



Assignment - Lab 5

ME16B172 - Sushant Uttam Wadaokar

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

The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes 'Google Cloud Platform', 'My First Project', and a search icon. Below the navigation bar, the 'BigQuery' section is active, showing 'FEATURES & INFO' and 'SHORTCUTS'. The left sidebar contains a 'Query history' panel with sections for 'Saved queries', 'Job history', 'Transfers', 'Scheduled queries', 'Reservations', 'BI Engine', and 'Resources'. The main area is the 'Query editor', which displays the SQL query from the previous block. Below the editor, there are buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. The 'Query results' section at the bottom shows 'Query complete (0.0 sec elapsed, cached)' and 'Job information', 'Results', 'JSON', and 'Execution details'. A warning icon indicates 'This query returned no results.'

- b.

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

a. Logistic Regression:

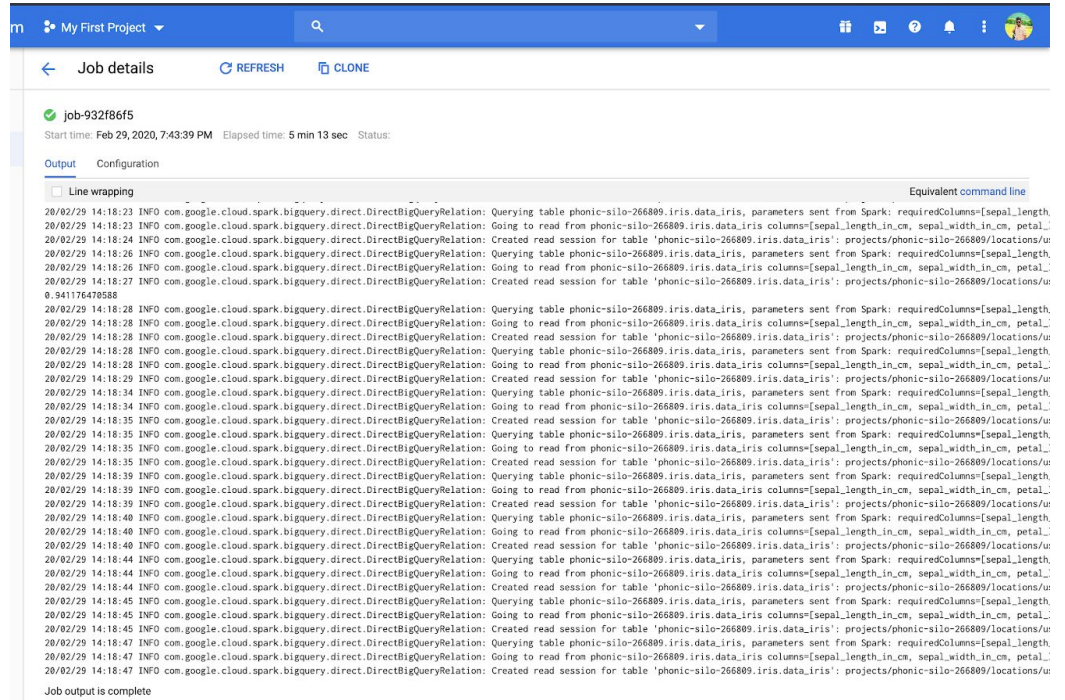
i. Accuracy = 1.0

```
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935827 (rel: 8.60e-07) 0.000358535
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935827 (rel: 4.13e-07) 0.000498472
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.5000
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935826 (rel: 4.86e-07) 0.000859234
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.5000
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935825 (rel: 1.15e-06) 0.000770577
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935824 (rel: 1.24e-06) 0.000134293
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 1.000
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935824 (rel: 6.67e-07) 0.000916133
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Step Size: 0.5000
20/02/29 14:24:36 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935823 (rel: 9.61e-07) 0.000829430
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935821 (rel: 1.60e-06) 0.000194476
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935821 (rel: 3.90e-07) 0.000578916
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935820 (rel: 5.77e-07) 0.000523647
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Step Size: 0.2500
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Val and Grad Norm: 0.0935819 (rel: 1.37e-06) 0.000371060
20/02/29 14:24:37 INFO breeze.optimize.OMLQN: Converged because max iterations reached
20/02/29 14:24:37 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table phonic-silo-266809.iris.data.iris, parameters sent from Spark: requiredColumns=[sepal_length
20/02/29 14:24:37 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from phonic-silo-266809.iris.data.iris columns=[sepal_length_in_cm, sepal_width_in_cm, petal_
20/02/29 14:24:37 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'phonic-silo-266809.iris.data.iris': projects/phonic-silo-266809/locations/us
Coefficients:DenseMatrix([[ -0.04377941, 3.63562951, -1.95634943, -4.02381071],
[ 0. , -3.06428955, 4.25988971, 8.66558682]])
20/02/29 14:24:37 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table phonic-silo-266809.iris.data.iris, parameters sent from Spark: requiredColumns=[sepal_length
20/02/29 14:24:37 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from phonic-silo-266809.iris.data.iris columns=[sepal_length_in_cm, sepal_width_in_cm, petal_
20/02/29 14:24:38 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'phonic-silo-266809.iris.data.iris': projects/phonic-silo-266809/locations/us
1.0
20/02/29 14:24:39 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@4ee40eac(HTTP/1.1,[http/1.1])(0.0.0.0:4040)
Job output is complete
```

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.