

Experiment 8

Aim: Machine Learning Model Training with Spark MLlib: Train machine learning models on large datasets using Spark's MLlib library, and evaluate model performance using techniques such as cross-validation and model selection.

Spark Machine Learning Model:

```
Dataset Schema:
root
|-- sepal length (cm): double (nullable = true)
|-- sepal width (cm): double (nullable = true)
|-- petal length (cm): double (nullable = true)
|-- petal width (cm): double (nullable = true)
|-- target: long (nullable = true)

First 5 rows of the dataset:
```

```
24/11/24 18:46:05 INFO CodeGenerator: Code generated in 51.6571 ms
+-----+-----+-----+-----+-----+
|sepal length (cm)|sepal width (cm)|petal length (cm)|petal width (cm)|target|
+-----+-----+-----+-----+-----+
|5.1|3.5|1.4|0.2|0|
|4.9|3.0|1.4|0.2|0|
|4.7|3.2|1.3|0.2|0|
|4.6|3.1|1.5|0.2|0|
|5.0|3.6|1.4|0.2|0|
+-----+-----+-----+-----+-----+
only showing top 5 rows

Classes in the dataset:
Class 0: setosa
Class 1: versicolor
Class 2: virginica
24/11/24 18:46:06 INFO CodeGenerator: Code generated in 43.4243 ms
```

Fig 1: Dataset Schema

Fig 2: First Rows and Classes of Dataset

```
24/11/24 18:46:40 INFO DAGScheduler: Job 10 finished: collect
24/11/24 18:46:40 INFO TorrentBroadcast: Destroying Broadcast
24/11/24 18:46:40 INFO RandomForest: Internal timing for Dec
24/11/24 18:46:40 INFO BlockManagerInfo: Removed broadcast_2
24/11/24 18:46:40 INFO RandomForest: init: 0.0031386
total: 7.8614862
findBestSplits: 7.8489939
chooseSplits: 7.8430047
24/11/24 18:46:40 INFO MapPartitionsRDD: Removing RDD 29 from
24/11/24 18:46:40 INFO BlockManager: Removing RDD 29
24/11/24 18:46:40 INFO TorrentBroadcast: Destroying Broadcast
24/11/24 18:46:40 INFO Instrumentation: [7527c3f4] training
24/11/24 18:46:40 INFO BlockManagerInfo: Removed broadcast_7
24/11/24 18:46:41 INFO BlockManagerInfo: Removed broadcast_1
24/11/24 18:46:41 INFO BlockManagerInfo: Removed broadcast_1
```

Fig 3: Best Models and Parameters

```
24/11/24 18:46:44 INFO CodeGenerator: Code
+-----+-----+-----+
|features|label|prediction|
+-----+-----+-----+
|[4.6,3.1,1.5,0.2]|0.0|0.0|
|[4.8,3.4,1.6,0.2]|0.0|0.0|
|[4.9,3.1,1.5,0.1]|0.0|0.0|
|[5.4,3.7,1.5,0.2]|0.0|0.0|
|[4.6,3.6,1.0,0.2]|0.0|0.0|
|[5.0,3.0,1.6,0.2]|0.0|0.0|
|[5.0,3.2,1.2,0.2]|0.0|0.0|
|[5.4,3.4,1.5,0.4]|0.0|0.0|
|[4.4,3.2,1.3,0.2]|0.0|0.0|
|[5.0,3.5,1.3,0.3]|0.0|0.0|
+-----+-----+-----+
only showing top 10 rows
```

Fig 4: Classification Report with prediction

```
24/11/24 18:46:51 INFO DAGScheduler: Job 15 is finished. Cancelling potential
24/11/24 18:46:51 INFO DAGScheduler:
Model Accuracy: 100.00%

Classification Report:
24/11/24 18:46:51 INFO BlockManagerInfo: Removed broadcast_1
24/11/24 18:46:51 INFO BlockManagerInfo: Removed broadcast_1
24/11/24 18:46:51 INFO BlockManagerInfo: Removed broadcast_1
```

Fig 5: Model Accuracy

```
24/11/24 18:46:58 INFO DAGScheduler: Job 15 is finished. Cancelling potential
24/11/24 18:46:58 INFO TaskSchedulerImpl: Killing all running tasks in stage
24/11/24 18:46:58 INFO DAGScheduler: Job 15 finished: showString at NativeMet
+-----+-----+-----+
|label|prediction|count|
+-----+-----+-----+
|0.0|0.0|13|
|1.0|1.0|8|
|2.0|2.0|13|
+-----+-----+-----+

24/11/24 18:46:58 INFO SparkContext: SparkContext is stopping with exitCode 0
24/11/24 18:46:58 INFO SparkUI: Stopped Spark web UI at http://MADHURIMA-RAWA
24/11/24 18:46:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMaster
24/11/24 18:46:58 INFO MemoryStore: MemoryStore cleared
24/11/24 18:46:58 INFO BlockManager: BlockManager stopped
```

Fig 6: Label Count and Prediction