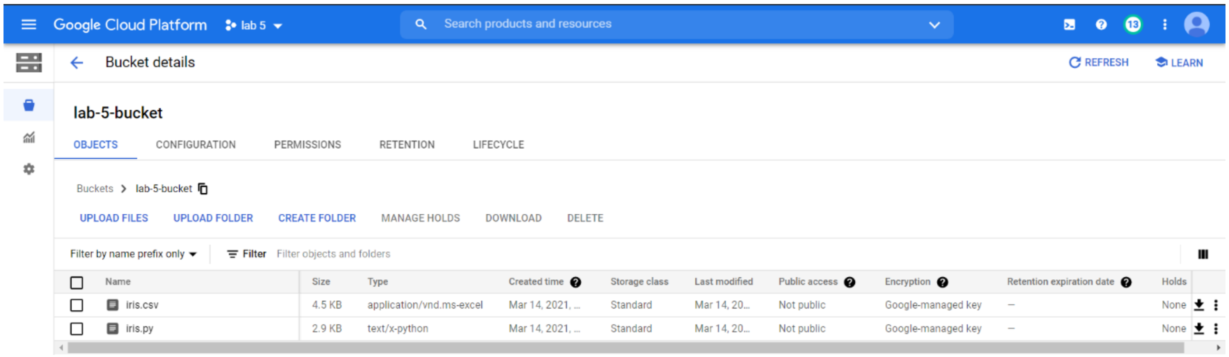
1) The dataset has 5 columns. 4 of them are features and last one is the class which the flower belongs to. The 4 features are: sepal length, sepal width, petal length, petal width (all are measured in cm)

   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.

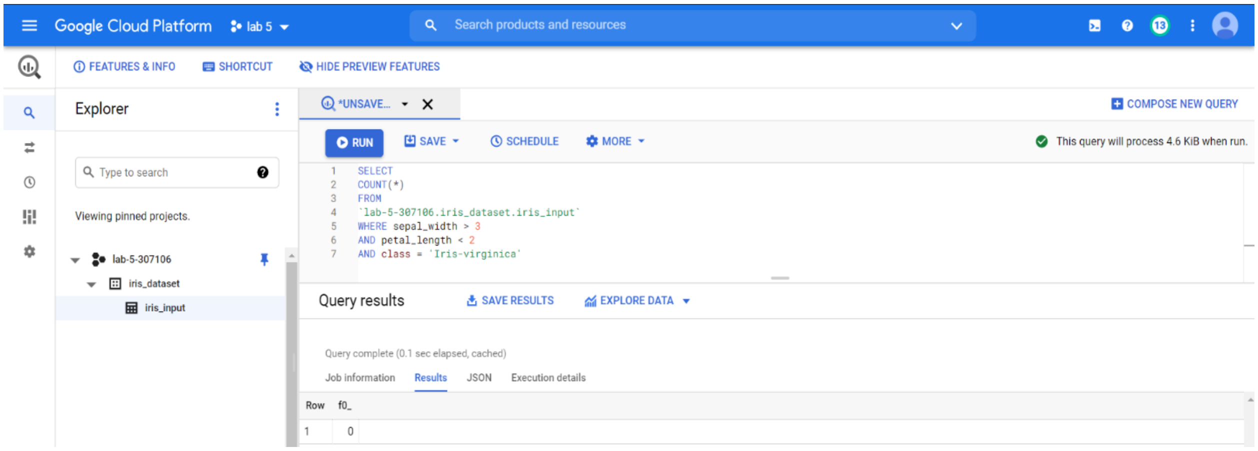


lab-5-bucket

Buckets > lab-5-bucket

	Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
□	iris.csv	4.5 KB	application/vnd.ms-excel	Mar 14, 2021, ...	Standard	Mar 14, 20...	Not public	Google-managed key	–	None
□	iris.py	2.9 KB	text/x-python	Mar 14, 2021, ...	Standard	Mar 14, 20...	Not public	Google-managed key	–	None

2)



Viewing pinned projects.

lab-5-307106
iris_dataset
iris_input

This query will process 4.6 KiB when run.

```
SELECT
COUNT(*)
FROM
`lab-5-307106.iris_dataset.iris_input`
WHERE sepal_width > 3
AND petal_length < 2
AND class = 'Iris-virginica'
```

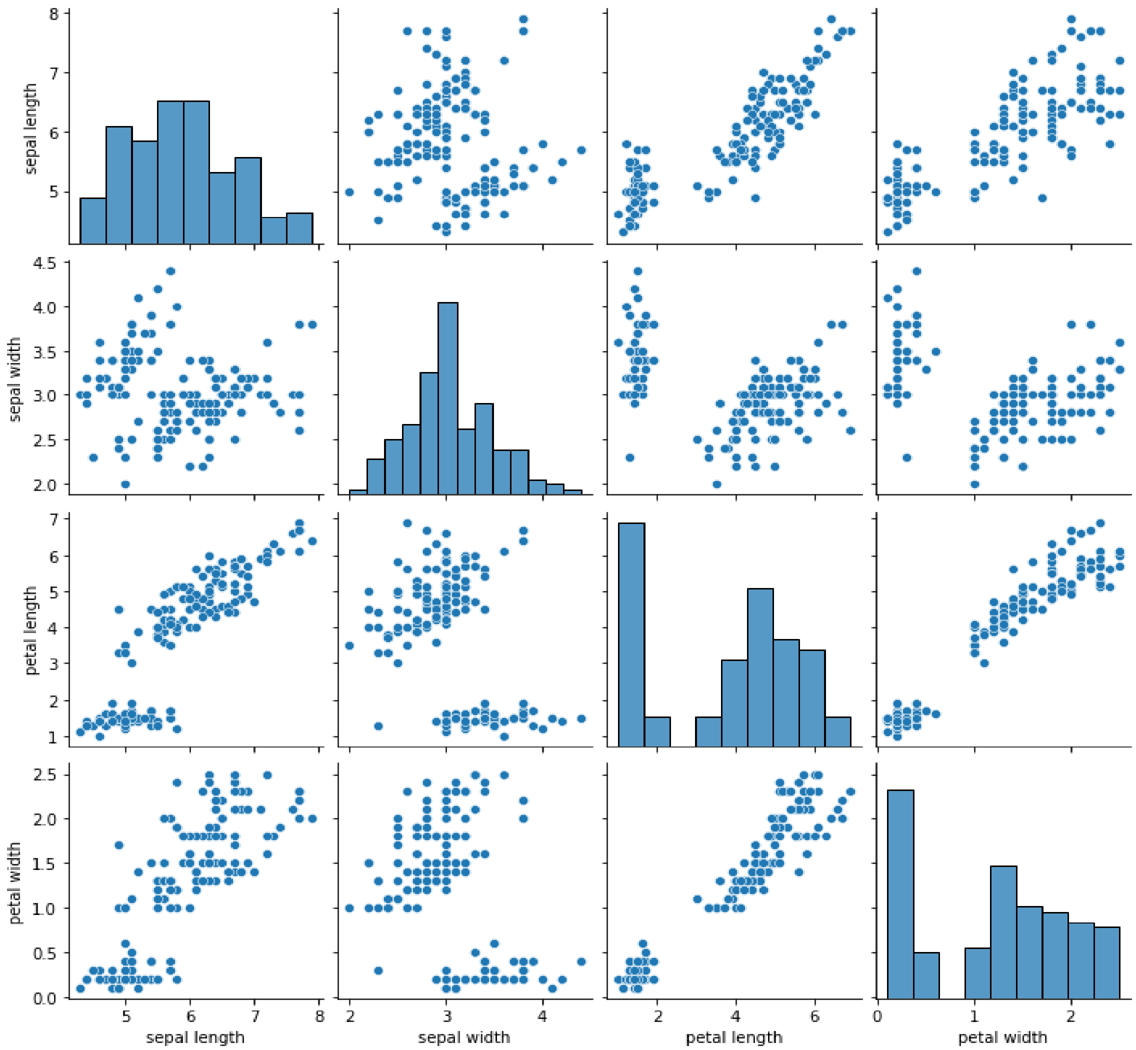
Query results

Query complete (0.1 sec elapsed, cached)

Row	f0_
1	0

The count for the given question = 0

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



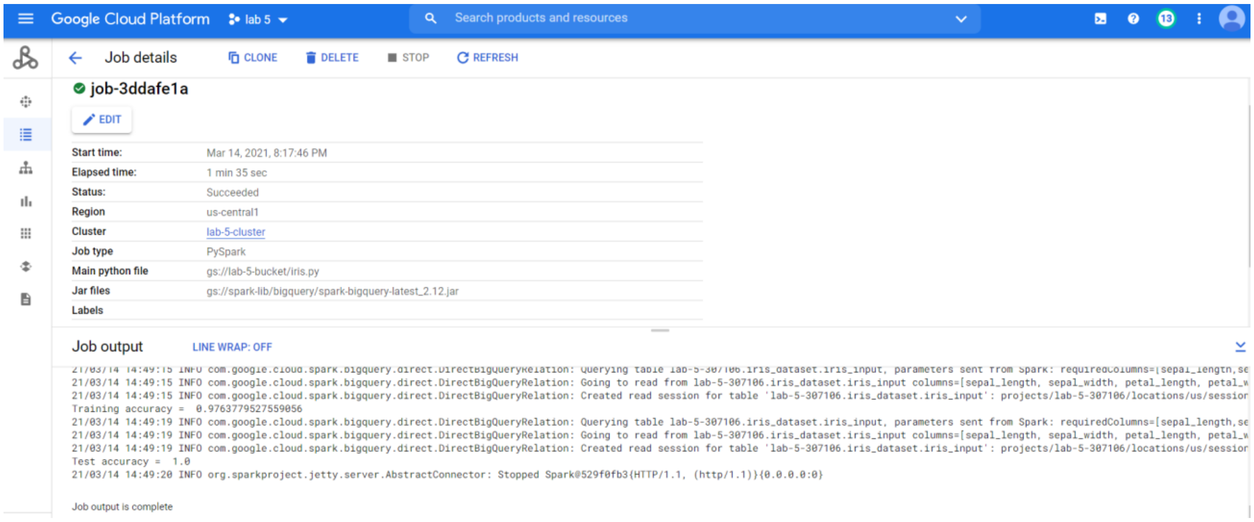
The dataset is very simple and not very completed data pre-processing techniques are required for obtaining good accuracy. While training some models Min-Max scaling was used. To model the input vs output 4 well known models are used. They are:

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



job-3ddafe1a

Start time:	Mar 14, 2021, 8:17:46 PM
Elapsed time:	1 min 35 sec
Status:	Succeeded
Region	us-central1
Cluster	lab-5-cluster
Job type	PySpark
Main python file	gs://lab-5-bucket/iris.py
Jar files	gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar
Labels	

Job output

```
21/03/14 14:49:15 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table lab-5-307106.iris_dataset.iris_input, parameters sent from Spark: requiredColumns=[sepal_length,se
21/03/14 14:49:15 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from lab-5-307106.iris_dataset.iris_input columns=[sepal_length, sepal_width, petal_length, petal_w
21/03/14 14:49:15 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'lab-5-307106.iris_dataset.iris_input': projects/lab-5-307106/locations/us/sessior
Training accuracy =  0.9763779527559056
21/03/14 14:49:19 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table lab-5-307106.iris_dataset.iris_input, parameters sent from Spark: requiredColumns=[sepal_length,se
21/03/14 14:49:19 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from lab-5-307106.iris_dataset.iris_input columns=[sepal_length, sepal_width, petal_length, petal_w
21/03/14 14:49:19 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'lab-5-307106.iris_dataset.iris_input': projects/lab-5-307106/locations/us/sessior
Test accuracy =  1.0
21/03/14 14:49:20 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@529f0fb3{HTTP/1.1, (http/1.1)}{0.0.0.0:0}

Job output is complete
```

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