JUNAID GIRKAR

60004190057

TE COMPS A4

**EXPERIMENT 8**

**AIM:** Execute any two ML Algorithms using Apache Spark MLlib and compare the results.

**THEORY:**

**Spark MLlib:**

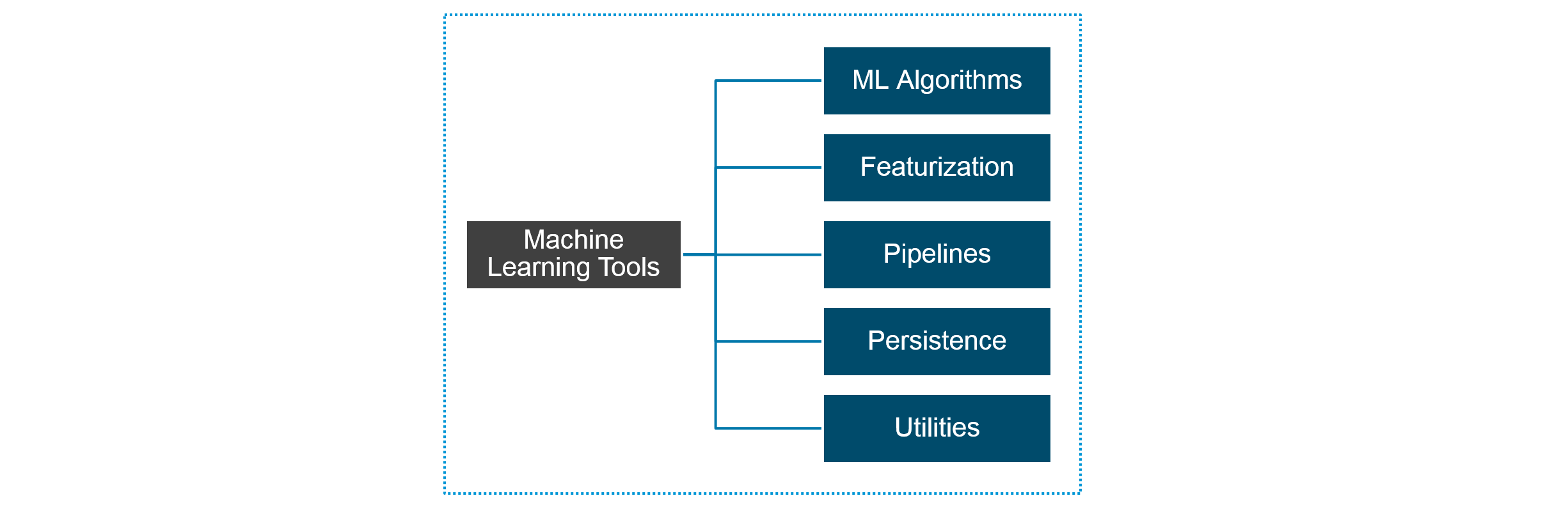
Spark MLlib is Apache Spark’s Machine Learning component. MLlib consists of popular algorithms and utilities. MLlib in Spark is a scalable Machine learning library that discusses both high-quality algorithm and high speed. The machine learning algorithms like regression, classification, clustering, pattern mining, and collaborative filtering. Lower level machine learning primitives like generic gradient descent optimization algorithm are also present in MLlib.

spark.mllib contains the original API built on top of RDDs. It is currently in maintenance mode.

spark.ml provides higher level API built on top of DataFrames for constructing ML pipelines. spark.ml is the primary Machine Learning API for Spark at the moment.

**Spark MLlib Tools :**

* **ML Algorithms:** ML Algorithms form the core of MLlib. These include common learning algorithms such as classification, regression, clustering and collaborative filtering.
* **Featurization:** Featurization includes feature extraction, transformation, dimensionality reduction and selection.
* **Pipelines:** Pipelines provide tools for constructing, evaluating and tuning ML Pipelines.
* **Persistence:** Persistence helps in saving and loading algorithms, models and Pipelines.
* **Utilities:** Utilities for linear algebra, statistics and data handling.



**MLlib Algorithms:**

The popular algorithms and utilities in Spark MLlib are:

* Basic Statistics
* Regression
* Classification
* Recommendation System
* Clustering
* Dimensionality Reduction
* Feature Extraction
* Optimization

**CODE:**

| !apt-get install openjdk-8-jdk-headless -qq |
| --- |

| !apt-get·install·openjdk-8-jdk-headless·-qq |
| --- |

| Selecting previously unselected package openjdk-8-jre-headless:amd64. (Reading database ... 155632 files and directories currently installed.) Preparing to unpack .../openjdk-8-jre-headless\_8u312-b07-0ubuntu1~18.04\_amd64.deb ... Unpacking openjdk-8-jre-headless:amd64 (8u312-b07-0ubuntu1~18.04) ... Selecting previously unselected package openjdk-8-jdk-headless:amd64. Preparing to unpack .../openjdk-8-jdk-headless\_8u312-b07-0ubuntu1~18.04\_amd64.deb ... Unpacking openjdk-8-jdk-headless:amd64 (8u312-b07-0ubuntu1~18.04) ... Setting up openjdk-8-jre-headless:amd64 (8u312-b07-0ubuntu1~18.04) ... update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertool (servertool) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv (tnameserv) in auto mode Setting up openjdk-8-jdk-headless:amd64 (8u312-b07-0ubuntu1~18.04) ... update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/idlj to provide /usr/bin/idlj (idlj) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsimport to provide /usr/bin/wsimport (wsimport) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jsadebugd to provide /usr/bin/jsadebugd (jsadebugd) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/native2ascii to provide /usr/bin/native2ascii (native2ascii) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/javah to provide /usr/bin/javah (javah) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/hsdb to provide /usr/bin/hsdb (hsdb) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/clhsdb to provide /usr/bin/clhsdb (clhsdb) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/xjc to provide /usr/bin/xjc (xjc) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/schemagen to provide /usr/bin/schemagen (schemagen) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/extcheck to provide /usr/bin/extcheck (extcheck) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jhat to provide /usr/bin/jhat (jhat) in auto mode update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in auto mode |
| --- |

| !wget -q https://archive.apache.org/dist/spark/spark-3.0.0/spark-3.0.0-bin-hadoop3.2.tgz |
| --- |

| !tar xf spark-3.0.0-bin-hadoop3.2.tgz |
| --- |

| !pip install -q findspark |
| --- |

| import os os.environ["JAVA\_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64" os.environ["SPARK\_HOME"] = "/content/spark-3.0.0-bin-hadoop3.2"  import findspark findspark.init()  findspark.find() |
| --- |

| **/content/spark-3.0.0-bin-hadoop3.2** |
| --- |

| from pyspark.sql import SparkSession  spark = SparkSession.builder\  .master("local")\  .appName("Colab")\  .config('spark.ui.port', '4050')\  .getOrCreate()  spark |
| --- |

| **SparkSession - in-memory  SparkContext  Spark UI  Version v3.0.0 Master local AppName Colab** |
| --- |

| **!wget https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip !unzip ngrok-stable-linux-amd64.zip** |
| --- |

| **--2022-06-15 14:23:32-- https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip Resolving bin.equinox.io (bin.equinox.io)... 54.161.241.46, 18.205.222.128, 54.237.133.81, ... Connecting to bin.equinox.io (bin.equinox.io)|54.161.241.46|:443... connected. HTTP request sent, awaiting response... 200 OK Length: 13832437 (13M) [application/octet-stream] Saving to: 'ngrok-stable-linux-amd64.zip'  ngrok-stable-linux- 100%[===================>] 13.19M 51.8MB/s in 0.3s   2022-06-15 14:23:32 (51.8 MB/s) - 'ngrok-stable-linux-amd64.zip' saved [13832437/13832437]  Archive: ngrok-stable-linux-amd64.zip  inflating: ngrok** |
| --- |

| **get\_ipython().system\_raw('./ngrok http 4050 &') !curl -s http://localhost:4040/api/tunnels** |
| --- |

| **{"tunnels":[],"uri":"/api/tunnels"}** |
| --- |

| **!/content/ngrok authtoken 28vGHKeB8z7Z5fpbJUql9gT8jl5\_Qm96Hsui43Um6XjhYq3o** |
| --- |

| **Authtoken saved to configuration file: /root/.ngrok2/ngrok.yml** |
| --- |

| **!ls -la** |
| --- |

| **!wget https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip !unzip ngrok-stable-linux-amd64.zip** |
| --- |

| **--2022-06-15 14:23:32-- https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip Resolving bin.equinox.io (bin.equinox.io)... 54.161.241.46, 18.205.222.128, 54.237.133.81, ... Connecting to bin.equinox.io (bin.equinox.io)|54.161.241.46|:443... connected. HTTP request sent, awaiting response... 200 OK Length: 13832437 (13M) [application/octet-stream] Saving to: 'ngrok-stable-linux-amd64.zip'  ngrok-stable-linux- 100%[===================>] 13.19M 51.8MB/s in 0.3s   2022-06-15 14:23:32 (51.8 MB/s) - 'ngrok-stable-linux-amd64.zip' saved [13832437/13832437]  Archive: ngrok-stable-linux-amd64.zip  inflating: ngrok** |
| --- |

| **get\_ipython().system\_raw('./ngrok·http·4050·&') !curl·-s·http://localhost:4040/api/tunnels** |
| --- |

| **{"tunnels":[],"uri":"/api/tunnels"}** |
| --- |

| **!/content/ngrok·authtoken·28vGHKeB8z7Z5fpbJUql9gT8jl5\_Qm96Hsui43Um6XjhYq3o** |
| --- |

| **Authtoken saved to configuration file: /root/.ngrok2/ngrok.yml** |
| --- |

| **!ls·-la·** |
| --- |

| **total 262080 drwxr-xr-x 1 root root 4096 Jun 15 14:23 . drwxr-xr-x 1 root root 4096 Jun 15 14:20 .. drwxr-xr-x 4 root root 4096 Jun 1 13:49 .config -rwxr-xr-x 1 root root 30053267 May 4 2021 ngrok -rw-r--r-- 1 root root 13832437 Jun 15 14:23 ngrok-stable-linux-amd64.zip drwxr-xr-x 1 root root 4096 Jun 1 13:50 sample\_data drwxr-xr-x 13 1000 1000 4096 Jun 6 2020 spark-3.0.0-bin-hadoop3.2 -rw-r--r-- 1 root root 224453229 Jun 6 2020 spark-3.0.0-bin-hadoop3.2.tgz** |
| --- |

| **!wget https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data** |
| --- |

| **--2022-06-15 14:23:33-- https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data Resolving archive.ics.uci.edu (archive.ics.uci.edu)... 128.195.10.252 Connecting to archive.ics.uci.edu (archive.ics.uci.edu)|128.195.10.252|:443... connected. HTTP request sent, awaiting response... 200 OK Length: 4551 (4.4K) [application/x-httpd-php] Saving to: 'iris.data'  iris.data 100%[===================>] 4.44K --.-KB/s in 0s   2022-06-15 14:23:34 (61.3 MB/s) - 'iris.data' saved [4551/4551]** |
| --- |

| **!wget https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names** |
| --- |

EDA:

| !cat dataset/iris.names |
| --- |

| cat: dataset/iris.names: No such file or directory |
| --- |

| import pandas as pd import seaborn as sns from sklearn import preprocessing |
| --- |

| df = pd.read\_csv('iris.data', header=None, names=["D1", "D2", "D3", "D4", "Class"])  le = preprocessing.LabelEncoder() le.fit(df.iloc[:,4]) |
| --- |

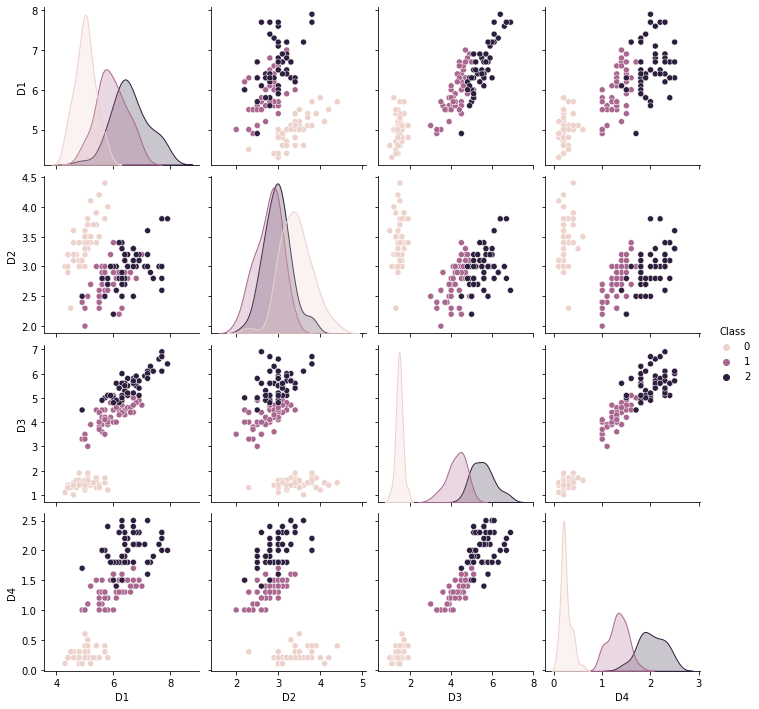
| LabelEncoder() |
| --- |

| df['Class'] = le.transform(df['Class']) df |
| --- |

| **D1** | **D2** | **D3** | **D4** | **Class** |  |
| --- | --- | --- | --- | --- | --- |
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | 0 |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | 0 |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | 0 |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | 0 |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | 0 |
| **...** | ... | ... | ... | ... | ... |
| **145** | 6.7 | 3.0 | 5.2 | 2.3 | 2 |
| **146** | 6.3 | 2.5 | 5.0 | 1.9 | 2 |
| **147** | 6.5 | 3.0 | 5.2 | 2.0 | 2 |
| **148** | 6.2 | 3.4 | 5.4 | 2.3 | 2 |
| **149** | 5.9 | 3.0 | 5.1 | 1.8 | 2 |

150 rows × 5 columns

| **df[:,].to\_csv('iris-encoded.csv', header=None) sns.pairplot(df, hue='Class')** |
| --- |

****

| **!sudo apt-get install libgfortran3** |
| --- |

| Reading package lists... Done Building dependency tree  Reading state information... Done The following package was automatically installed and is no longer required:  libnvidia-common-460 Use 'sudo apt autoremove' to remove it. The following additional packages will be installed:  gcc-6-base The following NEW packages will be installed:  gcc-6-base libgfortran3 0 upgraded, 2 newly installed, 0 to remove and 45 not upgraded. Need to get 286 kB of archives. After this operation, 1,394 kB of additional disk space will be used. Get:1 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 gcc-6-base amd64 6.5.0-2ubuntu1~18.04 [16.7 kB] Get:2 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 libgfortran3 amd64 6.5.0-2ubuntu1~18.04 [270 kB] Fetched 286 kB in 1s (290 kB/s) debconf: unable to initialize frontend: Dialog debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76, <> line 2.) debconf: falling back to frontend: Readline debconf: unable to initialize frontend: Readline debconf: (This frontend requires a controlling tty.) debconf: falling back to frontend: Teletype dpkg-preconfigure: unable to re-open stdin:  Selecting previously unselected package gcc-6-base:amd64. (Reading database ... 155975 files and directories currently installed.) Preparing to unpack .../gcc-6-base\_6.5.0-2ubuntu1~18.04\_amd64.deb ... Unpacking gcc-6-base:amd64 (6.5.0-2ubuntu1~18.04) ... Selecting previously unselected package libgfortran3:amd64. Preparing to unpack .../libgfortran3\_6.5.0-2ubuntu1~18.04\_amd64.deb ... Unpacking libgfortran3:amd64 (6.5.0-2ubuntu1~18.04) ... Setting up gcc-6-base:amd64 (6.5.0-2ubuntu1~18.04) ... Setting up libgfortran3:amd64 (6.5.0-2ubuntu1~18.04) ... Processing triggers for libc-bin (2.27-3ubuntu1.3) ... /sbin/ldconfig.real: /usr/local/lib/python3.7/dist-packages/ideep4py/lib/libmkldnn.so.0 is not a symbolic link |
| --- |

| **sc = spark.sparkContext  from pyspark.mllib.clustering import KMeans, GaussianMixture from numpy import array from math import sqrt  # Load and parse the data data = sc.textFile("iris-encoded.csv") parsedData = data.map(lambda line: array([float(x) for x in line.split(',') if x]))** |
| --- |

| **clusters = KMeans.train(parsedData, 2, maxIterations=10)  def error(point):  center = clusters.centers[clusters.predict(point)]  return sqrt(sum([x\*\*2 for x in (point - center)]))  WSSSE = parsedData.map(lambda point: error(point)).reduce(lambda x, y: x + y) print("KMeans: Sum of Squared Error = " + str(WSSSE))** |
| --- |

| **KMeans: Sum of Squared Error = 2828.724531888541** |
| --- |

| **clusters2 = GaussianMixture.train(parsedData, 1, maxIterations=10)  def error2(point):  center = clusters2.predict(point)  return sqrt(sum([x\*\*2 for x in (point - center)]))  WSSSE = parsedData.map(lambda point: error2(point)).reduce(lambda x, y: x + y) print("Gaussian Mixture: Sum of Squared Error = " + str(WSSSE))** |
| --- |

| **Gaussian Mixture: Sum of Squared Error = 3165.92681973472** |
| --- |

**CONCLUSION:** We successfully trained two Machine Learning models, one using KMeans algorithm and the other using Gaussian Mixture algorithm and KMeans had a lower Sum of Squared error therefore it is the better of the two algorithms.