Sheet

```python
from pyspark.sql import SparkSession
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.classification import  RandomForestClassifier
from pyspark.ml.evaluation import BinaryClassificationEvaluator

spark = SparkSession.builder.getOrCreate()

marksDF = spark.read.csv("teach_scores_1.csv",header = True, inferSchema = True)

marksDF.printSchema()

"""
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)
"""
```

```
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)



'\nroot\n |-- subject_1_gp: double (nullable = true)\n |-- subject_2_gp: double (nullable = true)\n |-- subject_3_gp: double (nullab
```

```python
marksDF.show(10,False)

"""
+------------+------------+------------+------------+------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|
+------------+------------+------------+------------+------------+-----+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |
+------------+------------+------------+------------+------------+-----+
"""
```

```
+------------+------------+------------+------------+------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|
+------------+------------+------------+------------+------------+-----+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |
+------------+------------+------------+------------+------------+-----+
only showing top 10 rows



'\n+------------+------------+------------+------------+------------+-----+\n|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|su
```

```
marksDF.describe("subject_1_gp").show()

"""
+-------+------------------+
|summary|      subject_1_gp|
+-------+------------------+
|  count|               172|
|   mean| 6.663953488372093|
| stddev|2.1988786504628766|
|    min|               0.0|
|    max|               9.9|
+-------+------------------+
"""
```

```
+-------+------------------+
|summary|      subject_1_gp|
+-------+------------------+
|  count|               172|
|   mean| 6.663953488372093|
| stddev|2.1988786504628766|
|    min|               0.0|
|    max|               9.9|
+-------+------------------+



'\n+-------+------------------+\n|summary|      subject_1_gp|\n+-------+------------------+\n|  count|               172|\n|   mean|
```

```
inputCols = ["subject_1_gp","subject_2_gp","subject_3_gp","subject_4_gp","subject_5_gp"]

outputCol = "features"

marksDF_assembler = VectorAssembler(inputCols = inputCols,outputCol = outputCol)

featuresDf = marksDF_assembler.transform(marksDF)

print("featuresDF printSchema")

featuresDf.printSchema()

"""
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)
 |-- features: vector (nullable = true)
"""
```

```
featuresDF printSchema
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)
 |-- features: vector (nullable = true)



'\nroot\n |-- subject_1_gp: double (nullable = true)\n |-- subject_2_gp: double (nullable = true)\n |-- subject_3_gp: double (nullab
featuresDf.show(10,False)

print("featureDf show")

"""
+------------+------------+------------+------------+------------+-----+---------------------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |
+------------+------------+------------+------------+------------+-----+---------------------+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |[9.0,9.0,9.0,9.0,9.0]|
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |[8.0,8.0,8.0,8.0,8.0]|
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |[7.0,7.0,7.0,7.0,7.0]|
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |[6.0,6.0,6.0,6.0,6.0]|
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |[5.0,5.0,5.0,5.0,5.0]|
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|
+------------+------------+------------+------------+------------+-----+---------------------+
"""
```

```
+------------+------------+------------+------------+------------+-----+---------------------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |
+------------+------------+------------+------------+------------+-----+---------------------+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |[9.0,9.0,9.0,9.0,9.0]|
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |[8.0,8.0,8.0,8.0,8.0]|
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |[7.0,7.0,7.0,7.0,7.0]|
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |[6.0,6.0,6.0,6.0,6.0]|
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |[5.0,5.0,5.0,5.0,5.0]|
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|
+------------+------------+------------+------------+------------+-----+---------------------+
only showing top 10 rows

featureDf show


'\n+------------+------------+------------+------------+------------+-----+---------------------+\n|subject_1_gp|subject_2_gp|subjec
```

```
grade_indexer = StringIndexer(inputCol = "grade", outputCol = "label")

label_df = grade_indexer.fit(featuresDf).transform(featuresDf)

print("after adding label")

label_df.printSchema()

"""
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)
 |-- features: vector (nullable = true)
 |-- label: double (nullable = false)
"""
```

```
after adding label
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
 |-- grade: string (nullable = true)
 |-- features: vector (nullable = true)
 |-- label: double (nullable = false)
```

```
'\nroot\n |-- subject_1_gp: double (nullable = true)\n |-- subject_2_gp: double (nullable = true)\n |-- subject_3_gp: double (nullab
```

```
print("label included df")

label_df.createOrReplaceGlobalTempView("main_df")

label_df.show(10,False)

"""
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |[9.0,9.0,9.0,9.0,9.0]|6.0  |
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |[8.0,8.0,8.0,8.0,8.0]|1.0  |
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |[7.0,7.0,7.0,7.0,7.0]|2.0  |
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |[6.0,6.0,6.0,6.0,6.0]|3.0  |
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |[5.0,5.0,5.0,5.0,5.0]|4.0  |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |
+------------+------------+------------+------------+------------+-----+---------------------+-----+
"""
```

```
label included df
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|9.0         |9.0         |9.0         |9.0         |9.0         |A+   |[9.0,9.0,9.0,9.0,9.0]|6.0  |
|8.0         |8.0         |8.0         |8.0         |8.0         |A    |[8.0,8.0,8.0,8.0,8.0]|1.0  |
|7.0         |7.0         |7.0         |7.0         |7.0         |B    |[7.0,7.0,7.0,7.0,7.0]|2.0  |
|6.0         |6.0         |6.0         |6.0         |6.0         |C    |[6.0,6.0,6.0,6.0,6.0]|3.0  |
|5.0         |5.0         |5.0         |5.0         |5.0         |D    |[5.0,5.0,5.0,5.0,5.0]|4.0  |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |
+------------+------------+------------+------------+------------+-----+---------------------+-----+
only showing top 10 rows



'\n+------------+------------+------------+------------+------------+-----+---------------------+-----+\n|subject_1_gp|subject_2_gp|
```

```
trainingData,testdata = label_df.randomSplit([0.7,0.3],seed = 42)

print("display training data")

trainingData.show(10,False)
"""
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|0.0         |0.0         |0.0         |0.0         |0.0         |F    |(5,[],[])            |0.0  |
|0.0         |0.0         |0.0         |0.0         |0.0         |F    |(5,[],[])            |0.0  |
|1.0         |1.0         |1.0         |1.0         |1.0         |F    |[1.0,1.0,1.0,1.0,1.0]|0.0  |
|2.0         |2.0         |2.0         |2.0         |2.0         |F    |[2.0,2.0,2.0,2.0,2.0]|0.0  |
|2.0         |2.0         |2.0         |2.0         |2.0         |F    |[2.0,2.0,2.0,2.0,2.0]|0.0  |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |
+------------+------------+------------+------------+------------+-----+---------------------+-----+
"""
```

```
display training data
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|
+------------+------------+------------+------------+------------+-----+---------------------+-----+
|0.0         |0.0         |0.0         |0.0         |0.0         |F    |(5,[],[])            |0.0  |
|0.0         |0.0         |0.0         |0.0         |0.0         |F    |(5,[],[])            |0.0  |
|1.0         |1.0         |1.0         |1.0         |1.0         |F    |[1.0,1.0,1.0,1.0,1.0]|0.0  |
|2.0         |2.0         |2.0         |2.0         |2.0         |F    |[2.0,2.0,2.0,2.0,2.0]|0.0  |
|2.0         |2.0         |2.0         |2.0         |2.0         |F    |[2.0,2.0,2.0,2.0,2.0]|0.0  |
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|2.1         |2.0         |2.4         |3.5         |3.0         |F    |[2.1,2.0,2.4,3.5,3.0]|0.0  |
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |
+------------+------------+------------+------------+------------+-----+---------------------+-----+
only showing top 10 rows



'\n+------------+------------+------------+------------+------------+-----+---------------------+-----+\n|subject_1_gp|subject_2_gp|
```

```python
ran_for_regression = RandomForestClassifier().setImpurity("gini").setMaxDepth(10).setNumTrees(20).setFeatureSubsetStrategy("auto").set

ran_for_Model = ran_for_regression .fit(trainingData)

predictionDf = ran_for_Model.transform(testdata)

print("RandomForestClassifier prediction")

predictionDf.show(10,False)

"""
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------------
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|rawPrediction
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------------
|1.0         |1.0         |1.0         |1.0         |1.0         |F    |[1.0,1.0,1.0,1.0,1.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |[0.0,0.0,0.0,0.0,1.116071428571428
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |[0.0,0.0,0.0,0.0,1.116071428571428
|4.1         |4.1         |4.1         |4.1         |4.1         |E    |[4.1,4.1,4.1,4.1,4.1]|5.0  |[0.0,0.0,0.0,0.0,1.116071428571428
|4.2         |4.2         |4.2         |4.2         |4.2         |E    |[4.2,4.2,4.2,4.2,4.2]|5.0  |[0.0,0.0,0.0,0.0,1.116071428571428
|4.3         |4.3         |4.3         |4.3         |4.3         |E    |[4.3,4.3,4.3,4.3,4.3]|5.0  |[0.0,0.0,0.0,0.0,1.116071428571428
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------------
"""
```

```
RandomForestClassifier prediction
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------+-
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|grade|features             |label|rawPrediction                |p
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------+-
|1.0         |1.0         |1.0         |1.0         |1.0         |F    |[1.0,1.0,1.0,1.0,1.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]|[
|2.0         |2.0         |2.0         |3.0         |3.0         |F    |[2.0,2.0,2.0,3.0,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]|[
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]|[
|2.0         |2.1         |3.5         |2.4         |3.0         |F    |[2.0,2.1,3.5,2.4,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]|[
|3.0         |3.0         |3.0         |3.0         |3.0         |F    |[3.0,3.0,3.0,3.0,3.0]|0.0  |[20.0,0.0,0.0,0.0,0.0,0.0,0.0]|[
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |[0.0,0.0,0.0,0.0,0.0,20.0,0.0]|[
|4.0         |4.0         |4.0         |4.0         |4.0         |E    |[4.0,4.0,4.0,4.0,4.0]|5.0  |[0.0,0.0,0.0,0.0,0.0,20.0,0.0]|[
|4.1         |4.1         |4.1         |4.1         |4.1         |E    |[4.1,4.1,4.1,4.1,4.1]|5.0  |[0.0,0.0,0.0,0.0,0.0,20.0,0.0]|[
|4.2         |4.2         |4.2         |4.2         |4.2         |E    |[4.2,4.2,4.2,4.2,4.2]|5.0  |[0.0,0.0,0.0,0.0,0.0,20.0,0.0]|[
|4.3         |4.3         |4.3         |4.3         |4.3         |E    |[4.3,4.3,4.3,4.3,4.3]|5.0  |[0.0,0.0,0.0,0.0,0.0,20.0,0.0]|[
+------------+------------+------------+------------+------------+-----+---------------------+-----+-----------------------------+-
only showing top 10 rows



'\n+------------+------------+------------+------------+------------+-----+---------------------+-----+------------------------------
```

```python
evaluator = BinaryClassificationEvaluator() .setLabelCol("label").setRawPredictionCol("prediction").setMetricName("areaUnderROC")

accuracy = evaluator.evaluate(predictionDf)

print("accuracy of the model")

print(accuracy * 100)
```

```
accuracy of the model
100.0
```

```
df1 = spark.createDataFrame(
    [
      (9.1,9.2,9.3,9.4,9.5),
      (9.0,9.0,9.0,9.0,9.0),
      (2.1,2.0,2.4,3.5,3.0),
      (8.0,8.1,8.2,8.3,8.4),
      (7.0,7.1,7.2,7.3,7.35),
      (6.0,6.1,6.2,6.3,6.4),
      (5.0,5.1,5.2,5.3,5.4)
      ],
    ["subject_1_gp","subject_2_gp","subject_3_gp","subject_4_gp","subject_5_gp"]
    )

print("new values for prediction")

df1.printSchema()

df1.show(10,False)

"""
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)
"""

"""
+------------+------------+------------+------------+------------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|
+------------+------------+------------+------------+------------+
|9.1         |9.2         |9.3         |9.4         |9.5         |
|9.0         |9.0         |9.0         |9.0         |9.0         |
|2.1         |2.0         |2.4         |3.5         |3.0         |
|8.0         |8.1         |8.2         |8.3         |8.4         |
|7.0         |7.1         |7.2         |7.3         |7.35        |
|6.0         |6.1         |6.2         |6.3         |6.4         |
|5.0         |5.1         |5.2         |5.3         |5.4         |
+------------+------------+------------+------------+------------+
"""
```

```
new values for prediction
root
 |-- subject_1_gp: double (nullable = true)
 |-- subject_2_gp: double (nullable = true)
 |-- subject_3_gp: double (nullable = true)
 |-- subject_4_gp: double (nullable = true)
 |-- subject_5_gp: double (nullable = true)

+------------+------------+------------+------------+------------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|
+------------+------------+------------+------------+------------+
|9.1         |9.2         |9.3         |9.4         |9.5         |
|9.0         |9.0         |9.0         |9.0         |9.0         |
|2.1         |2.0         |2.4         |3.5         |3.0         |
|8.0         |8.1         |8.2         |8.3         |8.4         |
|7.0         |7.1         |7.2         |7.3         |7.35        |
|6.0         |6.1         |6.2         |6.3         |6.4         |
|5.0         |5.1         |5.2         |5.3         |5.4         |
+------------+------------+------------+------------+------------+



'\n+------------+------------+------------+------------+------------+\n|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_
```

```
df2 = marksDF_assembler.transform(df1)

df3 = ran_for_Model.transform(df2)

df3.createOrReplaceTempView("input_marks_view")

print("prediction of given data")

df3.show()

"""
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+-----
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|            features|       rawPrediction|         probability|predi
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+-----
|         9.1|         9.2|         9.3|         9.4|         9.5|[9.1,9.2,9.3,9.4,...|[0.0,0.0588235294...|[0.0,0.0029411764...|
|         9.0|         9.0|         9.0|         9.0|         9.0|[9.0,9.0,9.0,9.0,...|[0.75,5.058823529...|[0.0375,0.2529411...|
|         2.1|         2.0|         2.4|         3.5|         3.0|[2.1,2.0,2.4,3.5,...|[18.0,0.0,0.0,0.0...|[0.9,0.0,0.0,0.0,...|
|         8.0|         8.1|         8.2|         8.3|         8.4|[8.0,8.1,8.2,8.3,...|[0.0,20.0,0.0,0.0...|[0.0,1.0,0.0,0.0,...|
|         7.0|         7.1|         7.2|         7.3|        7.35|[7.0,7.1,7.2,7.3,...|[0.16666666666666...|[0.00833333333333...|
|         6.0|         6.1|         6.2|         6.3|         6.4|[6.0,6.1,6.2,6.3,...|[0.0,0.0,0.0,20.0...|[0.0,0.0,0.0,1.0,...|
|         5.0|         5.1|         5.2|         5.3|         5.4|[5.0,5.1,5.2,5.3,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+-----
"""
```

```
prediction of given data
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+---
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|            features|       rawPrediction|         probability|pre
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+---
|         9.1|         9.2|         9.3|         9.4|         9.5|[9.1,9.2,9.3,9.4,...|[0.0,0.0588235294...|[0.0,0.0029411764...|
|         9.0|         9.0|         9.0|         9.0|         9.0|[9.0,9.0,9.0,9.0,...|[0.75,5.058823529...|[0.0375,0.2529411...|
|         2.1|         2.0|         2.4|         3.5|         3.0|[2.1,2.0,2.4,3.5,...|[18.0,0.0,0.0,0.0...|[0.9,0.0,0.0,0.0,...|
|         8.0|         8.1|         8.2|         8.3|         8.4|[8.0,8.1,8.2,8.3,...|[0.0,20.0,0.0,0.0...|[0.0,1.0,0.0,0.0,...|
|         7.0|         7.1|         7.2|         7.3|        7.35|[7.0,7.1,7.2,7.3,...|[0.16666666666666...|[0.00833333333333...|
|         6.0|         6.1|         6.2|         6.3|         6.4|[6.0,6.1,6.2,6.3,...|[0.0,0.0,0.0,20.0...|[0.0,0.0,0.0,1.0,...|
|         5.0|         5.1|         5.2|         5.3|         5.4|[5.0,5.1,5.2,5.3,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|
+------------+------------+------------+------------+------------+--------------------+--------------------+--------------------+---
```

```
spark.sql("select subject_1_gp,subject_2_gp,subject_3_gp,subject_4_gp,subject_5_gp,prediction from input_marks_view").show()
"""
+------------+------------+------------+------------+------------+----------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|prediction|
+------------+------------+------------+------------+------------+----------+
|         9.1|         9.2|         9.3|         9.4|         9.5|       6.0|
|         9.0|         9.0|         9.0|         9.0|         9.0|       6.0|
|         2.1|         2.0|         2.4|         3.5|         3.0|       0.0|
|         8.0|         8.1|         8.2|         8.3|         8.4|       1.0|
|         7.0|         7.1|         7.2|         7.3|        7.35|       2.0|
|         6.0|         6.1|         6.2|         6.3|         6.4|       3.0|
|         5.0|         5.1|         5.2|         5.3|         5.4|       4.0|
+------------+------------+------------+------------+------------+----------+
"""
```

```
+------------+------------+------------+------------+------------+----------+
|subject_1_gp|subject_2_gp|subject_3_gp|subject_4_gp|subject_5_gp|prediction|
+------------+------------+------------+------------+------------+----------+
|         9.1|         9.2|         9.3|         9.4|         9.5|       6.0|
|         9.0|         9.0|         9.0|         9.0|         9.0|       6.0|
|         2.1|         2.0|         2.4|         3.5|         3.0|       0.0|
|         8.0|         8.1|         8.2|         8.3|         8.4|       1.0|
|         7.0|         7.1|         7.2|         7.3|        7.35|       2.0|
|         6.0|         6.1|         6.2|         6.3|         6.4|       3.0|
|         5.0|         5.1|         5.2|         5.3|         5.4|       4.0|
+------------+------------+------------+------------+------------+----------+
final_out =spark.sql ("SELECT main_df.subject_1_gp,main_df.subject_2_gp,main_df.subject_3_gp," +
      "main_df.subject_4_gp,main_df.subject_5_gp,main_df.grade,main_df.label,input_marks_df.prediction FROM main_df  " +
      "JOIN input_marks_df  ON main_df.subject_1_gp = input_marks_df.subject_1_gp AND main_df.subject_2_gp = input_marks_df.subject_2_
      "AND main_df.subject_3_gp = input_marks_df.subject_3_gp AND main_df.subject_4_gp = input_marks_view.subject_4_gp AND " +
      "main_df.subject_5_gp = input_marks_df.subject_5_gp  GROUP BY main_df.subject_1_gp,main_df.subject_2_gp," +
      "main_df.subject_3_gp,main_df.subject_4_gp,main_df.subject_5_gp,main_df.grade,input_marks_df.prediction,main_df.label")
```

```
AnalysisException: Table or view not found: main_df; line 1 pos 171;
'Aggregate ['main_df.subject_1_gp, 'main_df.subject_2_gp, 'main_df.subject_3_gp, 'main_df.subject_4_gp, 'main_df.subject_5_gp, 'main
+- 'Join Inner, (((('main_df.subject_1_gp = 'input_marks_df.subject_1_gp) AND ('main_df.subject_2_gp = 'input_marks_df.subject_2_gp)
   :- 'UnresolvedRelation [main_df], [], false
   +- 'UnresolvedRelation [input_marks_df], [], false
```

```
final_out.describe()
```

```
NameError: name 'final_out' is not defined
```