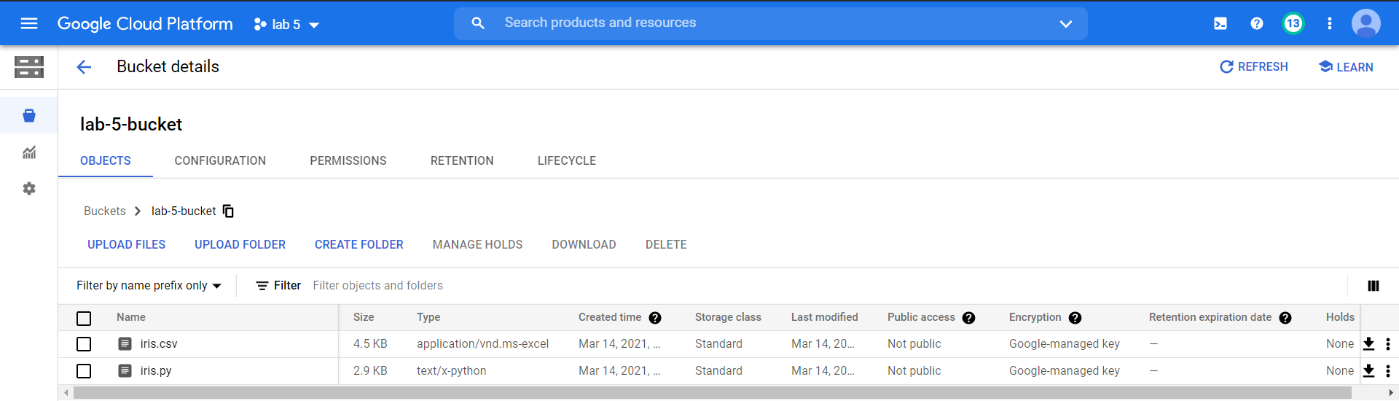
1. The dataset has 5 columns. 4 of them are features and last one is the class which the flower belongs to.

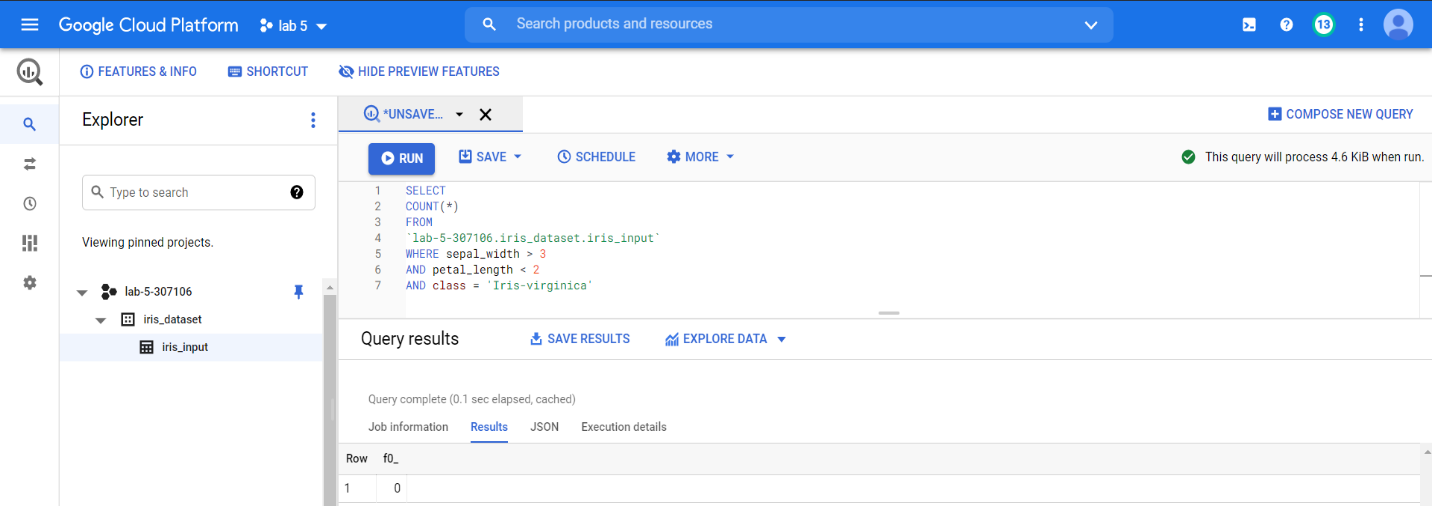
The 4 features are: sepal length, sepal width, petal length, petal width (all are measured in cm)

The output feature is the class which the flower belongs to: Iris-setosa, Iris-versicolor, Iris-virginica.

After downloading the data from the website, It is converted to csv format and is uploaded to the google cloud storage.

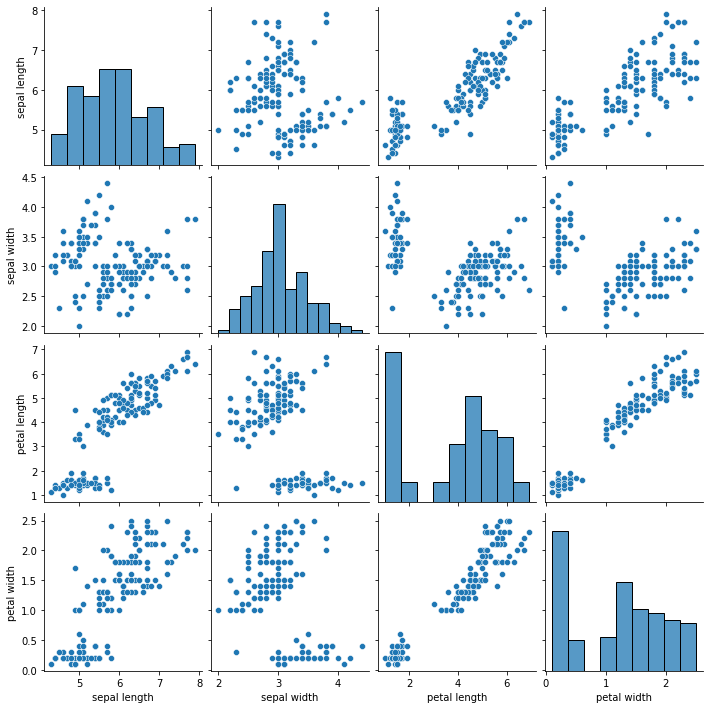






The count for the given question = 0

1. The given data is very simple and has only 4 features. The pair-plot for the given data is given below:



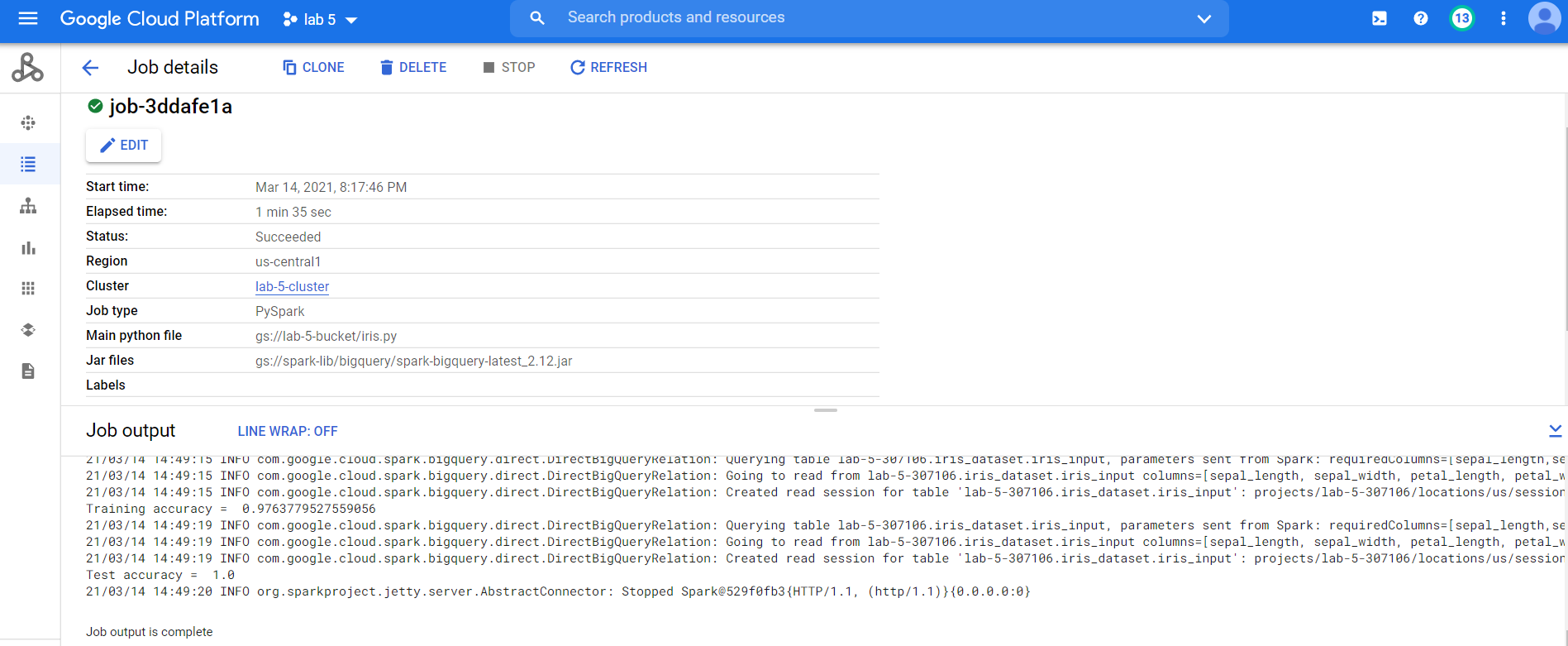
The dataset is very simple and not very completed data pre-processing techniques are required for obtaining good accuracy. While training some models Min-Max scaling was used. To model the input vs output 4 well known models are used. They are:

Random Forest Classifier, Logistic Regression, Decision Tree Classifier, One-vs-Rest Classifier.

Except for random forest for remaining models the input features are scaled.

The input data is split into **80:20** ratio. The various hyperparameters used for various models can be inferred from the code.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Random Forest** | **Logistic Regression** | **Decision Tree** | **One-vs-Rest** |
| **Train Accuracy** | 0.9921259842519685 | 0.8503937007874016 | 1.0 | 0.9763779527559056 |
| **Test Accuracy** | 1.0 | 0.9130434782608695 | 1.0 | 1.0 |



Note: The whole output log of the Dataproc job is attached in a .txt file. The code is also attached with the submission.