

Big Data Lab - lab 5 Assignment

Vasudev Gupta (ME18B182)

1) Iris data is downloaded from the website, converted to CSV and uploaded to the GCS bucket.

The screenshot shows the Google Cloud Storage interface for the bucket 'big-data-cs4830'. The bucket is located in 'us (multiple regions in United States)' with a 'Standard' storage class, 'Subject to object ACLs' public access, and 'None' protection. The 'OBJECTS' tab is selected, showing a list of objects. Two objects are visible: 'iris.csv' (4.4 KB, application/octet-stream) and 'q2.sql' (136 B, application/x-sql). Both were created on March 5, 2022, and are not public.

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Subject to object ACLs	None

Filter by name prefix only	Filter	Filter objects and folders	Show deleted data
<input type="checkbox"/>	Name	Size	Type
<input type="checkbox"/>	iris.csv	4.4 KB	application/octet-stream
<input type="checkbox"/>	q2.sql	136 B	application/x-sql

2) The count for the given question is 0

The screenshot shows the Google Cloud Data Studio interface. A SQL query is entered in the editor, and the 'Query results' tab is selected. The query is a SELECT statement that counts the number of rows in the 'iris_table' where 'sepal_width' is greater than 3 and 'petal_length' is less than 2, and the 'class' is 'Iris-virginica'. The query is complete, and the results show a count of 0.

```

1 SELECT COUNT(*)
2 FROM `gsoc-wav2vec2.big_data_lab5.iris_table`
3 WHERE sepal_width > 3 AND petal_length < 2 AND class = 'Iris-virginica';

```

Processing location: US

Query results

Query complete (0.3 sec elapsed, 4.6 KB processed)

Row	count
1	0

Code for question-2 lies in **q2.sql**

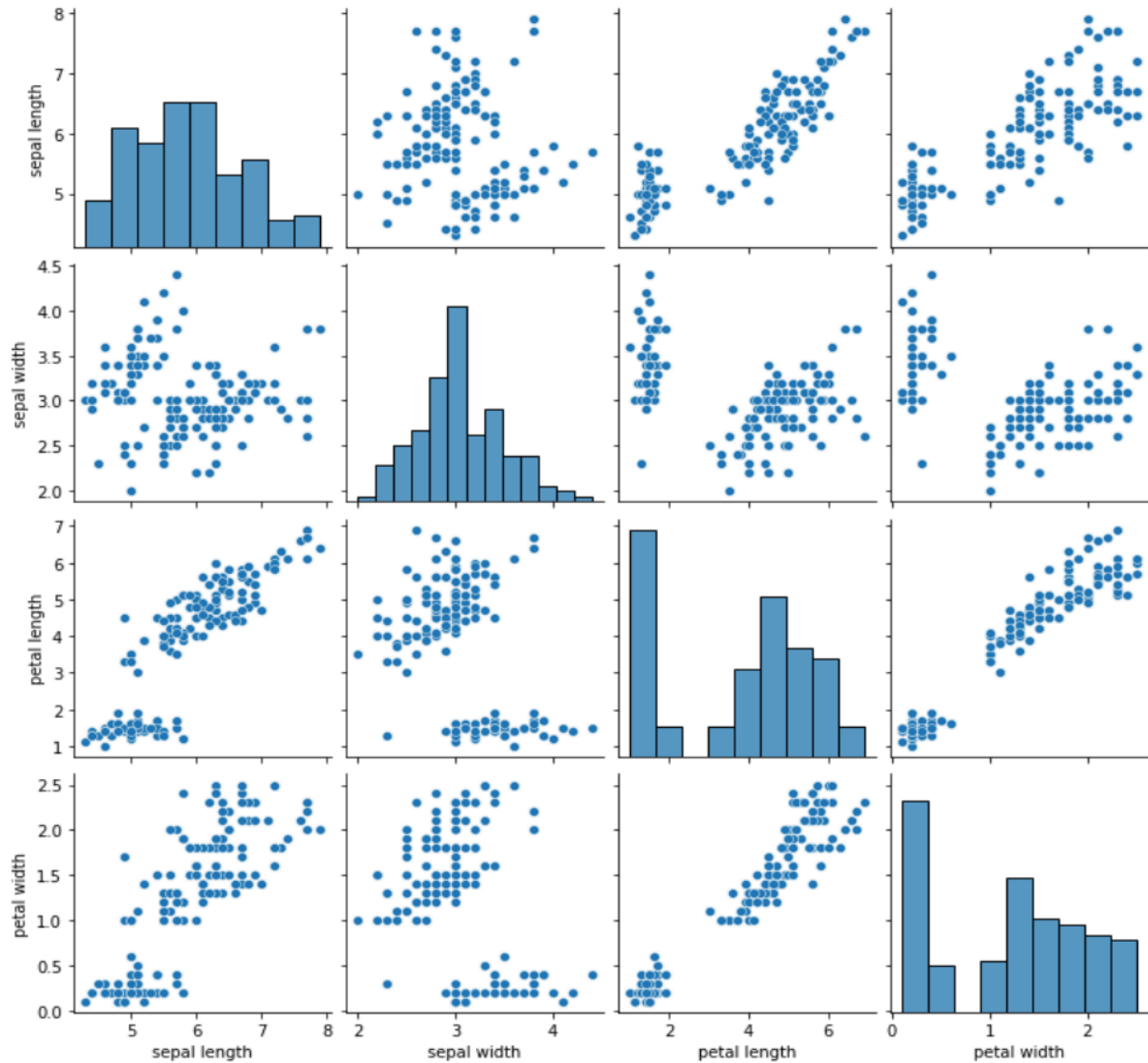
3) Dataset has 4 features and task is to predict 1 target variable. Models like Logistic Regression, Decision Trees, Random Forest Classifier and One-vs-Rest model are chosen for experimentation. The data pipeline is kept simple considering the complexity of the dataset and mix-max scalar was used for Logistic Regression & One-vs-Rest model.

Results obtained using these models are reported in the following table:

	Logistic Regression	Decision Trees	Random Forest	One-vs-Rest
Train Accuracy	85.3%	99.1%	99.13%	100%
Test Accuracy	88.2%	97.05%	97.1%	97.06%

The dataset is split into 80:20 for validating the models properly. The hyperparameters for the model can be inferred from the code directly. Code for question-3 lies in **q3.py** and logging from dataproc cluster lies in **dataproc_job.txt**

The pair plot for the the dataset can be found below:



Screen shot of the dataproc job can be found below:

[←](#) Job details [CLONE](#) [DELETE](#) [STOP](#) [REFRESH](#)

[EDIT](#)

Start time:	Mar 7, 2022, 12:58:06 AM
Elapsed time:	1 min 36 sec
Status:	Succeeded
Region	europe-west4
Cluster	lab5
Job type	PySpark
Main python file	gs://dataproc-staging-europe-west4-290596474871-iiqgnvt5/google-cloud-dataproc-metainfo/79bc07bb-6d9b-40c5-a812-f11b14978a7c/jobs/62308ff7dd5a4d1ca971fcfe06ff9699/staging/q3.py
Jar files	gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar
Labels	

Output

LINE WRAP: OFF

```
22/03/06 19:29:39 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table gsoc-wavZ
22/03/06 19:29:39 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from gsoc-v
22/03/06 19:29:40 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for
Test Accuracy: 0.9705882352941176
*****
22/03/06 19:29:40 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@51e03029{HTTP/1.1, (http,
```