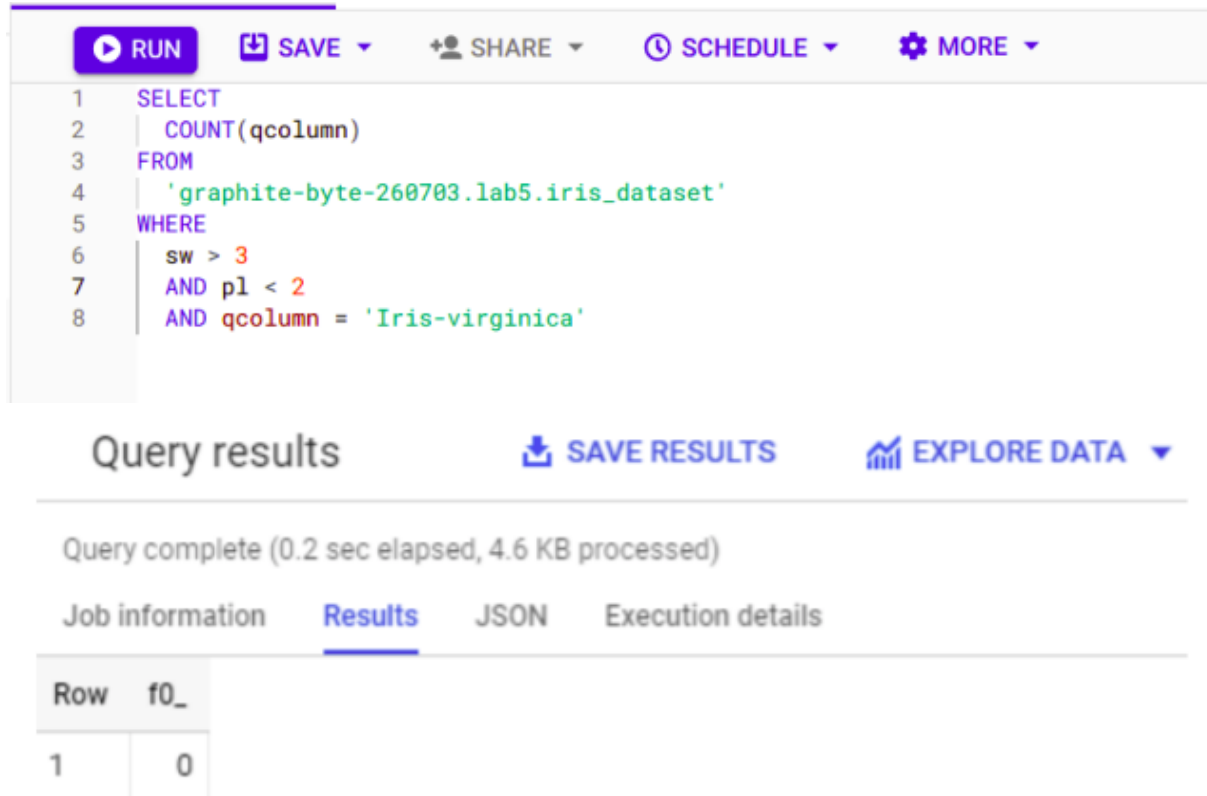


CS4830: Lab Assignment-5

Shania Mitra CH18B067

Q1. Count using BigQuery the number of Iris Virginica flowers which have sepal width greater than 3 cm and petal length smaller than 2 cm



The screenshot displays the Google BigQuery web interface. At the top, there are buttons for 'RUN', 'SAVE', 'SHARE', 'SCHEDULE', and 'MORE'. Below these, a SQL query is entered in the editor:

```
1 SELECT
2   COUNT(qcolumn)
3 FROM
4   'graphite-byte-260703.lab5.iris_dataset'
5 WHERE
6   sw > 3
7   AND pl < 2
8   AND qcolumn = 'Iris-virginica'
```

Below the query editor, the 'Query results' section is visible. It shows 'Query complete (0.2 sec elapsed, 4.6 KB processed)'. There are tabs for 'Job information', 'Results' (which is selected), 'JSON', and 'Execution details'. Under the 'Results' tab, a table is displayed with the following data:

Row	f0_
1	0

Thus, there are no flowers with sepal width greater than 3 cm and petal length smaller than 2 cm

Q2. Train a classification model on the dataset and report the accuracy for different preprocessing techniques and models. Provide the details of data exploration and feature engineering steps

The dataset consists of 4 features and 150 samples. Further, there are no missing values. On plotting we observe that there is a clear decision boundary. Thus, we use Logistic Regression and Random Forests to perform classification on the dataset. The classes are also well balanced making accuracy a suitable metric for evaluation.

Upon performing Standardisation and training the models we obtain:

<u>Model</u>	<u>Accuracy</u>
Random Forest	97.213%
Logistic Regression	96.458%

Thus, we observe that Random Forest performs better than Logistic regression giving us an accuracy of 97%
