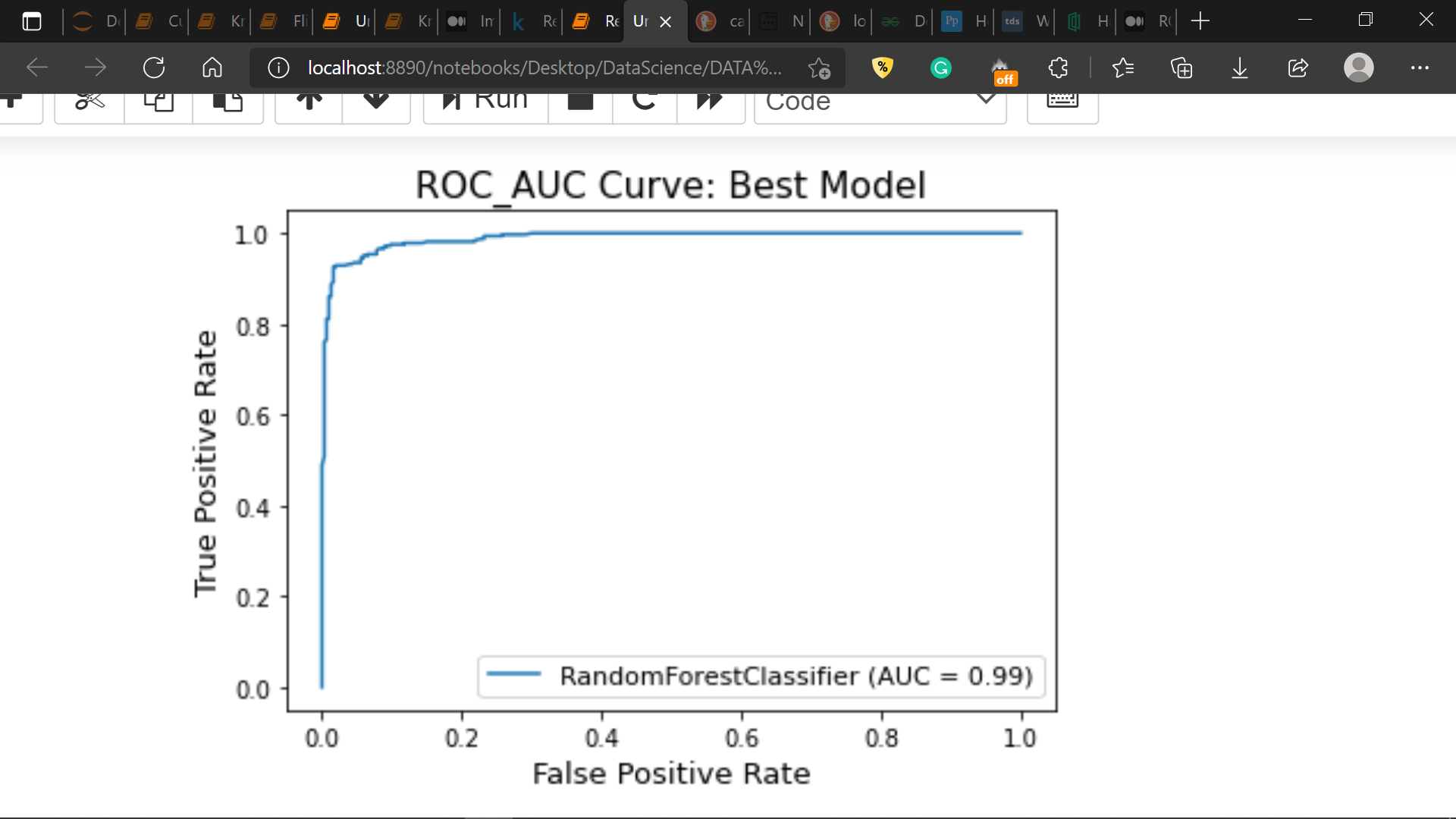
**Training the best model with best hyperparameters obtained above:**







The Random Forest Classifier when trained with the best hyperparameters as obtained above achives ‘Accuracy Score’ of 94.3%, ‘Cross Validation Score’ of 91.7% and ‘f1 Score’ of 94.5%. Also, it achives (precision,recall) of (0.95,0.93) for class ‘0’ and (0.94,0.95) for class ‘1’



**Above image shows the roc\_auc curve for the hyperparameter-tuned Random Forest Classifier model.**

**Conclusion:**

**In this case study, we used supervised machine learning to predict the quality of wine. Random Forest Classifier turned out to be the best performing algorithm/model for this dataset, achieving the best accuracy score of 94.275. The analysis suggested that the physiochemical parameters ‘Alcohol’, ‘Sulphates’ and ‘Volatile acidity’ affect the wine quality most. Higher values of ‘Alcohol’ and ‘Sulphates’ can be associated with ‘Good’ quality wine and on the other hand higher values of ‘volatile acidity’ are observed with the ‘Bad’ quality wines.**