SUPERVISED MACHINE LEARNING: CLASSIFICATION

COURSE PROJECT REPORT

Machine learning is connected with the field of education related to algorithms which continuously keeps on learning from various examples and then applying them to real-world problems.  Classification is a task of Machine Learning which assigns a label value to a specific class and then can identify a particular type to be of one kind or another. The most basic example can be of the mail spam filtration system where one can classify a mail as either “spam” or “not spam”. You will encounter multiple types of classification challenges and there exist some specific approaches for the type of model that might be used for each challenge.

ABOUT THE DATA

Available on the UCI machine learning repository (<https://archive.ics.uci.edu/ml/datasets/wine+quality>). The red wine samples were obtained from the north of Portugal to model red wine quality based on physicochemical tests. The dataset contains a total of 12 variables, which were recorded for 1,599 observations. The datasets are also available from <http://www3.dsi.uminho.pt/pcortez/wine/>

1. Alcohol: the amount of alcohol in wine
2. Volatile acidity: acetic acid content which leading to an unpleasant vinegar taste
3. Sulphates: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
4. Citric Acid: acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)
5. Total Sulfur Dioxide: is the amount of SO2
6. Density: sweeter wines have a higher density
7. Chlorides: the amount of salt
8. Fixed acidity: are non-volatile acids that do not evaporate easily
9. pH: the level of acidity
10. Free Sulfur Dioxide: it prevents microbial growth and the oxidation of wine
11. Residual sugar: is the amount of sugar remaining after fermentation stops. (Wines > 45g/ltrs are sweet)

OBJECTIVES FROM THE ANALYSIS

To classify the color of wine as red or white

1. Train test split

2. Simple EDA

* Descriptive statistics and data cleaning
* Numerical features
* Categorical features

3. Model variations

* Apply One-hot encoding

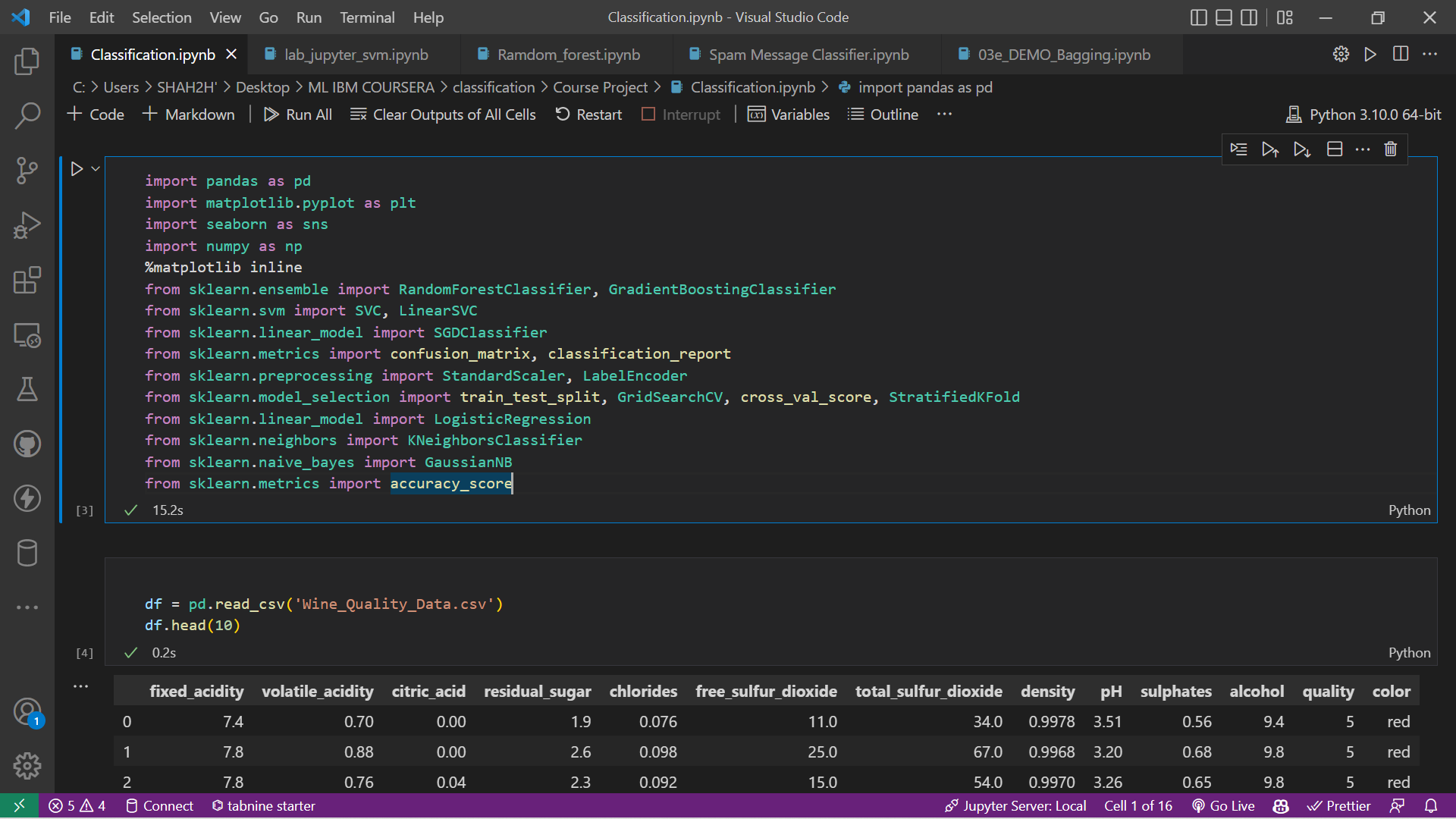
4. Applying classification Algorithms

* Logistic Regression
* KNN Classifier
* Svc Classifier
* SVM kernel Classifier
* Gaussian Nave Bayes
* Decision Tree
* Random Forest
* Compare the metrics

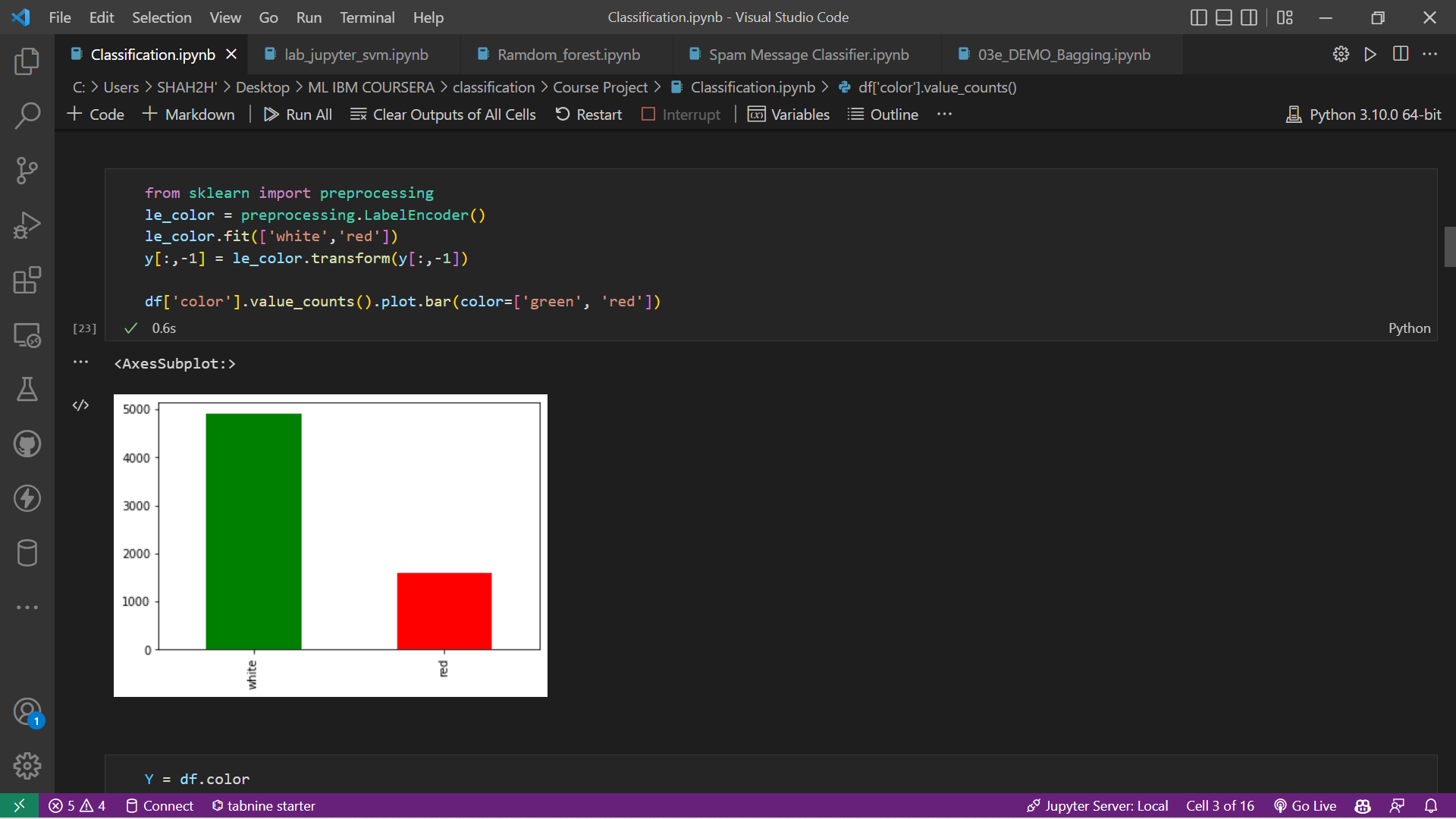
5. Conclusion

METHODS USED

1. Import all the required libraries and data

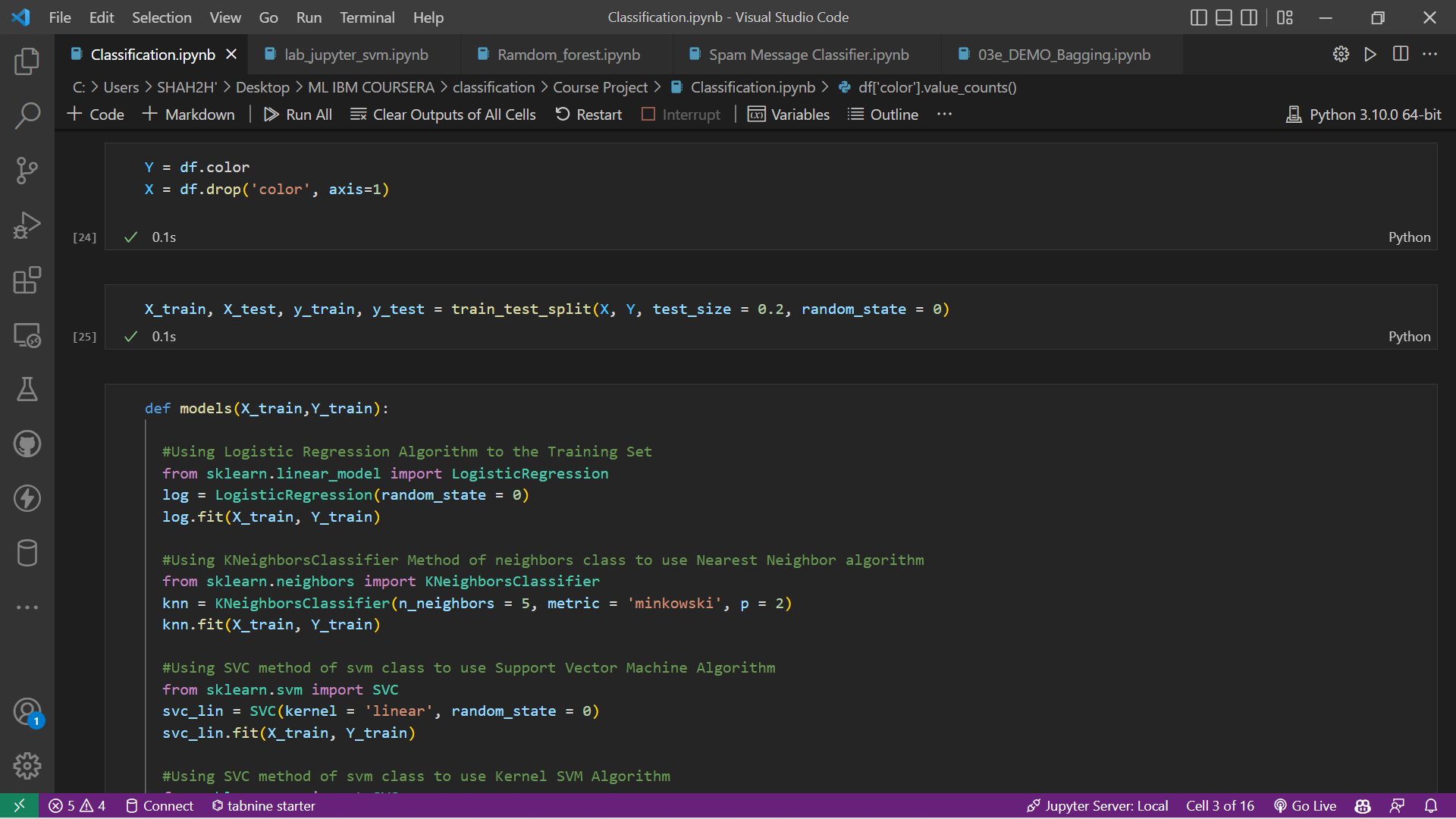


1. Find the value counts and the name of columns.
2. We then apply label encoder on the color column and plot a graph

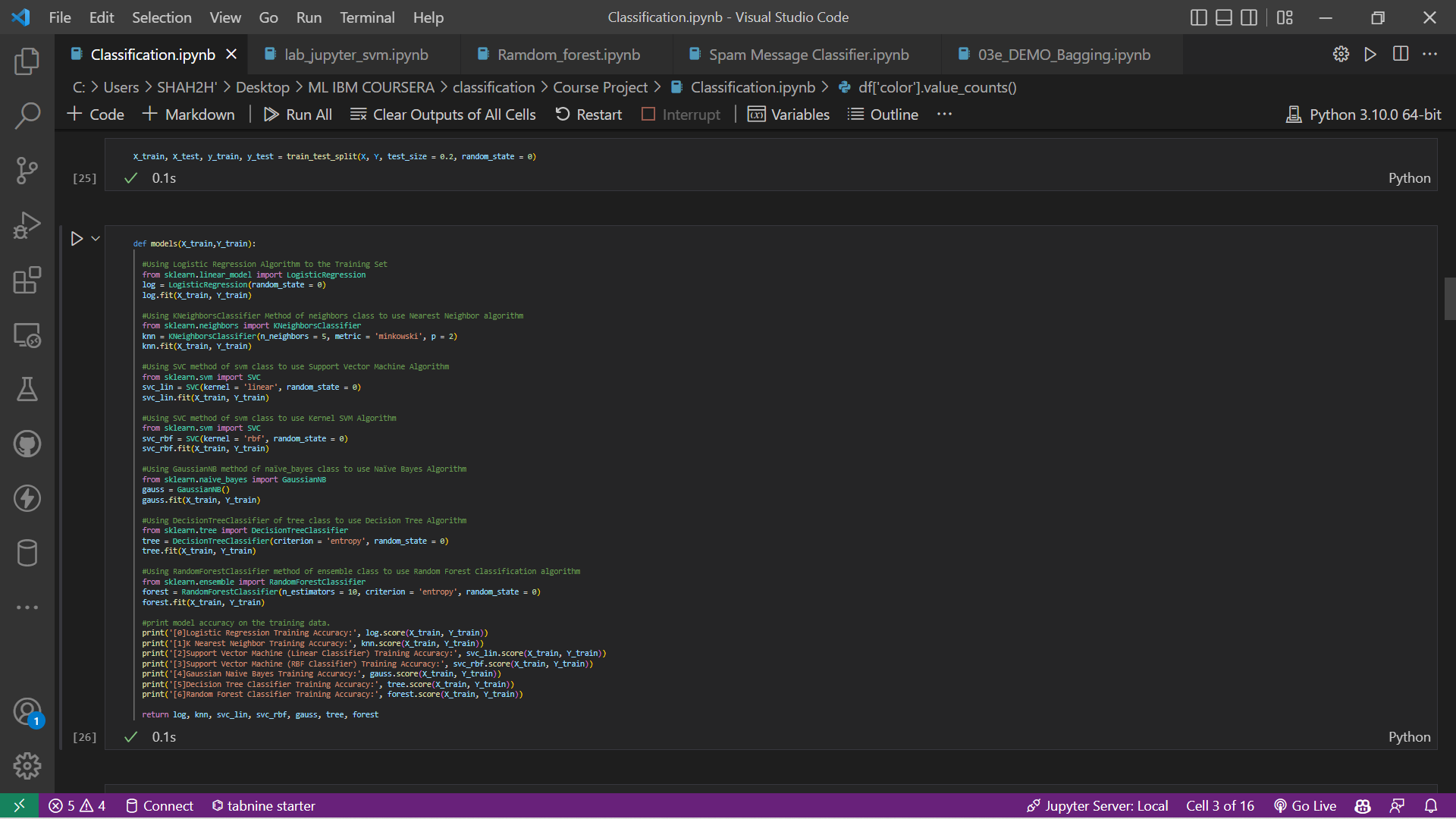


We can see that the data is not highly skewed so we go ahead with train test split

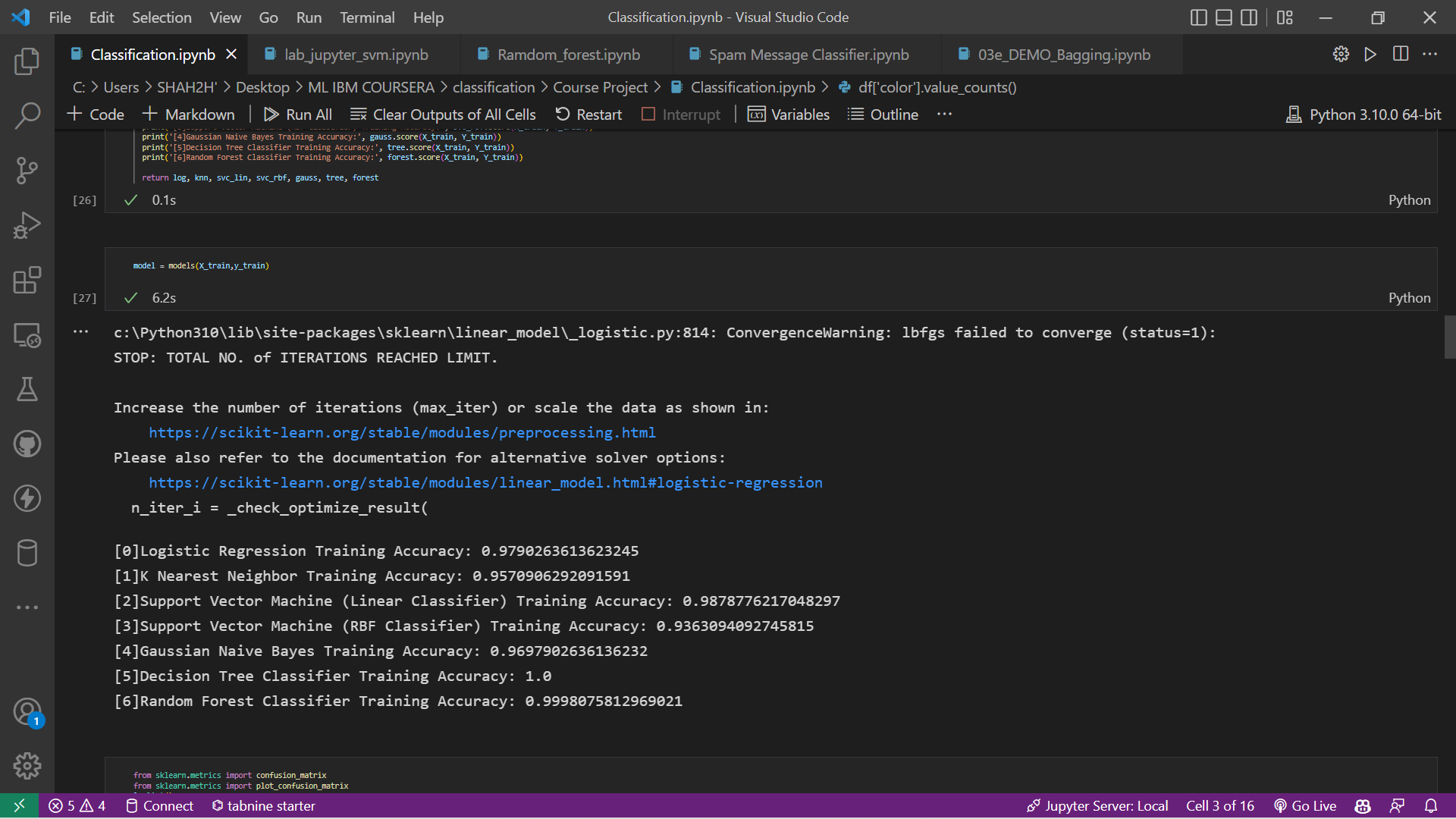
1. Now we apply train test split to the data



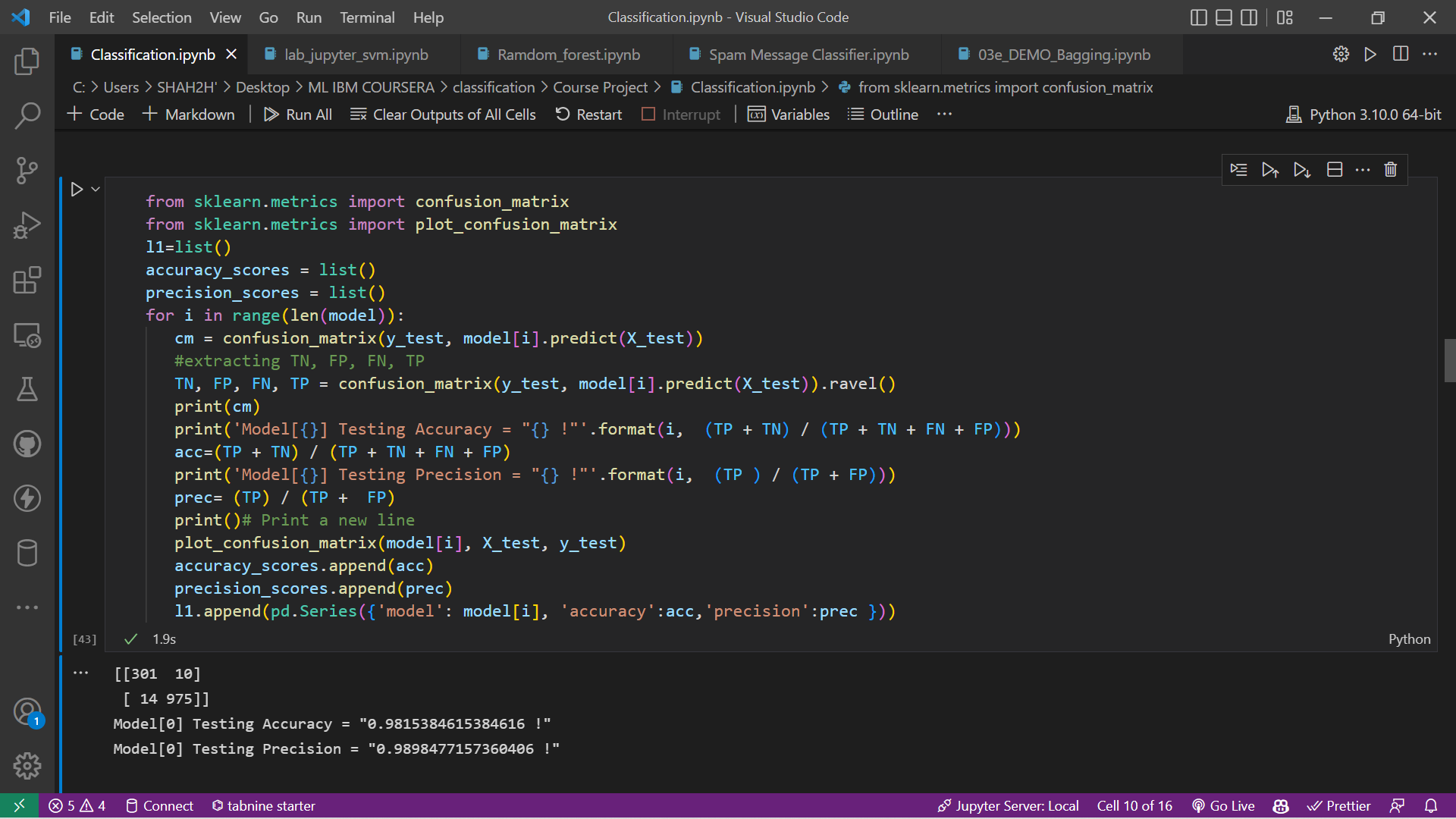
1. Now we apply all algorithm models



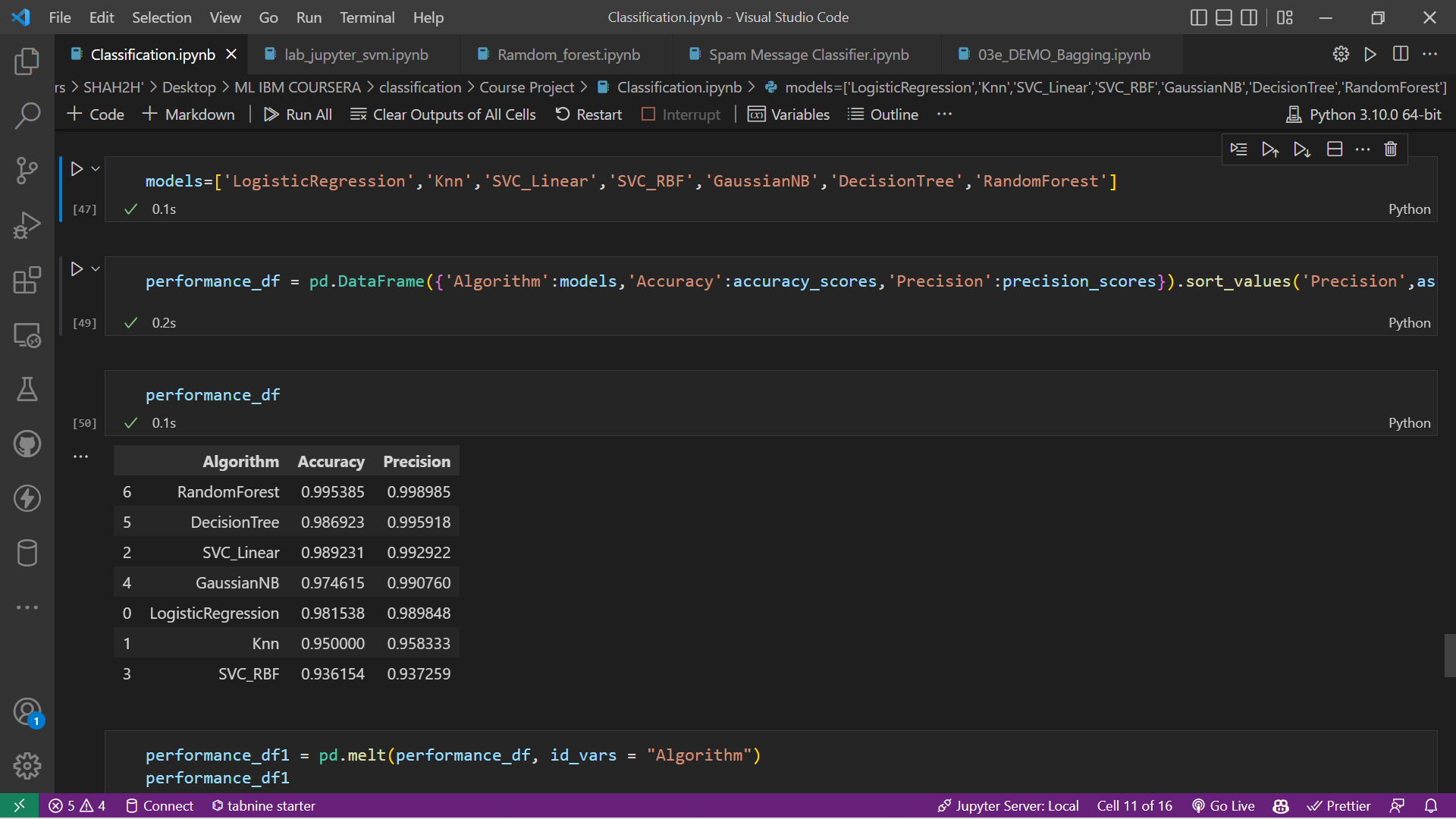
1. Then we print the training accuracy of all the classifiers



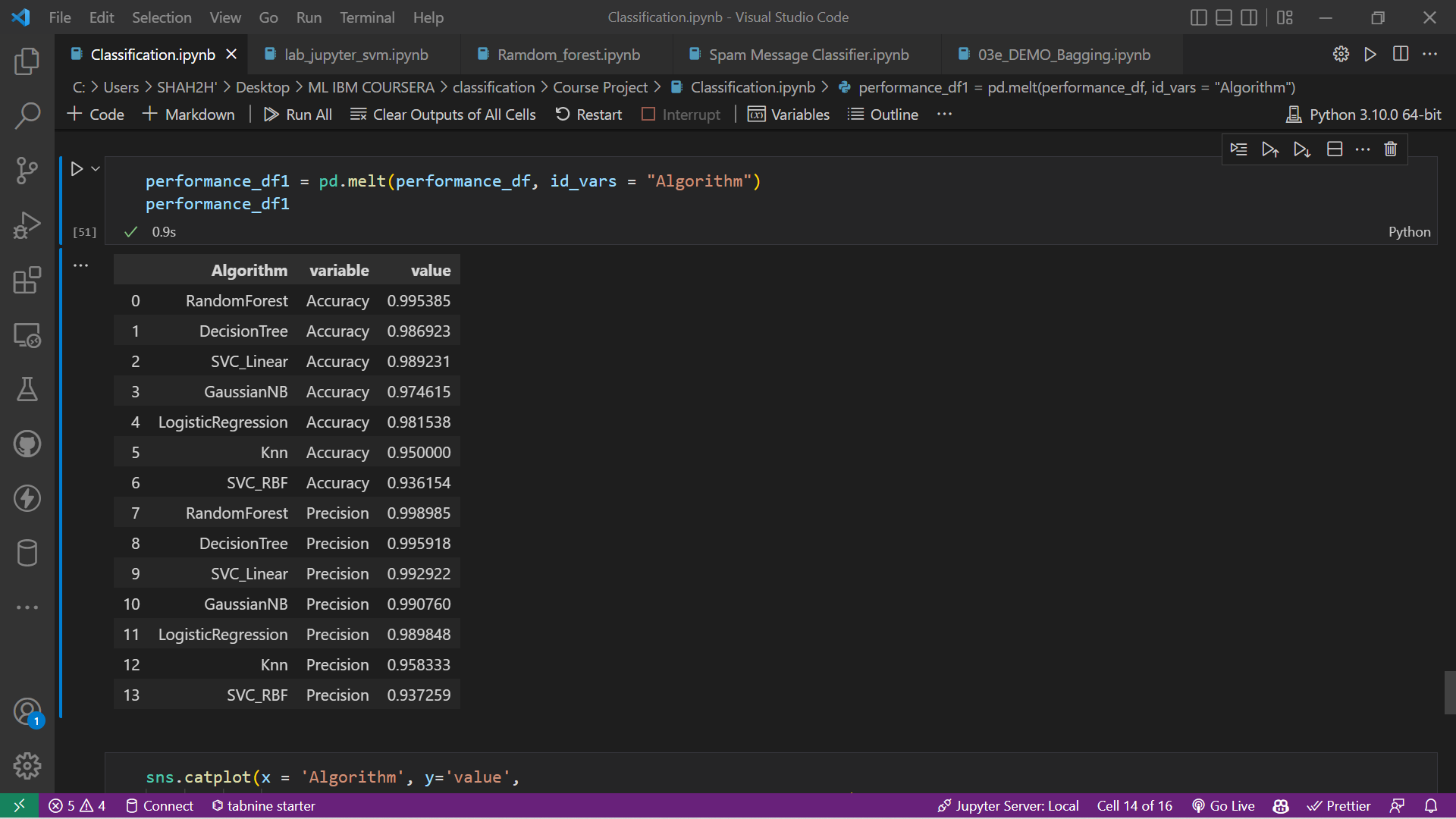
1. Now we print the confusion matrix and accuracy and precision for all the classifiers



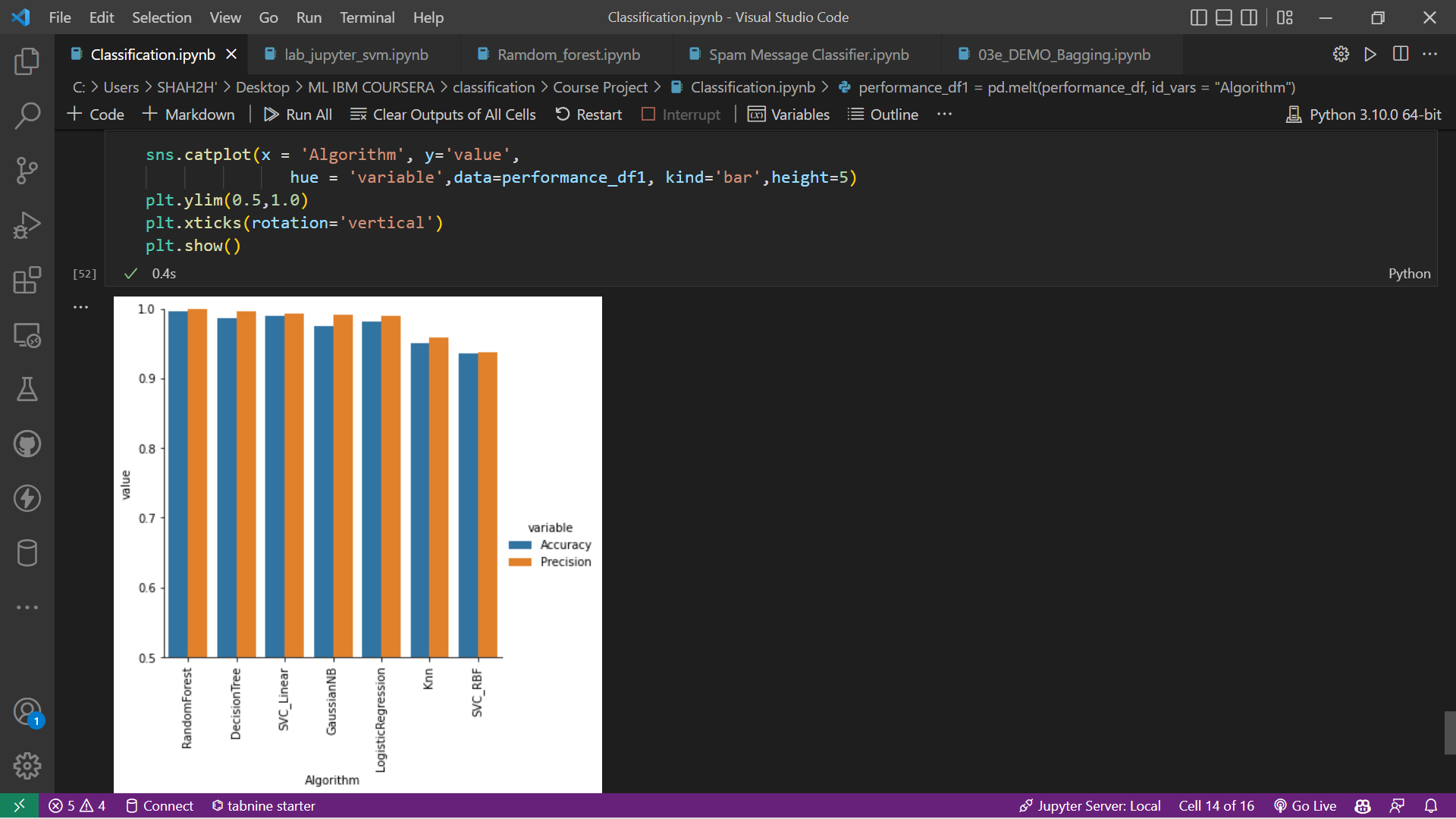
1. Now we extract all this information in table format



1. For plotting the graph we reformat the table as follows:



1. Now we finally print the comparison graph



CONCLUSION

We can now compare the metrics and see that the accuracy of our model is very good and all classifiers perform a great task at it.