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

RUN SAVE SHARE SCHEDULE MORE

```sql
SELECT
  COUNT(qcolumn)
FROM
  'graphite-byte-260703.lab5.iris_dataset'
WHERE
  sw > 3
  AND pl < 2
  AND qcolumn = 'Iris-virginica'
```

Query results SAVE RESULTS EXPLORE DATA

Query complete (0.2 sec elapsed, 4.6 KB processed)

Job information Results JSON Execution details

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:

Model	**Accuracy**
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%
