In [ ]:

```
#Ashiqur RahmanKhan
#online ApacheSpark installation

!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.6.tgz
!tar xvf spark-2.4.5-bin-hadoop2.6.tgz
!pip install -q findspark
```

In [0]:

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.5-bin-hadoop2.6"

import findspark
findspark.init()

from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

from pyspark import SparkContext, SparkConf
from pyspark.sql import SQLContext, SparkSession
from pyspark.sql.types import StructType, StructField, DoubleType, IntegerType, StringType
sc = SparkContext.getOrCreate(SparkConf().setMaster("local[*]"))
from pyspark.sql import SparkSession
spark = SparkSession \
    .builder \
    .getOrCreate()
```

In [3]:

```
!wget https://github.com/shatiilrahman/Machine-Learning/blob/master/titanic.parquet?raw=true
!mv titanic.parquet?raw=true titanic.parquet
```

```
--2020-05-11 21:06:06--  https://github.com/shatiilrahman/Machine-Learning/blob/master/titanic.parquet?raw=true
Resolving github.com (github.com)... 140.82.113.3
Connecting to github.com (github.com)|140.82.113.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://github.com/shatiilrahman/Machine-Learning/raw/master/titanic.parquet [following]
--2020-05-11 21:06:06--  https://github.com/shatiilrahman/Machine-Learning/raw/master/titanic.parquet
Reusing existing connection to github.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/shatiilrahman/Machine-Learning/master/titanic.parquet [following]
--2020-05-11 21:06:06--  https://raw.githubusercontent.com/shatiilrahman/Machine-Learning/master/titanic.parquet
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.0.133, 151.101.64.133, 151.101.128.133, ...
```

```
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.0.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 39334 (38K) [application/octet-stream]
Saving to: ‘titanic.parquet?raw=true’

titanic.parquet?raw 100%[===================>]  38.41K  --.-KB/s    in 0.01s

2020-05-11 21:06:07 (3.16 MB/s) - ‘titanic.parquet?raw=true’ saved [39334/39334]
```

In [0]:

```python
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer
from pyspark.ml.feature import OneHotEncoder
from pyspark.ml.linalg import Vector
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.feature import Normalizer
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.classification import OneVsRest
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

In [0]:

```python
#LogisticRegression with OneVsRest
```

In [0]:

```python
#loading dataset
d=0
d = spark.read.parquet('/content/titanic.parquet')
d.createOrReplaceTempView("titanic")
(d_train,d_test) = d.randomSplit([0.7, 0.3])
```

In [18]:

```python
d.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|          9|     1.0|   3.0|Johnson, Mrs. Osc...|female|27.0|  0.0|  2.0|          347742|11.1333| null|       S|
|         64|     0.0|   3.0|Skoog, Master. Ha...|  male| 4.0|  3.0|  2.0|          347088|   27.9| null|       S|
|        168|     0.0|   3.0|Skoog, Mrs. Willi...|female|45.0|  1.0|  4.0|          347088|   27.9| null|       S|
|        228|     0.0|   3.0|"Lovell, Mr. John...|  male|20.5|  0.0|  0.0|       A/5 21173|   7.25| null|       S|
|        322|     0.0|   3.0|    Danoff, Mr. Yoto|  male|27.0|  0.0|  0.0|          349219| 7.8958| null|       S|
|        440|     0.0|   2.0|Kvillner, Mr. Joh...|  male|31.0|  0.0|  0.0|      C.A. 18723|   10.5| null|       S|
|          2|     1.0|   1.0|Cumings, Mrs. Joh...|female|38.0|  1.0|  0.0|        PC 17599|71.2833|  C85|       C|
|         22|     1.0|   2.0|Beesley, Mr. Lawr...|  male|34.0|  0.0|  0.0|         248698|   13.0|  D56|       S|
|        138|     0.0|   1.0|Futrelle, Mr. Jac...|  male|37.0|  1.0|  0.0|         113803|   53.1| C123|       S|
|        541|     1.0|   1.0|Crosby, Miss. Har...|female|36.0|  0.0|  2.0|      WE/P 5735|   71.0|  B22|       S|
|        652|     1.0|   2.0| Doling, Miss. Elsie|female|18.0|  0.0|  1.0|         231919|   23.0| null|       S|
|        677|     0.0|   3.0|Sawyer, Mr. Frede...|  male|24.5|  0.0|  0.0|         342826|   8.05| null|       S|
|        828|     1.0|   2.0|Mallet, Master. A...|  male| 1.0|  0.0|  2.0|S.C./PARIS 2079|37.0042| null|       C|
|        883|     0.0|   3.0|Dahlberg, Miss. G...|female|22.0|  0.0|  0.0|           7552|10.5167| null|       S|
|        185|     1.0|   3.0|Kink-Heilmann, Mi...|female| 4.0|  0.0|  2.0|         315153| 22.025| null|       S|
|        523|     0.0|   3.0|  Lahoud, Mr. Sarkis|  male|null|  0.0|  0.0|           2624|  7.225| null|       C|
|        692|     1.0|   3.0|  Karun, Miss. Manca|female| 4.0|  0.0|  1.0|         349256|13.4167| null|       C|
|        779|     0.0|   3.0|Kilgannon, Mr. Th...|  male|null|  0.0|  0.0|          36865| 7.7375| null|       Q|
|        819|     0.0|   3.0|Holm, Mr. John Fr...|  male|43.0|  0.0|  0.0|         C 7075|   6.45| null|       S|
|        838|     0.0|   3.0| Sirota, Mr. Maurice|  male|null|  0.0|  0.0|         392092|   8.05| null|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+---------------+-------+-----+--------+
only showing top 20 rows
```

In [0]:

```
#pipeline formation
indexer = StringIndexer(inputCol = "Sex" , outputCol = "label")
vectorAssembler = VectorAssembler(inputCols=["Survived","Pclass","SibSp","Parch","Fare"],outputCol="features")
normalizer = Normalizer(inputCol="features", outputCol="features_norm",p=1.0)
classifier = LogisticRegression(maxIter = 20, regParam = 0.3, elasticNetParam = 0.8)
ovr = OneVsRest(classifier=classifier)
pipelineLo = Pipeline(stages=[indexer, vectorAssembler, normalizer,ovr])
```

In [9]:

```
#model_fit
modelLo = pipelineLo.fit(d_train)
prediction = modelLo.transform(d_test)
eval = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="accuracy")
eval.evaluate(prediction)
accuracyLo = eval.evaluate(prediction)
print("LogisticRegressionOneVsRest Accuracy : ",accuracyLo)
```

```
LogisticRegressionOneVsRest Accuracy :  0.6752767527675276
```

In [0]:

```
#RandomForestClassifier
```

In [0]:

```
#loading dataset
df_temp = prediction
data = df_temp.drop("prediction")
(trainingData, testData) = data.randomSplit([0.7, 0.3])
```

In [17]:

```
data.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+---------------+-------+-----+--------+-----+-----------
---------+--------------------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|         Ticket|   Fare|Cabin|Embarked|label|
features|       features_norm|
+-----------+--------+------+--------------------+------+----+-----+-----+---------------+-------+-----+--------+-----+-----------
---------+--------------------+
|        106|     0.0|   3.0|Mionoff, Mr. Stoy...|  male|28.0|  0.0|  0.0|         349207| 7.8958| null|       S|  0.0|(5,[1,4],[3
.0,7.8...|(5,[1,4],[0.27533...|
|        108|     1.0|   3.0|Moss, Mr. Albert ...|  male|null|  0.0|  0.0|         312991|  7.775| null|       S|  0.0|[1.0,3.0,0
.0,0.0,...|[0.08492569002123...|
|        125|     0.0|   1.0|White, Mr. Perciv...|  male|54.0|  0.0|  1.0|          35281|77.2875|  D26|       S|  0.0|[0.0,1.0,0
.0,1.0,...|[0.0,0.0126123285...|
|        133|     0.0|   3.0|Robins, Mrs. Alex...|female|47.0|  1.0|  0.0|      A/5. 3337|   14.5| null|       S|  1.0|[0.0,3.0,1
.0,0.0,...|[0.0,0.1621621621...|
|        139|     0.0|   3.0| Osen, Mr. Olaf Elon|  male|16.0|  0.0|  0.0|           7534| 9.2167| null|       S|  0.0|(5,[1,4],[
3.0,9.2...|(5,[1,4],[0.24556...|
|        140|     0.0|   1.0|  Giglio, Mr. Victor|  male|24.0|  0.0|  0.0|       PC 17593|   79.2|  B86|       C|  0.0|(5,[1,4],[1
.0,79.2])|(5,[1,4],[0.01246...|
|        141|     0.0|   3.0|Boulos, Mrs. Jose...|female|null|  0.0|  2.0|           2678|15.2458| null|       C|  1.0|[0.0,3.0,0
.0,2.0,...|[0.0,0.1481788815...|
|        146|     0.0|   2.0|Nicholls, Mr. Jos...|  male|19.0|  1.0|  1.0|     C.A. 33112|  36.75| null|       S|  0.0|[0.0,2.0,1.
0,1.0,...|[0.0,0.0490797546...|
|         15|     0.0|   3.0|Vestrom, Miss. Hu...|female|14.0|  0.0|  0.0|         350406| 7.8542| null|       S|  1.0|(5,[1,4],[3
.0,7.8...|(5,[1,4],[0.27639...|
|        150|     0.0|   2.0|Byles, Rev. Thoma...|  male|42.0|  0.0|  0.0|         244310|   13.0| null|       S|  0.0|(5,[1,4],[
2.0,13.0])|(5,[1,4],[0.13333...|
|        151|     0.0|   2.0|Bateman, Rev. Rob...|  male|51.0|  0.0|  0.0|    S.O.P. 1166| 12.525| null|       S|  0.0|(5,[1,4],[2
.0,12....|(5,[1,4],[0.13769...|
|        156|     0.0|   1.0|Williams, Mr. Cha...|  male|51.0|  0.0|  1.0|       PC 17597|61.3792| null|       C|  0.0|[0.0,1.0,0.
0,1.0,...|[0.0,0.0157780470...|
|        158|     0.0|   3.0|     Corn, Mr. Harry|  male|30.0|  0.0|  0.0|SOTON/OQ 392090|   8.05| null|       S|  0.0|(5,[1,4],[3
.0,8.05])|(5,[1,4],[0.27149...|
|         16|     1.0|   2.0|Hewlett, Mrs. (Ma...|female|55.0|  0.0|  0.0|         248706|   16.0| null|       S|  1.0|[1.0,2.0,0.
0,0.0,...|[0.05263157894736...|
|        161|     0.0|   3.0|Cribb, Mr. John H...|  male|44.0|  0.0|  1.0|         371362|   16.1| null|       S|  0.0|[0.0,3.0,0
.0,1.0,...|[0.0,0.1492537313...|
|        163|     0.0|   3.0|Bengtsson, Mr. Jo...|  male|26.0|  0.0|  0.0|         347068|  7.775| null|       S|  0.0|(5,[1,4],[
3.0,7.7...|(5,[1,4],[0.27842...|
|        172|     0.0|   3.0|Rice, Master. Arthur|  male| 4.0|  4.0|  1.0|         382652| 29.125| null|       Q|  0.0|[0.0,3.0,4.
0,1.0,...|[0.0,0.0808080808...|
|        173|     1.0|   3.0|Johnson, Miss. El...|female| 1.0|  1.0|  1.0|         347742|11.1333| null|       S|  1.0|[1.0,3.0,1.
0,1.0,...|[0.05836587230714...|
|        175|     0.0|   1.0|Smith, Mr. James ...|  male|56.0|  0.0|  0.0|          17764|30.6958|   A7|       C|  0.0|(5,[1,4],[
1.0,30....|(5,[1,4],[0.03154...|
|        176|     0.0|   3.0|Klasen, Mr. Klas ...|  male|18.0|  1.0|  1.0|         350404| 7.8542| null|       S|  0.0|[0.0,3.0,1.
0,1.0,...|[0.0,0.2333867529...|
+-----------+--------+------+--------------------+------+----+-----+-----+---------------+-------+-----+--------+-----+-----------
```

```
------------+--------------------+
only showing top 20 rows
```

In [0]:

```
#pipeline formation
labelIndexer = StringIndexer(inputCol="label", outputCol="indexedLabel").fit(data)
featureIndexer =\
    VectorIndexer(inputCol="features", outputCol="indexedFeatures", maxCategories=4).fit(data)
rf = RandomForestClassifier(labelCol="indexedLabel", featuresCol="indexedFeatures", numTrees=10)
pipelineRa = Pipeline(stages=[labelIndexer, featureIndexer, rf])
```

In [0]:

```
#model_fit
modelRa = pipelineRa.fit(trainingData)
prediction = modelRa.transform(testData)
```

In [13]:

```
evalua = MulticlassClassificationEvaluator(labelCol="indexedLabel", predictionCol="prediction", metricName="accuracy")
evalua.evaluate(prediction)
accuracyRa = evalua.evaluate(prediction)
print("RandomForestClassifier Accuracy : ",accuracyRa)
```

```
RandomForestClassifier Accuracy :  0.7204301075268817
```