Assignment - Lab 5

ME16B172 - Sushant Uttam Wadavkar

1. Dataset has been downloaded and uploaded to my google cloud bucket.

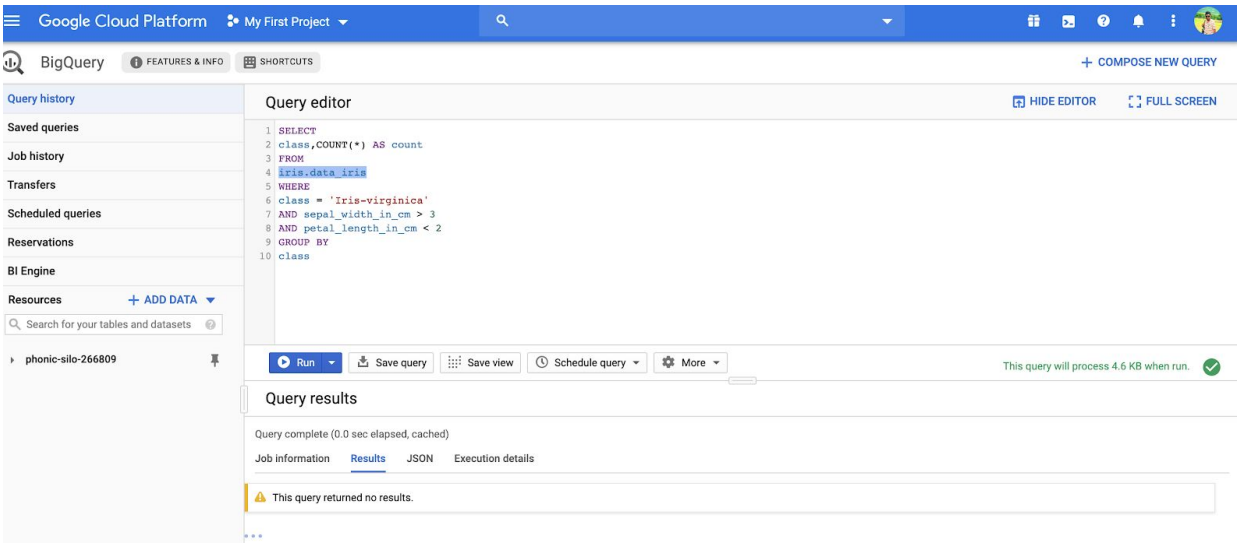
2. Bigquery code has been attached in zip

DL5_2_iris_bigquery.txt

```
SELECT
class,COUNT(*) AS count
FROM
iris.data_iris
WHERE
class = 'Iris-virginica'
AND sepal_width_in_cm > 3
AND petal_length_in_cm < 2
GROUP BY
class
```

   a. Count the number of Iris Virginica flowers which have sepal width greater than 3 cm and petal length smaller than 2 cm. The count = 0

   b.

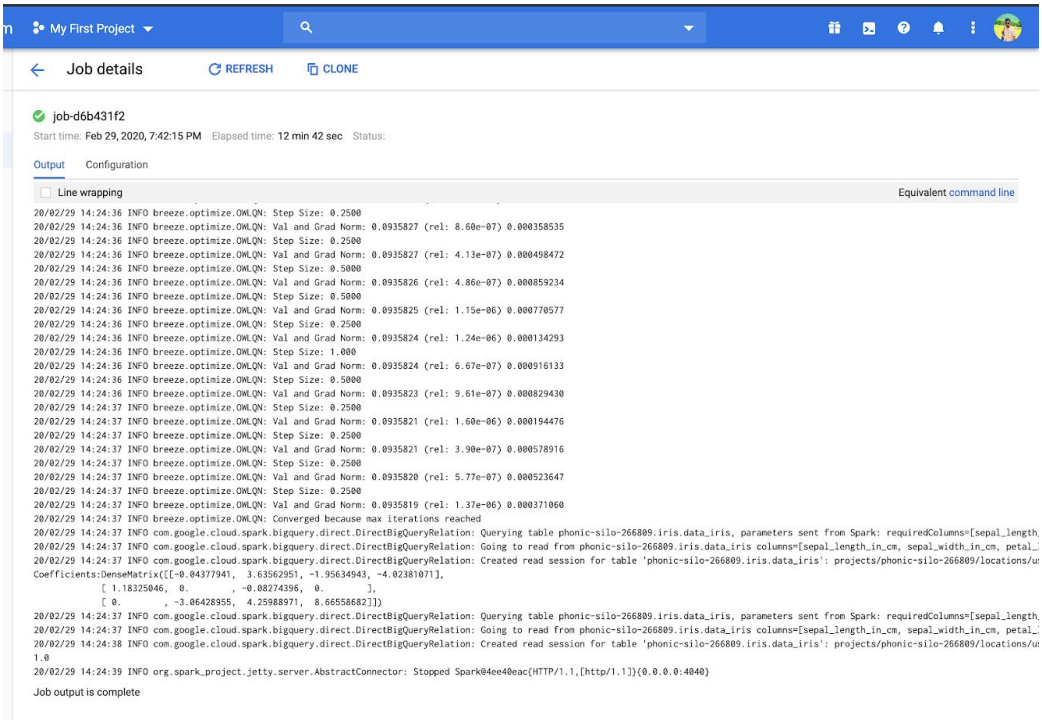


3. The classification models have been trained. The details are as follows,

    a. Logistic Regression:

        i. Accuracy = 1.0

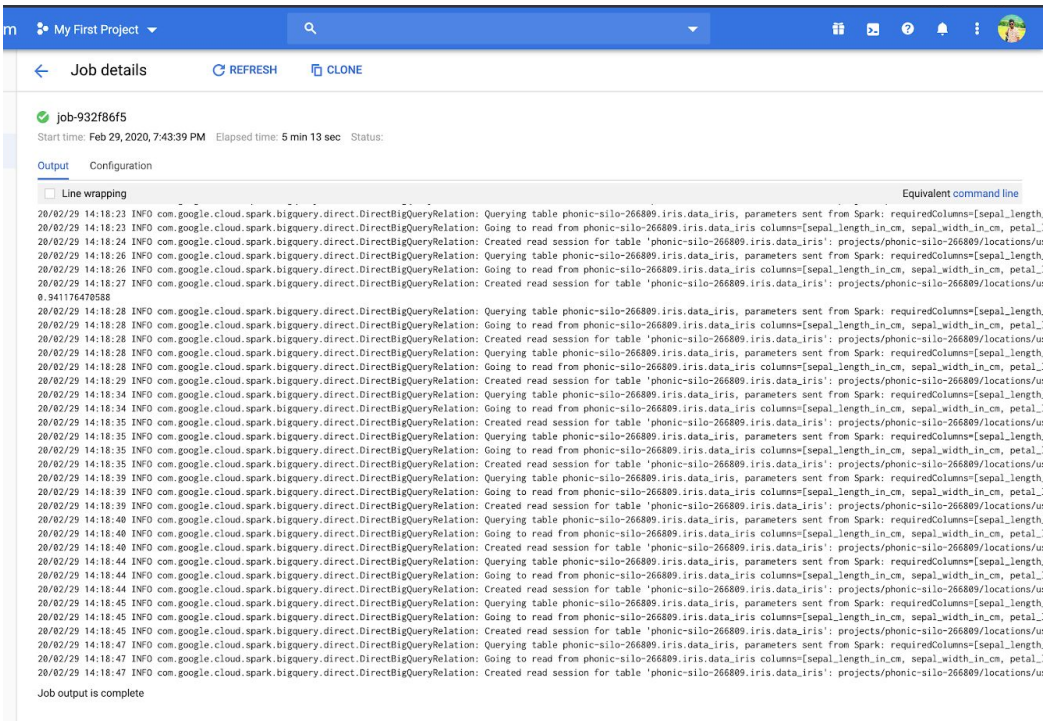
        ii.



b. Decision Tree Classifier:

i. Accuracy = 0.941176470588

ii.



The details of data exploration and feature engineering steps:

1. In the first model, I used Polynomial Expansion with degree 2 (given in the code).
2. Then it is processed with StandardScaler to normalize the distribution.
3. As I'm using logistic regression, I had to use MulticlassClassificationEvaluator, the regression parameters have been taken as default.
4. In the second model, I used DecisionTreeClassifier, with the polynomial expansion degree as 3 (as specified in the code).
5. The maxDepth has been set to 2.
6. The observation is that the model with Logistic Regression gives better accuracy than the Decision Tree Classification technique.

Codes for Part 2 and 3 have been attached in the zip file.