HW2_Part1

October 23, 2022

1 Part 1

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.types import LongType, StringType, StructField, StructType,␣
 ↪BooleanType, ArrayType, IntegerType, DoubleType
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer, OneHotEncoder, VectorAssembler
from pyspark.ml.linalg import Vectors
from pyspark.ml.classification import LogisticRegression,␣
 ↪RandomForestClassifier, NaiveBayes, DecisionTreeClassifier, GBTClassifier,\
MultilayerPerceptronClassifier, LinearSVC, OneVsRest
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

Schema = StructType([
    StructField("age", IntegerType(), True),
    StructField("workclass", StringType(), True),
    StructField("fnlwgt", DoubleType(), True),
    StructField("education", StringType(), True),
    StructField("education_num", DoubleType(), True),
    StructField("marital_status", StringType(), True),
    StructField("occupation", StringType(), True),
    StructField("relationship", StringType(), True),
    StructField("race", StringType(), True),
    StructField("sex", StringType(), True),
    StructField("capital_gain", DoubleType(), True),
    StructField("capital_loss", DoubleType(), True),
    StructField("hours_per_week", DoubleType(), True),
    StructField("native_country", StringType(), True),
    StructField("income", StringType(), True),
])

spark = SparkSession \
    .builder \
    .getOrCreate()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
```

```
22/10/23 14:34:48 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/10/23 14:34:48 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/10/23 14:34:48 INFO org.apache.spark.SparkEnv: Registering
BlockManagerMasterHeartbeat
22/10/23 14:34:48 INFO org.apache.spark.SparkEnv: Registering
OutputCommitCoordinator
```

1. Data loading

[2]:
```
df = spark.read.option("inferSchema", "True").schema(Schema).csv("gs://
↪eecs6893_data/notebooks/jupyter/EECS_6893_Big_Data_Analysis/HW2/adult.data")
df.take(3)
df.printSchema()
```

```
[Stage 0:>                                                          (0 + 1) / 1]

root
 |-- age: integer (nullable = true)
 |-- workclass: string (nullable = true)
 |-- fnlwgt: double (nullable = true)
 |-- education: string (nullable = true)
 |-- education_num: double (nullable = true)
 |-- marital_status: string (nullable = true)
 |-- occupation: string (nullable = true)
 |-- relationship: string (nullable = true)
 |-- race: string (nullable = true)
 |-- sex: string (nullable = true)
 |-- capital_gain: double (nullable = true)
 |-- capital_loss: double (nullable = true)
 |-- hours_per_week: double (nullable = true)
 |-- native_country: string (nullable = true)
 |-- income: string (nullable = true)
```

2. Data preprocessing

[3]:
```
indexer = StringIndexer(inputCols=["workclass", "education", "marital_status",\
                                   "occupation", "relationship", "race", "sex",\
                                  "native_country"],\
                        outputCols=["workclassIndex", "educationIndex",␣
↪"marital_statusIndex",\
                                   "occupationIndex", "relationshipIndex",␣
↪"raceIndex", "sexIndex",\
                                  "native_countryIndex"])
# indexed = indexer.fit(df).transform(df)
```

[4]:
```
encoder = OneHotEncoder(inputCols=["workclassIndex", "educationIndex",␣
↪"marital_statusIndex",\
                                   "occupationIndex", "relationshipIndex",␣
↪"raceIndex", "sexIndex",\
                                  "native_countryIndex"],
                        outputCols=["workclassVec", "educationVec",␣
↪"marital_statusVec",\
                                   "occupationVec", "relationshipVec",␣
↪"raceVec", "sexVec",\
                                  "native_countryVec"])
# encoded = encoder.fit(indexed).transform(indexed)
```

[5]:
```
assembler = VectorAssembler(
    inputCols=["age", "fnlwgt", "education_num", "capital_gain",␣
↪"capital_loss",\
               "hours_per_week", "workclassVec", "educationVec",␣
↪"marital_statusVec",\
               "occupationVec", "relationshipVec", "raceVec", "sexVec",\
              "native_countryVec"],
    outputCol="features")
# output = assembler.transform(encoded)
```

[6]:
```
labelGenerater = StringIndexer(inputCols=["income"],\
                        outputCols=["label"])
# preprossed = labelGenerater.fit(output).transform(output)
```

[7]:
```
dataPreprocess = Pipeline(stages=[indexer, encoder, assembler, labelGenerater])
df = dataPreprocess.fit(df).transform(df)
df.printSchema()
training, test = df.randomSplit([0.7, 0.3], seed=100)
print(training.count())
print(test.count())
```

```
root
 |-- age: integer (nullable = true)
 |-- workclass: string (nullable = true)
 |-- fnlwgt: double (nullable = true)
 |-- education: string (nullable = true)
 |-- education_num: double (nullable = true)
 |-- marital_status: string (nullable = true)
 |-- occupation: string (nullable = true)
 |-- relationship: string (nullable = true)
 |-- race: string (nullable = true)
 |-- sex: string (nullable = true)
 |-- capital_gain: double (nullable = true)
```

```
 |-- capital_loss: double (nullable = true)
 |-- hours_per_week: double (nullable = true)
 |-- native_country: string (nullable = true)
 |-- income: string (nullable = true)
 |-- workclassIndex: double (nullable = false)
 |-- educationIndex: double (nullable = false)
 |-- marital_statusIndex: double (nullable = false)
 |-- occupationIndex: double (nullable = false)
 |-- relationshipIndex: double (nullable = false)
 |-- raceIndex: double (nullable = false)
 |-- sexIndex: double (nullable = false)
 |-- native_countryIndex: double (nullable = false)
 |-- workclassVec: vector (nullable = true)
 |-- educationVec: vector (nullable = true)
 |-- marital_statusVec: vector (nullable = true)
 |-- occupationVec: vector (nullable = true)
 |-- relationshipVec: vector (nullable = true)
 |-- raceVec: vector (nullable = true)
 |-- sexVec: vector (nullable = true)
 |-- native_countryVec: vector (nullable = true)
 |-- features: vector (nullable = true)
 |-- label: double (nullable = false)
```

```
22/10/23 14:37:18 WARN org.apache.spark.sql.catalyst.util.package: Truncated the
string representation of a plan since it was too large. This behavior can be
adjusted by setting 'spark.sql.debug.maxToStringFields'.
```

```
22832
```

```
[Stage 10:>                                                         (0 + 1) / 1]
```

```
9729
```

3. Modeling

[8]:
```
modelAccuracy = {}
```

[9]:
```
lr = LogisticRegression(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[lr])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
```

```
modelAccuracy["LogisticRegression"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
22/10/23 14:37:28 WARN com.github.fommil.netlib.BLAS: Failed to load
implementation from: com.github.fommil.netlib.NativeSystemBLAS
22/10/23 14:37:28 WARN com.github.fommil.netlib.BLAS: Failed to load
implementation from: com.github.fommil.netlib.NativeRefBLAS
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 7812|
|       1.0| 1917|
+----------+-----+
```

```
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+------------
--+----------------+------+--------------+--------------+------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+---------------------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|         probability|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+------------
--+----------------+------+--------------+--------------+------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+---------------------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|          5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[7.59914926603844…|[0.99949937338691…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
```

```
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.68997409534302…|[0.99875822893237…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[7.87082968144825…|[0.99961842812584…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[6.16958616444024…|[0.99791226614392…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.67809346085757…|[0.99874340656756…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[8.23375073009264…|[0.99973453199508…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.662041833856,-…|[0.99872309940859…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
```

```
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[8.07681465205288…|[0.99968943746093…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[8.4294975613348,…|[0.99978171649723…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|         10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[8.36188920726137…|[0.99976645196451…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         20.0|
Peru| <=50K|           3.0|           5.0|               1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[8.83293681742222…|[0.99985417199170…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.62783300103822…|[0.99867872110765…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|         20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[8.04933605036915…|[0.99968078806835…|       0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
```

```
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[6.63134205442267…|[0.99868334332668…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[8.14442275541637…|[0.99970973407241…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[8.47201130114974…|[0.99979080014583…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[7.42433868432000…|[0.99940380040839…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.98560939771980…|[0.99907575564650…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[6.52671320926417…|[0.99853833041920…|       0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
```

```
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[7.64023852463578…|[0.99951951719839…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+-------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+-------------------+---
------------+-----------------+---------+--------+-------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+-------------------
-+--------------------+----------+
only showing top 20 rows

[Stage 136:>                                                        (0 + 1) / 1]
Test set accuracy = 0.8501387604070305
```

[10]:
```
rf = RandomForestClassifier(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[rf])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["RandomForestClassifier"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 8525|
|       1.0| 1204|
+----------+-----+



+---+---------+--------+-------------+-------------+--------------+----------+--
------------+-------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+-------------------+---
------------+-----------------+---------+--------+-------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+-------------------
```

```
-+--------------------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|         probability|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+--------------------+-------+------------+------------+-----------
--+----------------+------+--------------+--------------+-------------------+---
------------+-----------------+---------+--------+-------------------+----------
---+--------------+-----------------+--------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+-------------------
-+--------------------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1249505481369…|[0.95624752740684…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[18.2109205744704…|[0.91054602872352…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1699940509857…|[0.95849970254928…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
```

```
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.1237041679320…|[0.95618520839660…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.1237041679320…|[0.95618520839660…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1581839642711…|[0.95790919821355…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1581839642711…|[0.95790919821355…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
```

```
Peru| <=50K|           3.0|           5.0|                1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1117705432625…|[0.95558852716312…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.1237041679320…|[0.95618520839660…|       0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1249505481369…|[0.95624752740684…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1581839642711…|[0.95790919821355…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1581839642711…|[0.95790919821355…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
```

```
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1581839642711…|[0.95790919821355…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[19.0904707517978…|[0.95452353758989…|       0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[19.1450039593968…|[0.95725019796984…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+--------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+-------------------+---
------------+-----------------+---------+--------+-------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+-------------------
-+---------------------+----------+
only showing top 20 rows


[Stage 158:>                                                        (0 + 1) / 1]

Test set accuracy = 0.8296844485558639
```

[11]:
```
nb = NaiveBayes(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[nb])
model = pipeline.fit(training)
predictions = model.transform(test)
```

```
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["NaiveBayes"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 8836|
|       1.0|  893|
+----------+-----+
```

```
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+-----------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|probability|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+-----------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|          5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-483.33612716071…|  [1.0,0.0]|
```

```
0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-795.25348381951…|  [1.0,0.0]|
0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[-499.68885477319…|  [1.0,0.0]|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-861.91687380403…|  [1.0,0.0]|
0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-831.48128549490…|  [1.0,0.0]|
0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-647.62644264750…|  [1.0,0.0]|
0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-880.42775887499…|  [1.0,0.0]|
0.0|
```

```
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-720.85198405938…|  [1.0,0.0]|
0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-646.69564071260…|  [1.0,0.0]|
0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-690.72601267990…|  [1.0,0.0]|
0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
Peru| <=50K|           3.0|           5.0|               1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-812.71689639804…|  [1.0,0.0]|
0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-984.74127953645…|  [1.0,0.0]|
0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-804.64290729117…|  [1.0,0.0]|
0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
```

```
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-1002.5532126991…|  [1.0,0.0]|
0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-823.37095531303…|  [1.0,0.0]|
0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-760.25107398618…|  [1.0,0.0]|
0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-1117.6809115486…|  [1.0,0.0]|
0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-1109.7343776724…|  [1.0,0.0]|
0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[-1293.0874176433…|  [1.0,0.0]|
0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
```

```
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[-1695.4784324160…|  [1.0,0.0]|
0.0|
+---+---------+--------+------------+------------+-------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+-------------+-------------+------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+-------------+---------------+------------
-+-------------+-----------------+-------------------+-----+------------------
-+-----------+----------+
only showing top 20 rows


[Stage 169:>                                                        (0 + 1) / 1]

Test set accuracy = 0.7829170521122417
```

[12]:
```
dtc = DecisionTreeClassifier(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[dtc])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["DecisionTreeClassifier"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 8308|
|       1.0| 1421|
+----------+-----+



+---+---------+--------+------------+------------+-------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+-------------+-------------+------------------+---
```

```
-----------+----------------+---------+--------+------------------+----------
---+--------------+----------------+--------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+--------------+---
-----------------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|              race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|  rawPrediction|         probability|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+------------------+-------+-----------+-----------+-----------
--+----------------+------+--------------+--------------+-------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+----------------+--------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+--------------+---
-----------------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[11495.0,611.0]|[0.94952915909466…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
```

```
0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
```

```
0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
Peru| <=50K|           3.0|           5.0|                1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
```

```
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[11495.0,611.0]|[0.94952915909466…|
0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+-------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+--------------+---
-----------------+----------+
only showing top 20 rows

Test set accuracy = 0.837804502004317
```

[13]:
```
gbtc = GBTClassifier(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[gbtc])
model = pipeline.fit(training)
```

```
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["GBTClassifier"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 7970|
|       1.0| 1759|
+----------+-----+
```

```
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+--------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+--------------------+---
------------+-----------------+---------+--------+--------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+--------------------
-+---------------------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|         probability|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+--------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+--------------------+---
------------+-----------------+---------+--------+--------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+--------------------
-+---------------------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|               White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
```

```
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52891824356618…|[0.95511964684074…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.52891824356618…|[0.95511964684074…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.51057975503622…|[0.95352094056127…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
```

```
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
Peru| <=50K|           3.0|           5.0|                1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52891824356618…|[0.95511964684074…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
```

```
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.52829596537235…|[0.95506626732494…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[1.51057975503622…|[0.95352094056127…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.52891824356618…|[0.95511964684074…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[1.51120203323005…|[0.95357606663068…|       0.0|
```

```
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.78990689985924…|[0.82917814586072…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+-----------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+------------------
-+--------------------+----------+
only showing top 20 rows


[Stage 403:>                                                        (0 + 1) / 1]

Test set accuracy = 0.8505499023537877
```

[14]:
```
mpc = MultilayerPerceptronClassifier(layers=[100, 2, 2], seed=100)
pipeline = Pipeline(stages=[mpc])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["MultilayerPerceptronClassifier"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 9726|
|       1.0|    3|
+----------+-----+



+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+-----------
```

```
--+----------------+------+--------------+-------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+----------------+-------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+------------------
-+--------------------+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|         probability|prediction|
+---+---------+--------+------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+-------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+----------------+-------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+------------------
-+--------------------+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
```

```
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
```

```
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
Peru| <=50K|           3.0|           5.0|                1.0|            7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
```

```
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|
0.0|[0.40646876155099…|[0.76090274269450…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+--------------------+----------+
only showing top 20 rows

[Stage 448:>                                                        (0 + 1) / 1]

Test set accuracy = 0.7570151094665434
```

[15]:
```
lsvc = LinearSVC(labelCol="label", featuresCol="features")
pipeline = Pipeline(stages=[lsvc])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["LinearSVC"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 7911|
|       1.0| 1818|
+----------+-----+
```

```
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
```

```
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.08546793419189…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.70641115312414…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[3.62766356362195…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[3.48939903811833…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.70223101390926…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[4.37964678648136…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.69658333275436…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
```

```
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[4.30998479785766…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.44124122604992…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|         10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.41167564008095…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         20.0|
Peru| <=50K|           3.0|           5.0|                1.0|            7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.75690587595968…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.68454713353185…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|         20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[4.30031659532112…|       0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[3.65828092613439…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|         20.0|
```

```
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.28673408598901…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.46486637279695…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[3.99513428996327…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.85376343754010…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|         40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[3.64896866274293…|       0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|         40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[4.07019156248438…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+------------+------------+------------
--+----------------+------+--------------+--------------+------------------+---
------------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+--------------+------------
-+-------------+-----------------+------------------+-----+------------------
-+----------+
only showing top 20 rows
```

```
[Stage 664:>                                                        (0 + 1) / 1]
```

```
Test set accuracy = 0.8485969781066913
```

[16]:
```
ovr = OneVsRest(labelCol="label", featuresCol="features", classifier=lr)
pipeline = Pipeline(stages=[ovr])
model = pipeline.fit(training)
predictions = model.transform(test)
predictions.groupby("prediction").count().show()
predictions.show()
evaluator = MulticlassClassificationEvaluator(labelCol="label",␣
 ↪predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

modelAccuracy["OneVsRest"] = accuracy
print(f"Test set accuracy = {accuracy}")
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|       0.0| 7812|
|       1.0| 1917|
+----------+-----+
```

```
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+------------
--+----------------+------+--------------+--------------+------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+--------------+---------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+----------+
|age|workclass|  fnlwgt|    education|education_num|marital_status|occupation|
relationship|                race|
sex|capital_gain|capital_loss|hours_per_week|  native_country|income|workclassIn
dex|educationIndex|marital_statusIndex|occupationIndex|relationshipIndex|raceInd
ex|sexIndex|native_countryIndex| workclassVec|  educationVec|marital_statusVec|
occupationVec|relationshipVec|      raceVec|       sexVec|native_countryVec|
features|label|       rawPrediction|prediction|
+---+---------+--------+-------------+-------------+--------------+----------+--
------------+------------------+-------+-----------+-----------+------------
--+----------------+------+--------------+--------------+------------------+---
-----------+-----------------+---------+--------+------------------+----------
```

```
---+--------------+----------------+--------------+--------------+------------
-+-------------+-----------------+--------------------+-----+------------------
-+----------+
| 17|        ?| 47407.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[7.59914231450615…|       0.0|
| 17|        ?| 89870.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.68996764295798…|       0.0|
| 17|        ?| 94366.0|         10th|          6.0| Never-married|         ?|
Other-relative|              White|   Male|         0.0|         0.0|
6.0|   United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              5.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[],[])|(4,[0],[1.0])|(1,[0],[1.0])|   (41,[0],[1.0])|(100,[0,1,2,5,9,2…|
0.0|[7.87083047830965…|       0.0|
| 17|        ?|110998.0| Some-college|         10.0| Never-married|         ?|
Own-child| Asian-Pac-Islander| Female|         0.0|         0.0|          40.0|
Philippines| <=50K|           3.0|           1.0|                1.0|
7.0|              2.0|      2.0|     1.0|
3.0|(8,[3],[1.0])|(15,[1],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[2],[1.0])|    (1,[],[])|
(41,[3],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[6.16958441292778…|       0.0|
| 17|        ?|112942.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.67808709490210…|       0.0|
| 17|        ?|139183.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          15.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[8.23374475518700…|       0.0|
| 17|        ?|144114.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
```

```
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.66203558467335…|       0.0|
| 17|        ?|158762.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[8.07680885660970…|       0.0|
| 17|        ?|170320.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           8.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[8.42949185113601…|       0.0|
| 17|        ?|187539.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          10.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[8.36188360401352…|       0.0|
| 17|        ?|198797.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          20.0|
Peru| <=50K|           3.0|           5.0|                1.0|           7.0|
2.0|      0.0|     0.0|               26.0|(8,[3],[1.0])|(15,[5],[1.0])|
(6,[1],[1.0])|(14,[7],[1.0])|  (5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[26],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[8.83293119038318…|       0.0|
| 17|        ?|210547.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.62782700071894…|       0.0|
| 17|        ?|212125.0|         10th|          6.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[8.04933045482807…|       0.0|
| 17|        ?|215743.0|         11th|          7.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
```

```
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[6.63133637019830…|       0.0|
| 17|        ?|216595.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          20.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[8.14441692391919…|       0.0|
| 17|        ?|258872.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|           5.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[8.47200585900310…|       0.0|
| 17|        ?|297117.0|         11th|          7.0| Never-married|         ?|
Own-child|              White| Female|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      0.0|     1.0|
0.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|    (1,[],[])|
(41,[0],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[7.42433412826788…|       0.0|
| 17|        ?|371316.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          25.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.98560368130029…|       0.0|
| 17|        ?|406920.0|         10th|          6.0| Never-married|         ?|
Own-child|              White|   Male|         0.0|         0.0|          40.0|
United-States| <=50K|           3.0|           7.0|                1.0|
7.0|              2.0|      0.0|     0.0|
0.0|(8,[3],[1.0])|(15,[7],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[0],[1.0])|(1,[0],[1.0])|
(41,[0],[1.0])|(100,[0,1,2,5,9,2…|  0.0|[6.52670794457397…|       0.0|
| 17|        ?|659273.0|         11th|          7.0| Never-married|         ?|
Own-child|              Black| Female|         0.0|         0.0|          40.0|
Trinadad&Tobago| <=50K|           3.0|           5.0|                1.0|
7.0|              2.0|      1.0|     1.0|
33.0|(8,[3],[1.0])|(15,[5],[1.0])|    (6,[1],[1.0])|(14,[7],[1.0])|
(5,[2],[1.0])|(4,[1],[1.0])|    (1,[],[])|
(41,[33],[1.0])|(100,[0,1,2,5,9,1…|  0.0|[7.64023183995303…|       0.0|
+---+---------+--------+-------------+-------------+--------------+----------+--
-------------+------------------+-------+------------+------------+------------
```

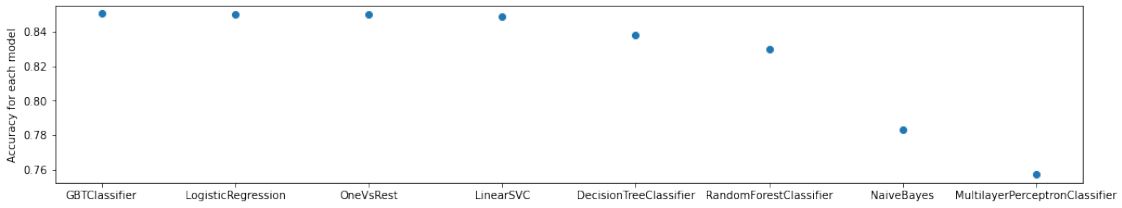
```
--+----------------+------+--------------+--------------+--------------------+---
-----------+-----------------+---------+--------+------------------+----------
---+--------------+-----------------+-------------+--------------+-----------
-+-------------+-----------------+--------------------+-----+------------------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

[Stage 909:>                                                        (0 + 1) / 1]
Test set accuracy = 0.8501387604070305
```

4. Comparison and analysis

[17]:
```
sorted_modelAccuracy = dict(sorted(modelAccuracy.items(), key=lambda item:␣
↪item[1], reverse=True))
print(sorted_modelAccuracy)
```

```
{'GBTClassifier': 0.8505499023537877, 'LogisticRegression': 0.8501387604070305,
'OneVsRest': 0.8501387604070305, 'LinearSVC': 0.8485969781066913,
'DecisionTreeClassifier': 0.837804502004317, 'RandomForestClassifier':
0.8296844485558639, 'NaiveBayes': 0.7829170521122417,
'MultilayerPerceptronClassifier': 0.7570151094665434}
```

[18]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(17, 3))
plt.plot(sorted_modelAccuracy.keys(),sorted_modelAccuracy.values(),'o')
plt.ylabel("Accuracy for each model")
plt.show()
```



As we can see from the plot above, the model with the highest accuracy is GBTClassifier. My assumption is that the defult hyper-paramiters are the best for each model. Therefor, to avoid unfair paramiters, I set only minimum hypr-paramiters and left as much defult as possible.