CS5542 Big Data Apps and Analytics LAB ASSIGNMENT #4

Name:Sidrah Junaid

Class Id:15

1. Write a spark program for the following Machine Learning Tasks

Answer

Selected data set contain different images of birds. By classifying the images we can easily identify the different birds.

Dataset:











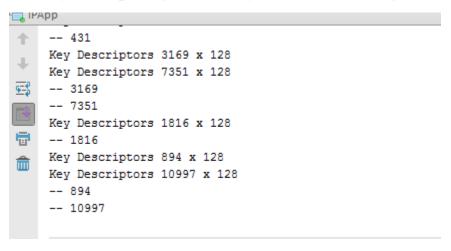






The dataset is split into 70% of Training data and 30% of test data.

The key descriptors generated by Decision Tree Algorithm are:



Histogram:

```
# Histogram size: (400, 1)
Histogram : [ 9.525143E-4, 0.0017684669, 0.001243233, 6.8017957E-4, 0.0019045029, 0.0014963951, 0.0014963951, 0.0058495444, 0.0023126106, 0.0017684669, 0.0017684669, 0.0044891

# Histogram : [ 9.5025143E-4, 0.0017684669, 0.001243233, 6.8017957E-4, 0.0019045029, 0.0014963951, 0.0014963951, 0.0058495444, 0.0023126106, 0.0017684669, 0.0017684669, 0.0044891

# Histogram size: (400, 1)
Histogram : [ 0.0021026898, 0.002208898, 0.0031555698, 0.0044177976, 0.0037866835, 0.0025244558, 0.0025244558, 3.1555697E-4, 0.002208988, 0.0012622279, 0.0025244558, 0.0028400

# Histogram : [ 0.002202643, 0.0033039646, 0.0, 0.0060572685, 5.5066077E-4, 0.0, 0.0, 0.0038546254, 5.5066077E-4, 0.0, 0.001103215, 0.0, 0.0, 0.002202643, 5.5066077E-4, 5.5066077

# Histogram : [ 0.002202643, 0.0033039646, 0.0, 0.0060572685, 5.5066077E-4, 0.0, 0.001185682, 0.0067114094, 0.0022371365, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.001185682, 0.00188857747, 0.-1

# Histogram size: (400, 1)
Histogram : [ 0.003276287, 0.00472073, 9.1779203E-4, 0.0032778287, 0.0012835841, 0.001311315, 0.0014422446, 0.005900915, 0.0011800183, 0.0011800183, 0.0015733577, 0.00445784

# Histogram : [ 0.003278287, 0.00472073, 9.1779203E-4, 0.0032778287, 0.0012835841, 0.001311315, 0.0014422446, 0.005900915, 0.0011800183, 0.0011800183, 0.0015733577, 0.00445784

# Histogram : [ 0.003276287, 0.00472073, 9.1779203E-4, 0.003278287, 0.0018258841, 0.001592061, 0.001592061, 0.00241473, 0.001592061, 0.0013822855, 0.0022638093, 0.00482946, 0.-1

# Histogram : [ 0.003276287, 0.00472073, 9.1779203E-4, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467532, 0.0032467
```

Confusion Matrix

```
Codecroor. Got prand-new decompressor ..dz
    1
        17/02/16 22:48:01 INFO InternalParquetRecordReader: block read in memory in 8 ms. row count = 2
         Predicting test image : sparrow as duck
+
         (0.0, 3)
III <u>5</u>
         (0.0,3)
   O
         (1.0,3)
         (1.0,3)
   +[]
         (3.0,2)
         (2.0, 2)
(2.0, 2)
         (2.0,1)
160
         (1.0, 1)
×
         (3.0, 1)
?
         (1.0.0)
         (1.0,0)
        AAAAAA0.23076923076923078
                             = Confusion matrix ====
        0.0 2.0 0.0 1.0
        0.0 0.0 2.0 1.0
        2.0 2.0 0.0 0.0
        0.23076923076923078
         17/02/16 22:48:03 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
        17/02/16 22:48:03 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
```

Accuracy

23.07%

Error Rate