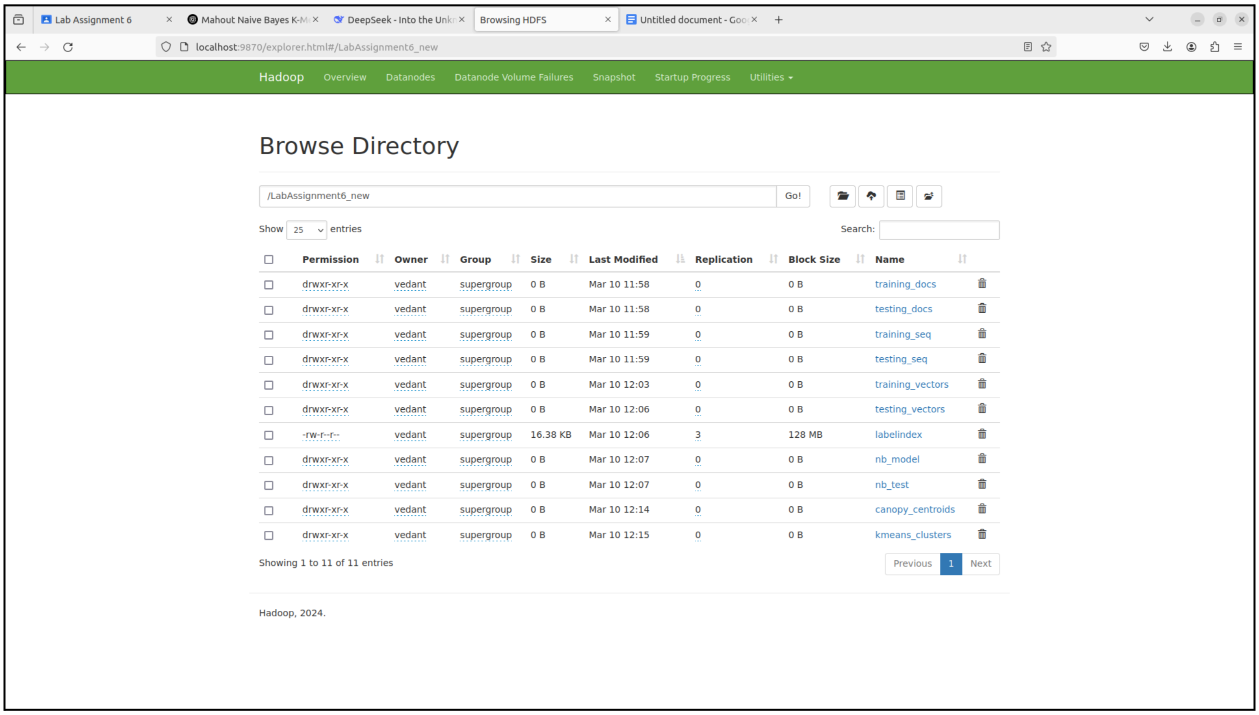
Lab Task 6

Name : Vedant Rahane
Roll No : MDE2024002

Hadoop Directory with all files used to implement task-



Browse Directory

/LabAssignment6_new Go!

Show 25 entries

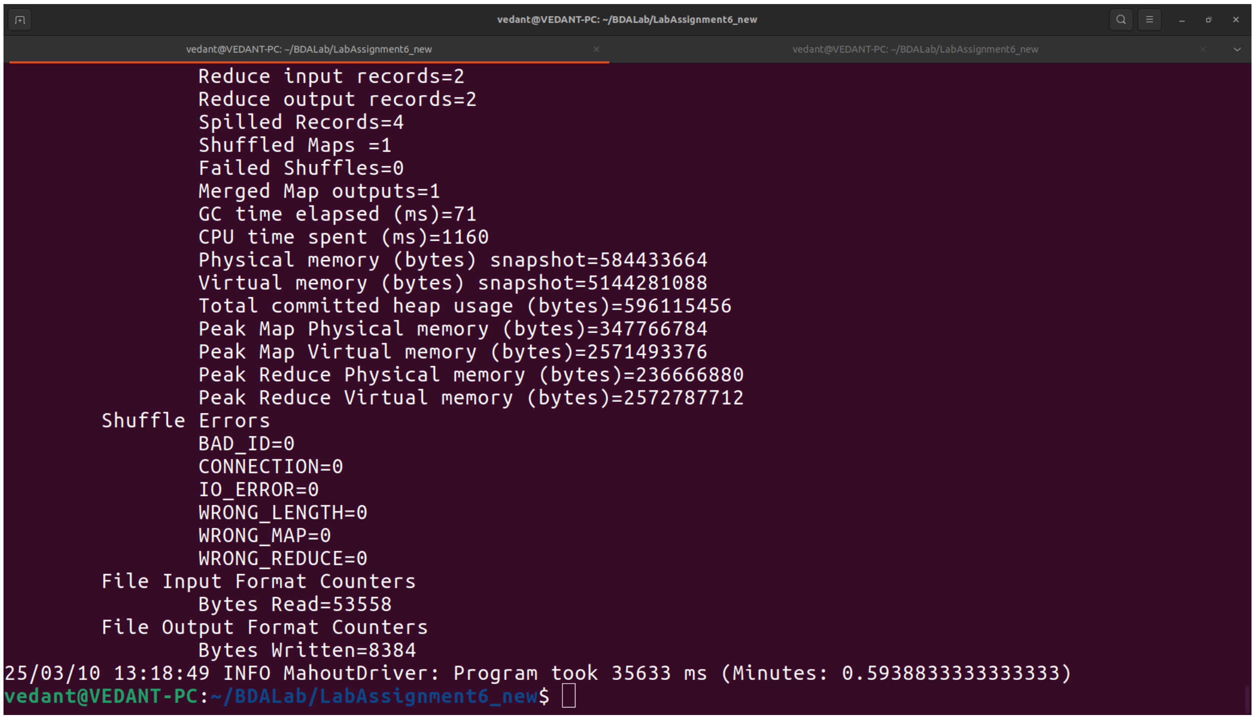
Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 11:58	0	0 B	training_docs
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 11:58	0	0 B	testing_docs
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 11:59	0	0 B	training_seq
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 11:59	0	0 B	testing_seq
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:03	0	0 B	training_vectors
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:06	0	0 B	testing_vectors
□	-rw-r--r--	vedant	supergroup	16.38 KB	Mar 10 12:06	3	128 MB	labelindex
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:07	0	0 B	nb_model
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:07	0	0 B	nb_test
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:14	0	0 B	canopy_centroids
□	drwxr-xr-x	vedant	supergroup	0 B	Mar 10 12:15	0	0 B	kmeans_clusters

Showing 1 to 11 of 11 entries

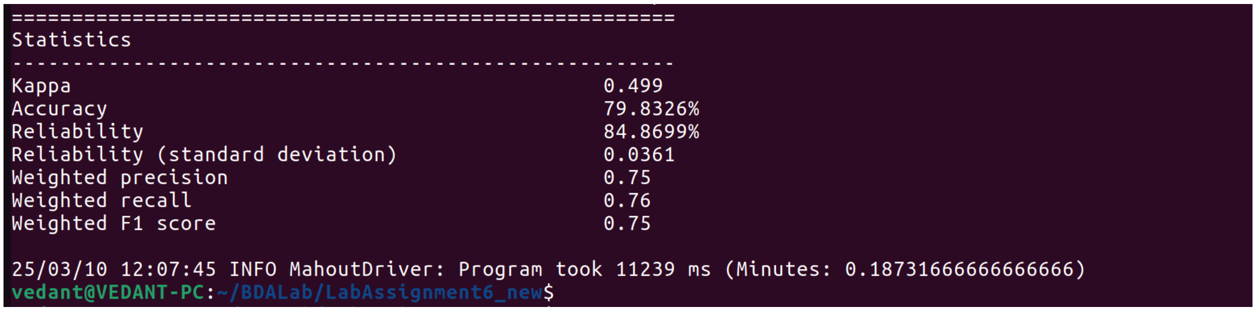
Hadoop, 2024.

Naive Bayes Classifier Training:


vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new

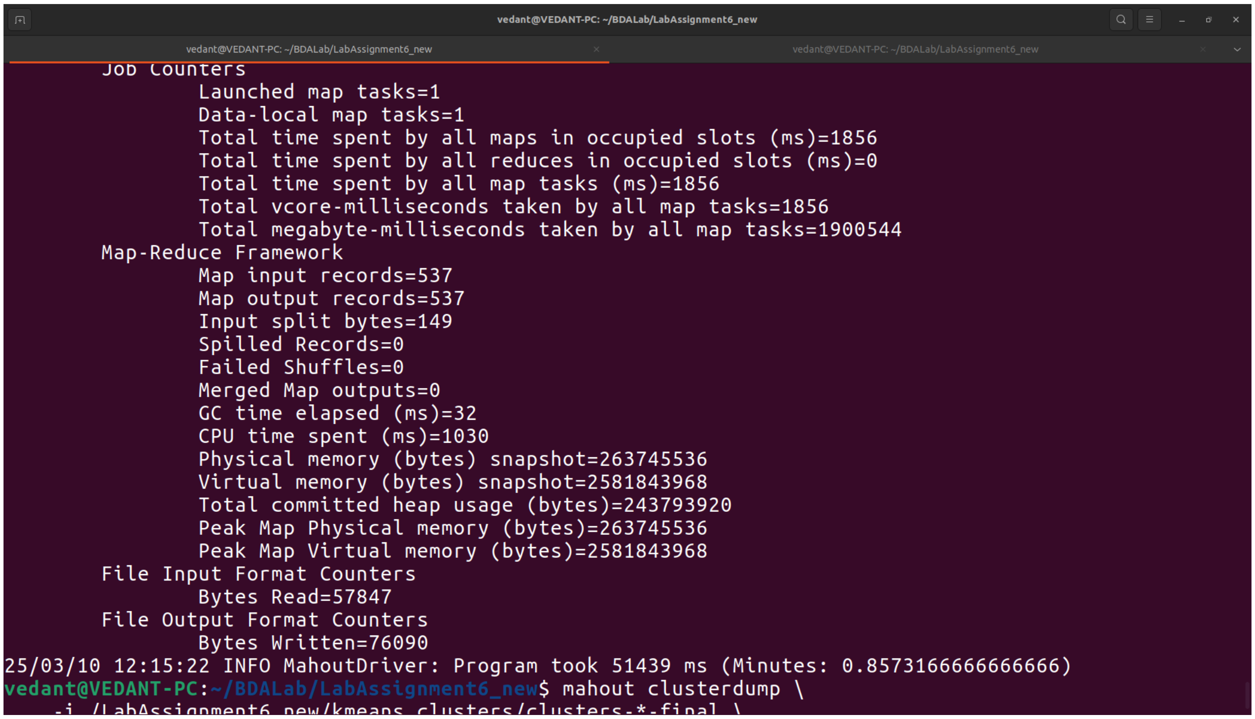
```
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=71
                CPU time spent (ms)=1160
                Physical memory (bytes) snapshot=584433664
                Virtual memory (bytes) snapshot=5144281088
                Total committed heap usage (bytes)=596115456
                Peak Map Physical memory (bytes)=347766784
                Peak Map Virtual memory (bytes)=2571493376
                Peak Reduce Physical memory (bytes)=236666880
                Peak Reduce Virtual memory (bytes)=2572787712
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=53558
        File Output Format Counters
                Bytes Written=8384
25/03/10 13:18:49 INFO MahoutDriver: Program took 35633 ms (Minutes: 0.5938833333333333)
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ 
```

Naive Bayes Classifier Output:



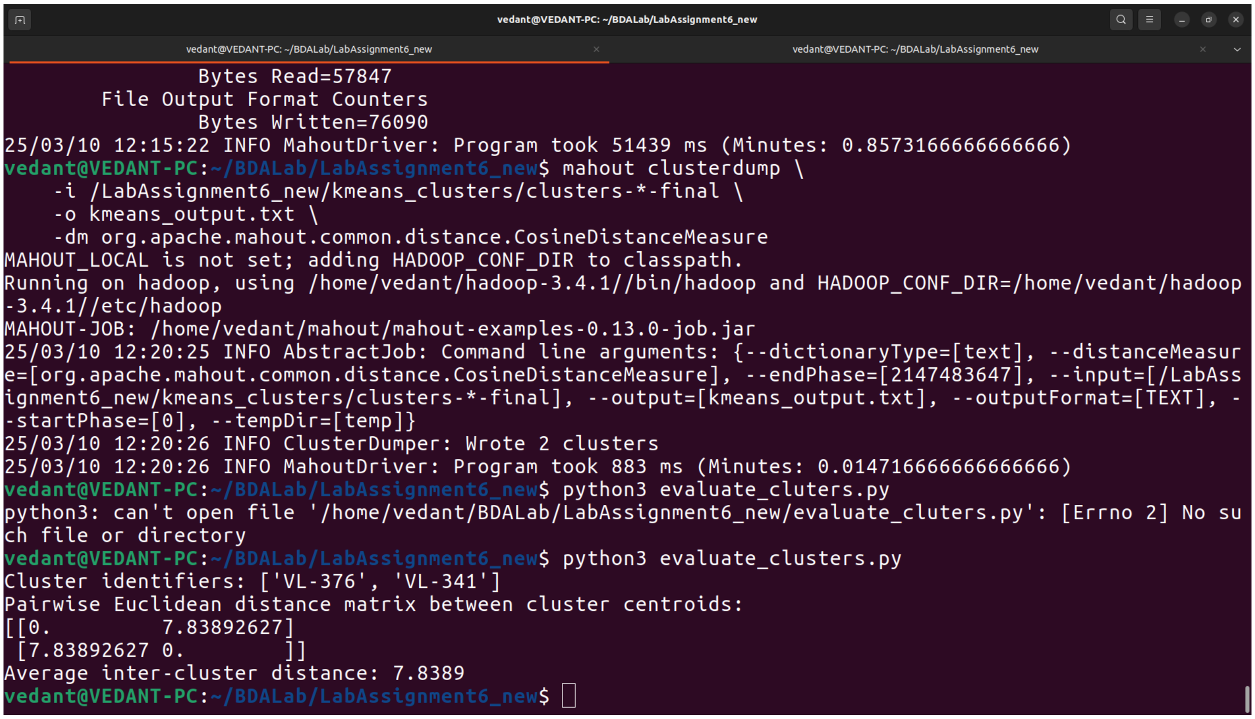
```
=======================================================
Statistics
-------------------------------------------------------
Kappa                                            0.499
Accuracy                                         79.8326%
Reliability                                      84.8699%
Reliability (standard deviation)                 0.0361
Weighted precision                               0.75
Weighted recall                                  0.76
Weighted F1 score                                0.75

25/03/10 12:07:45 INFO MahoutDriver: Program took 11239 ms (Minutes: 0.18731666666666666)
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$
```

K-Means Clustering Training:


vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new

```
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1856
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=1856
                Total vcore-milliseconds taken by all map tasks=1856
                Total megabyte-milliseconds taken by all map tasks=1900544
        Map-Reduce Framework
                Map input records=537
                Map output records=537
                Input split bytes=149
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=32
                CPU time spent (ms)=1030
                Physical memory (bytes) snapshot=263745536
                Virtual memory (bytes) snapshot=2581843968
                Total committed heap usage (bytes)=243793920
                Peak Map Physical memory (bytes)=263745536
                Peak Map Virtual memory (bytes)=2581843968
        File Input Format Counters
                Bytes Read=57847
        File Output Format Counters
                Bytes Written=76090
25/03/10 12:15:22 INFO MahoutDriver: Program took 51439 ms (Minutes: 0.8573166666666666)
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ mahout clusterdump \
    -i /LabAssignment6_new/kmeans_clusters/clusters-*-final \
```

K-Means Clustering Output:


vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new
vedant@VEDANT-PC: ~/BDALab/LabAssignment6_new

```
                Bytes Read=57847
        File Output Format Counters
                Bytes Written=76090
25/03/10 12:15:22 INFO MahoutDriver: Program took 51439 ms (Minutes: 0.8573166666666666)
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ mahout clusterdump \
    -i /LabAssignment6_new/kmeans_clusters/clusters-*-final \
    -o kmeans_output.txt \
    -dm org.apache.mahout.common.distance.CosineDistanceMeasure
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /home/vedant/hadoop-3.4.1//bin/hadoop and HADOOP_CONF_DIR=/home/vedant/hadoop
-3.4.1//etc/hadoop
MAHOUT-JOB: /home/vedant/mahout/mahout-examples-0.13.0-job.jar
25/03/10 12:20:25 INFO AbstractJob: Command line arguments: {--dictionaryType=[text], --distanceMeasur
e=[org.apache.mahout.common.distance.CosineDistanceMeasure], --endPhase=[2147483647], --input=[/LabAss
ignment6_new/kmeans_clusters/clusters-*-final], --output=[kmeans_output.txt], --outputFormat=[TEXT], -
-startPhase=[0], --tempDir=[temp]}
25/03/10 12:20:26 INFO ClusterDumper: Wrote 2 clusters
25/03/10 12:20:26 INFO MahoutDriver: Program took 883 ms (Minutes: 0.014716666666666666)
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ python3 evaluate_cluters.py
python3: can't open file '/home/vedant/BDALab/LabAssignment6_new/evaluate_cluters.py': [Errno 2] No su
ch file or directory
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ python3 evaluate_clusters.py
Cluster identifiers: ['VL-376', 'VL-341']
Pairwise Euclidean distance matrix between cluster centroids:
[[0.         7.83892627]
 [7.83892627 0.        ]]
Average inter-cluster distance: 7.8389
vedant@VEDANT-PC:~/BDALab/LabAssignment6_new$ 
```

Summary of Steps Performed

1. Data Preparation

- **Dataset Splitting:**
  - Downloaded `diabetes.csv` and split it into **training (80%)** and **testing (20%)** datasets.
  - Converted CSV into **Hadoop SequenceFiles** using `mahout seqdirectory`.
- **Sequence File Processing:**
  - Applied `mahout seq2sparse` to convert SequenceFiles into **TF-IDF sparse vectors** for feature extraction.
  - Generated **TF-IDF vector representations** of both training and testing datasets.

2. Naive Bayes Classification

- **Training:** Used `mahout trainnb` to train a Naive Bayes model on the **TF-IDF training vectors**.
- **Testing:** Used `mahout testnb` to evaluate the trained model on the **test set**.
- **Results:** Achieved **79.83% accuracy, F1-score: 0.75**, indicating good generalization.

3. K-Means Clustering

- Applied `mahout kmeans` to cluster the data into groups.
- Used `mahout clusterdump` to analyze cluster characteristics.
- Evaluated clustering with **centroid distances (~7.83 separation)**.

Comparison of Classification & Clustering

- **Classification (Naive Bayes)**
  Uses the known Outcome labels (supervised) and achieves ~80% accuracy.
- **Clustering (K-Means)**
  Automatically groups data into two clusters (unsupervised). While the centroid distance shows they are reasonably distinct.