

CS4830 Assignment 3

Shashank H S

BE18B006

2. Query and query result are shown in the picture below. No Iris virginica flowers fit the given constraints. Full picture in zipped folder.

The screenshot shows a web-based SQL query editor. At the top, there's a tab labeled '*QUERY L...' with a close button. Below the tab are buttons for 'RUN', 'SAVE', 'SHARE', and a clock icon. The query text is as follows:

```
1 select COUNT(*)
2 from Iris.irisdata
3 where (class = "Iris-virginica")
4 and (sepalwidth>3)
5 and (sepallen<2)
```

Below the query, it says 'Processing location: US'. Underneath, there's a section titled 'Query results' with a 'SAVE RESULTS' button. Below this, it states 'Query complete (0.3 sec elapsed, 4.6 KB processed)'. There are tabs for 'Job information', 'Results' (which is selected), 'JSON', and 'Execution details'. The 'Results' tab shows a table with one column 'f0_' and one row with the value '0'.

Row	f0_
1	0

3. 80% of the data was used for training and 20% was used for testing.

Preprocessing	Model	Test set accuracy
None	Random Forest classifier	0.92
None	Logistic regression	0.805556
None	Multilayer Perceptron	0.9375
Standard scaler	Multilayer Perceptron	0.933333
Standard scaler and PCA	Multilayer Perceptron	0.936

A screenshot for the Multilayer Perceptron training job in Dataproc.

```
Output LINE WRAP: OFF
22/03/12 18:20:27 INFO breeze.optimize.LBFGS: Val and Grad Norm: 9.52998e-10 (rel: 0.000861) 3.5534 accuracy
22/03/12 18:20:27 INFO breeze.optimize.LBFGS: Step Size: 1.000
22/03/12 18:20:27 INFO breeze.optimize.LBFGS: Val and Grad Norm: 4.74100e-10 (rel: 0.000425) 1.96312e-07
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Step Size: 1.000
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Val and Grad Norm: 2.34651e-10 (rel: 0.000213) 6.03655e-08
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Step Size: 1.000
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Val and Grad Norm: 1.08795e-10 (rel: 0.000112) 5.18279e-08
22/03/12 18:20:28 INFO breeze.optimize.StrongWolfeLineSearch: Line search t: 0.452950536227279 fval: 6.687026626369691E-7 rhs: 1.0872262156179486E-10 cdd: 1.8365
22/03/12 18:20:28 INFO breeze.optimize.StrongWolfeLineSearch: Line search t: 0.27303941992053743 fval: 7.3685354311642625E-9 rhs: 1.087512971729888E-10 cdd: 1.66
22/03/12 18:20:28 INFO breeze.optimize.StrongWolfeLineSearch: Line search t: 0.13883359221211625 fval: 4.442281721799699E-10 rhs: 1.0877268792534657E-10 cdd: 8.5
22/03/12 18:20:28 INFO breeze.optimize.StrongWolfeLineSearch: Line search t: 0.0393982009090819 fval: 9.406269675399626E-11 rhs: 1.0878853669933261E-10 cdd: 6.00
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Step Size: 0.03940
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Val and Grad Norm: 9.40627e-11 (rel: 1.31e-05) 1.29421e-08
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Step Size: 1.000
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Val and Grad Norm: 4.31361e-11 (rel: 4.52e-05) 3.78932e-09
22/03/12 18:20:28 INFO breeze.optimize.LBFGS: Converged because gradient converged
22/03/12 18:20:28 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table big-data-lab-341515.iris.irisdata, parameters sent from Spar
22/03/12 18:20:28 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from big-data-lab-341515.iris.irisdata columns=[sepalen, sep
22/03/12 18:20:28 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'big-data-lab-341515.iris.irisdata': project
22/03/12 18:20:29 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table big-data-lab-341515.iris.irisdata, parameters sent from Spar
22/03/12 18:20:29 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from big-data-lab-341515.iris.irisdata columns=[sepalen, sep
22/03/12 18:20:29 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'big-data-lab-341515.iris.irisdata': project
Test set accuracy = 0.9375
22/03/12 18:20:30 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@264398b1{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

Multilayer Perceptron performs the best. PCA reduces its accuracy since compression is not lossless.