1 VM:

URL: spark://master:7077
Alive Workers: 1
Cores in use: 8 Total, 0 Used
Memory in use: 4.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20251109020828-192.168.13.200-34165	192.168.13.200:34165	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	

▾ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

▾ Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20251109122752-0000	sat5165_stroke_pipeline	8	1024.0 MiB		2025/11/09 12:27:52	sat3812	FINISHED	54 min

```
[sat3812@master ~]$ $SPARK_HOME/sbin/start-master.sh
$SPARK_HOME/sbin/start-worker.sh spark://master:7077
org.apache.spark.deploy.master.Master running as process 128448.  Stop it first.
org.apache.spark.deploy.worker.Worker running as process 123127.  Stop it first.
[sat3812@master ~]$ sudo nano stokre_pipeline.py
[sudo] password for sat3812:
[sat3812@master ~]$ sudo nano stroke_pipeline.py
[sat3812@master ~]$ sudo nano stroke_pipeline.py
[sat3812@master ~]$ spark-submit --master spark://master:7077 stroke_pipeline.py
Loading: hdfs://master:9000/user/sat3812/stroke_project/data/healthcare-dataset-stroke-data.csv
Oversampled minority from 191 to ~4011 vs majority 3932.
[info] MLP input_dim = 20

Training Logistic Regression...
Total time: 242.91 seconds

Logistic Regression Performance Results
  accuracy: 0.727
  auc: 0.861
  precision: 0.154
  recall: 0.810
  specificity: 0.722

Training Linear SVM...
Total time: 1065.26 seconds

Linear SVM Performance Results
  accuracy: 0.719
  auc: 0.861
  precision: 0.155
  recall: 0.845
  specificity: 0.712

Training Gradient Boosted Trees...
Total time: 1545.94 seconds

Gradient Boosted Trees Performance Results
  accuracy: 0.819
  auc: 0.802
  precision: 0.162
  recall: 0.500
  specificity: 0.839

Training Multilayer Perceptron...
Total time: 258.94 seconds

Multilayer Perceptron Performance Results
  accuracy: 0.781
  auc: 0.697
  precision: 0.097
  recall: 0.328
  specificity: 0.809
MODEL COMPARISON SUMMARY
-------------------------------------------------------------------------
Model                      Acc    AUC    Prec   Recall   Spec   Time(s)
-------------------------------------------------------------------------
Logistic Regression        0.727  0.861  0.154  0.810    0.722   242.91
Linear SVM                 0.719  0.861  0.155  0.845    0.712  1065.26
Gradient Boosted Trees     0.819  0.802  0.162  0.500    0.839  1545.94
Multilayer Perceptron      0.781  0.697  0.097  0.328    0.809   258.94
[main] done.
[sat3812@master ~]$
```

2 VMS:

URL: spark://master:7077
Alive Workers: 2
Cores in use: 16 Total, 0 Used
Memory in use: 8.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20251109020828-192.168.13.200-34165	192.168.13.200:34165	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20251109134934-192.168.13.201-37849	192.168.13.201:37849	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	

▾ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

▾ Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20251109135056-0000	sat5165_stroke_pipeline	16	1024.0 MiB		2025/11/09 13:50:56	sat3812	FINISHED	37 min

```
[sat3812@master ~]$ spark-submit --master spark://master:7077 stroke_pipeline.py
Loading: hdfs://master:9000/user/sat3812/stroke_project/data/healthcare-dataset-stroke-data.csv
Oversampled minority from 191 to ~4011 vs majority 3932.
[info] MLP input_dim = 20

Training Logistic Regression...
Total time: 193.03 seconds

Logistic Regression Performance Results
  accuracy: 0.727
  auc: 0.861
  precision: 0.154
  recall: 0.810
  specificity: 0.722

Training Linear SVM...
Total time: 728.84 seconds

Linear SVM Performance Results
  accuracy: 0.719
  auc: 0.861
  precision: 0.155
  recall: 0.845
  specificity: 0.712

Training Gradient Boosted Trees...
Total time: 1039.26 seconds

Gradient Boosted Trees Performance Results
  accuracy: 0.819
  auc: 0.802
  precision: 0.162
  recall: 0.500
  specificity: 0.839

Training Multilayer Perceptron...
Total time: 179.01 seconds

Multilayer Perceptron Performance Results
  accuracy: 0.781
  auc: 0.697
  precision: 0.097
  recall: 0.328
  specificity: 0.809
MODEL COMPARISON SUMMARY
--------------------------------------------------------------------
Model                      Acc     AUC    Prec   Recall   Spec   Time(s)
--------------------------------------------------------------------
Logistic Regression        0.727   0.861  0.154  0.810    0.722  193.03
Linear SVM                 0.719   0.861  0.155  0.845    0.712  728.84
Gradient Boosted Trees     0.819   0.802  0.162  0.500    0.839  1039.26
Multilayer Perceptron      0.781   0.697  0.097  0.328    0.809  179.01
[main] done.
[sat3812@master ~]$
```

3 VMS:

```
Loading: hdfs://master:9000/user/sat3812/stroke_project/data/healthcare-dataset-stroke-c
Oversampled minority from 191 to ~4011 vs majority 3932.
[info] MLP input_dim = 20

Training Logistic Regression...
Total time: 203.43 seconds

Logistic Regression Performance Results
  accuracy: 0.727
  auc: 0.861
  precision: 0.154
  recall: 0.810
  specificity: 0.722

Training Linear SVM...
Total time: 697.93 seconds

Linear SVM Performance Results
  accuracy: 0.719
  auc: 0.861
  precision: 0.155
  recall: 0.845
  specificity: 0.712

Training Gradient Boosted Trees...
Total time: 916.42 seconds

Gradient Boosted Trees Performance Results
  accuracy: 0.819
  auc: 0.803
  precision: 0.162
  recall: 0.500
  specificity: 0.839

Training Multilayer Perceptron...
Total time: 181.76 seconds

Multilayer Perceptron Performance Results
  accuracy: 0.791
  auc: 0.696
  precision: 0.102
  recall: 0.328
  specificity: 0.820
MODEL COMPARISON SUMMARY
-----------------------------------------------------------------------
Model                      Acc     AUC     Prec    Recall   Spec   Time(s)
-----------------------------------------------------------------------
Logistic Regression        0.727   0.861   0.154   0.810   0.722   203.43
Linear SVM                 0.719   0.861   0.155   0.845   0.712   697.93
Gradient Boosted Trees     0.819   0.803   0.162   0.500   0.839   916.42
Multilayer Perceptron      0.791   0.696   0.102   0.328   0.820   181.76
Stroke Pipline Complete.
```

4 VMS:

▾ Workers (4)

Worker Id	Address	State	Cores	Memory	Resources
worker-20251109192507-192.168.13.201-33947	192.168.13.201:33947	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20251109192515-192.168.13.122-43307	192.168.13.122:43307	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20251109192515-192.168.13.123-42317	192.168.13.123:42317	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20251109192516-192.168.13.200-46823	192.168.13.200:46823	ALIVE	8 (0 Used)	4.0 GiB (0.0 B Used)	

▾ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

▾ Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20251109193815-0000	sat5165_stroke_pipeline	32	3.0 GiB		2025/11/09 19:38:15	sat3812	FINISHED	27 min

```
Loading: hdfs://master:9000/user/sat3812/stroke_project/data/healthcare-dataset-stroke-data.c
Oversampled minority from 191 to ~4011 vs majority 3932.
[info] MLP input_dim = 20

Training Logistic Regression...
Total time: 164.51 seconds

Logistic Regression Performance Results
  accuracy: 0.727
  auc: 0.861
  precision: 0.154
  recall: 0.810
  specificity: 0.722

Training Linear SVM...
Total time: 548.78 seconds

Linear SVM Performance Results
  accuracy: 0.719
  auc: 0.861
  precision: 0.155
  recall: 0.845
  specificity: 0.712

Training Gradient Boosted Trees...
Total time: 690.43 seconds

Gradient Boosted Trees Performance Results
  accuracy: 0.819
  auc: 0.802
  precision: 0.162
  recall: 0.500
  specificity: 0.839

Training Multilayer Perceptron...
Total time: 132.34 seconds

Multilayer Perceptron Performance Results
  accuracy: 0.779
  auc: 0.698
  precision: 0.096
  recall: 0.328
  specificity: 0.807
MODEL COMPARISON SUMMARY
-------------------------------------------------------------------------------
Model                      Acc     AUC     Prec   Recall    Spec   Time(s)
-------------------------------------------------------------------------------
Logistic Regression        0.727   0.861   0.154   0.810   0.722   164.51
Linear SVM                 0.719   0.861   0.155   0.845   0.712   548.78
Gradient Boosted Trees     0.819   0.802   0.162   0.500   0.839   690.43
Multilayer Perceptron      0.779   0.698   0.096   0.328   0.807   132.34
Stroke Pipline Complete.
[sat3812@master ~]$
```