Machine Learning with PySpark

January 31, 2021

0.0.1 ML with PySpark

- Classify/Predict

Datasource

- https://archive.ics.uci.edu/ml/datasets/HCV+data

[1]:
```
# Load our Pkgs
from pyspark import SparkContext
```

[2]:
```
sc = SparkContext(master='local[2]')
```

[3]:
```
# Spark UI
sc
```

[3]:
```
<SparkContext master=local[2] appName=pyspark-shell>
```

[4]:
```
# Load Pkgs
from pyspark.sql import SparkSession
```

[5]:
```
# Spark
spark = SparkSession.builder.appName("MLwithSpark").getOrCreate()
```

WorkFlow

- Data Prep
- Feature Engineering
- Build Model
- Evaluate

1 Task

- Predict if a patient is Hep or not based parameter
- The data set contains laboratory values of blood donors and Hepatitis C patients and demographic values like age.

[14]:
```
# Load our dataset
df = spark.read.csv("data/hcvdata.csv",header=True,inferSchema=True)
```

[15]:
```
# Preview Dataset
df.show()
```

```
+---+-------------+---+---+----+----+----+----+----+-----+----+-----+----+----+
|_c0|     Category|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|
+---+-------------+---+---+----+----+----+----+----+-----+----+-----+----+----+
|  1|0=Blood Donor| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|
|  2|0=Blood Donor| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|
|  3|0=Blood Donor| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|
|  4|0=Blood Donor| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|
|  5|0=Blood Donor| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|
|  6|0=Blood Donor| 32|  m|41.6|43.3|18.5|19.7|12.3| 9.92|6.05|111.0|91.0|  74|
|  7|0=Blood Donor| 32|  m|46.3|41.3|17.5|17.8| 8.5| 7.01|4.79| 70.0|16.9|74.5|
|  8|0=Blood Donor| 32|  m|42.2|41.9|35.8|31.1|16.1| 5.82| 4.6|109.0|21.5|67.1|
|  9|0=Blood Donor| 32|  m|50.9|65.5|23.2|21.2| 6.9| 8.69| 4.1| 83.0|13.7|71.3|
| 10|0=Blood Donor| 32|  m|42.4|86.3|20.3|20.0|35.2| 5.46|4.45| 81.0|15.9|69.9|
| 11|0=Blood Donor| 32|  m|44.3|52.3|21.7|22.4|17.2| 4.15|3.57| 78.0|24.1|75.4|
| 12|0=Blood Donor| 33|  m|46.4|68.2|10.3|20.0| 5.7| 7.36| 4.3| 79.0|18.7|68.6|
| 13|0=Blood Donor| 33|  m|36.3|78.6|23.6|22.0| 7.0| 8.56|5.38| 78.0|19.4|68.7|
| 14|0=Blood Donor| 33|  m|  39|51.7|15.9|24.0| 6.8| 6.46|3.38| 65.0| 7.0|70.4|
| 15|0=Blood Donor| 33|  m|38.7|39.8|22.5|23.0| 4.1| 4.63|4.97| 63.0|15.2|71.9|
| 16|0=Blood Donor| 33|  m|41.8|  65|33.1|38.0| 6.6| 8.83|4.43| 71.0|24.0|72.7|
| 17|0=Blood Donor| 33|  m|40.9|  73|17.2|22.9|10.0| 6.98|5.22| 90.0|14.7|72.4|
| 18|0=Blood Donor| 33|  m|45.2|88.3|32.4|31.2|10.1| 9.78|5.51|102.0|48.5|76.5|
| 19|0=Blood Donor| 33|  m|36.6|57.1|38.9|40.3|24.9| 9.62| 5.5|112.0|27.6|69.3|
| 20|0=Blood Donor| 33|  m|  42|63.1|32.6|34.9|11.2| 7.01|4.05|105.0|19.1|68.1|
+---+-------------+---+---+----+----+----+----+----+-----+----+-----+----+----+
only showing top 20 rows
```

[10]:
```
# check for columns
print(df.columns)
```

```
['_c0', 'Category', 'Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE',
'CHOL', 'CREA', 'GGT', 'PROT']
```

[16]:
```
# Rearrange
df = df.select('Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL',␣
 ↪'CREA', 'GGT', 'PROT','Category')
```

[12]:
```
df.show(5)
```

```
+---+---+----+----+----+----+----+-----+----+----+----+----+-------------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL|CREA| GGT|PROT|     Category|
+---+---+----+----+----+----+----+-----+----+----+----+----+-------------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23| 106|12.1|  69|0=Blood Donor|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8|  74|15.6|76.5|0=Blood Donor|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2|  86|33.2|79.3|0=Blood Donor|
```

```
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74|  80|33.8|75.7|0=Blood Donor|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32|  76|29.9|68.7|0=Blood Donor|
+---+---+----+----+----+----+----+-----+----+----+----+----+-------------+
only showing top 5 rows
```

[13]:
```
# Check for datatypes
# Before InferSchema=True
df.dtypes
```

[13]:
```
[('Age', 'string'),
 ('Sex', 'string'),
 ('ALB', 'string'),
 ('ALP', 'string'),
 ('ALT', 'string'),
 ('AST', 'string'),
 ('BIL', 'string'),
 ('CHE', 'string'),
 ('CHOL', 'string'),
 ('CREA', 'string'),
 ('GGT', 'string'),
 ('PROT', 'string'),
 ('Category', 'string')]
```

[17]:
```
# After InferSchema
df.dtypes
```

[17]:
```
[('Age', 'int'),
 ('Sex', 'string'),
 ('ALB', 'string'),
 ('ALP', 'string'),
 ('ALT', 'string'),
 ('AST', 'double'),
 ('BIL', 'double'),
 ('CHE', 'double'),
 ('CHOL', 'string'),
 ('CREA', 'double'),
 ('GGT', 'double'),
 ('PROT', 'string'),
 ('Category', 'string')]
```

[18]:
```
# Check for the Schema
df.printSchema()
```

```
root
 |-- Age: integer (nullable = true)
 |-- Sex: string (nullable = true)
 |-- ALB: string (nullable = true)
```

```
 |-- ALP: string (nullable = true)
 |-- ALT: string (nullable = true)
 |-- AST: double (nullable = true)
 |-- BIL: double (nullable = true)
 |-- CHE: double (nullable = true)
 |-- CHOL: string (nullable = true)
 |-- CREA: double (nullable = true)
 |-- GGT: double (nullable = true)
 |-- PROT: string (nullable = true)
 |-- Category: string (nullable = true)
```

[19]:
```
# Descriptive summary
print(df.describe().show())
```

```
+-------+-----------------+----+-----------------+-----------------+----------
--------+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+-------------+
|summary|              Age| Sex|              ALB|              ALP|
ALT|              AST|              BIL|              CHE|             CHOL|
CREA|              GGT|             PROT|     Category|
+-------+-----------------+----+-----------------+-----------------+----------
--------+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+-------------+
|  count|              615| 615|              615|              615|
615|              615|              615|              615|              615|
615|              615|              615|          615|
|   mean| 47.40813008130081|null|41.62019543973941| 68.28391959798999|
28.45081433224754|34.78634146341462|11.396747967479675| 8.196634146341458|
5.368099173553719|81.28780487804877|39.53317073170732|72.04413680781768|
null|
| stddev|10.055105445519239|null|5.780629404103076|26.028315300123676|25.4696888
13870942|33.09069033855156|19.673149805846588|2.2056572704292927|1.1327284311597
354|49.75616601234976|54.66107123891245|5.402635737104955|         null|
|    min|               19|   f|             14.9|            100.4|
0.9|             10.6|              0.8|             1.42|             1.43|
8.0|              4.5|             44.8|0=Blood Donor|
|    max|               77|   m|               NA|               NA|
NA|            324.0|            254.0|            16.41|               NA|
1079.1|            650.9|               NA|  3=Cirrhosis|
+-------+-----------------+----+-----------------+-----------------+----------
--------+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+-------------+

None
```

[21]:
```
# Value Count
df.groupBy('Category').count().show()
```

```
+--------------------+-----+
|            Category|count|
+--------------------+-----+
|       0=Blood Donor|  533|
|         3=Cirrhosis|   30|
|          2=Fibrosis|   21|
|0s=suspect Blood …|    7|
|         1=Hepatitis|   24|
+--------------------+-----+
```

[ ]:
```
#### Feature Engineering
+ Numberical Values
+ Vectorization
+ Scaling
```

[22]:
```
df.show(5)
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|     Category|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|0=Blood Donor|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
only showing top 5 rows
```

[23]:
```
import pyspark.ml
```

[24]:
```
dir(pyspark.ml)
```

[24]:
```
['Estimator',
 'Model',
 'Pipeline',
 'PipelineModel',
 'Transformer',
 'UnaryTransformer',
 '__all__',
 '__builtins__',
 '__cached__',
 '__doc__',
```

```
 '__file__',
 '__loader__',
 '__name__',
 '__package__',
 '__path__',
 '__spec__',
 'base',
 'classification',
 'clustering',
 'common',
 'evaluation',
 'feature',
 'fpm',
 'image',
 'linalg',
 'param',
 'pipeline',
 'recommendation',
 'regression',
 'stat',
 'tree',
 'tuning',
 'util',
 'wrapper']
```

[25]:
```
# Load ML Pkgs
from pyspark.ml.feature import VectorAssembler,StringIndexer
```

[26]:
```
df.show(4)
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|     Category|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+
only showing top 4 rows
```

[27]:
```
# Unique Values for Sex
df.select('Sex').distinct().show()
```

```
+---+
|Sex|
+---+
```

```
|  m|
|  f|
+---+
```

[29]:
```
# Convert the string into numerical code
# label encoding
genderEncoder = StringIndexer(inputCol='Sex',outputCol='Gender').fit(df)
```

[30]:
```
df = genderEncoder.transform(df)
```

[31]:
```
df.show(5)
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|
Category|Gender|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
0.0|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
0.0|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
0.0|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
0.0|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|0=Blood Donor|
0.0|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+
only showing top 5 rows
```

[32]:
```
# Encoding for Category
# Label Encoding
catEncoder = StringIndexer(inputCol='Category',outputCol='Target').fit(df)
df = catEncoder.transform(df)
```

[33]:
```
df.show(5)
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|
Category|Gender|Target|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
```

```
0.0|   0.0|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
0.0|   0.0|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
0.0|   0.0|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
0.0|   0.0|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|0=Blood Donor|
0.0|   0.0|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
only showing top 5 rows
```

[34]:
```
# Get the labels
catEncoder.labels
```

[34]:
```
['0=Blood Donor',
 '3=Cirrhosis',
 '1=Hepatitis',
 '2=Fibrosis',
 '0s=suspect Blood Donor']
```

[35]:
```
# IndexToString
from pyspark.ml.feature import IndexToString
```

[36]:
```
converter = IndexToString(inputCol='Target',outputCol='orig_cat')
```

[37]:
```
converted_df = converter.transform(df)
```

[38]:
```
converted_df.show()
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+-------------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|
Category|Gender|Target|     orig_cat|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+-------------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
```

```
| 32|  m|41.6|43.3|18.5|19.7|12.3| 9.92|6.05|111.0|91.0|  74|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|46.3|41.3|17.5|17.8| 8.5| 7.01|4.79| 70.0|16.9|74.5|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|42.2|41.9|35.8|31.1|16.1| 5.82| 4.6|109.0|21.5|67.1|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|50.9|65.5|23.2|21.2| 6.9| 8.69| 4.1| 83.0|13.7|71.3|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|42.4|86.3|20.3|20.0|35.2| 5.46|4.45| 81.0|15.9|69.9|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 32|  m|44.3|52.3|21.7|22.4|17.2| 4.15|3.57| 78.0|24.1|75.4|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|46.4|68.2|10.3|20.0| 5.7| 7.36| 4.3| 79.0|18.7|68.6|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|36.3|78.6|23.6|22.0| 7.0| 8.56|5.38| 78.0|19.4|68.7|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|  39|51.7|15.9|24.0| 6.8| 6.46|3.38| 65.0| 7.0|70.4|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|38.7|39.8|22.5|23.0| 4.1| 4.63|4.97| 63.0|15.2|71.9|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|41.8|  65|33.1|38.0| 6.6| 8.83|4.43| 71.0|24.0|72.7|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|40.9|  73|17.2|22.9|10.0| 6.98|5.22| 90.0|14.7|72.4|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|45.2|88.3|32.4|31.2|10.1| 9.78|5.51|102.0|48.5|76.5|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|36.6|57.1|38.9|40.3|24.9| 9.62| 5.5|112.0|27.6|69.3|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
| 33|  m|  42|63.1|32.6|34.9|11.2| 7.01|4.05|105.0|19.1|68.1|0=Blood Donor|
0.0|   0.0|0=Blood Donor|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+-------------+
only showing top 20 rows
```

[39]:
```
### Feature
df.show()
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|
Category|Gender|Target|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
| 32|  m|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|  69|0=Blood Donor|
0.0|   0.0|
| 32|  m|38.5|70.3|  18|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|0=Blood Donor|
0.0|   0.0|
```

```
| 32|  m|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|0=Blood Donor|
0.0|   0.0|
| 32|  m|43.2|  52|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|0=Blood Donor|
0.0|   0.0|
| 32|  m|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|0=Blood Donor|
0.0|   0.0|
| 32|  m|41.6|43.3|18.5|19.7|12.3| 9.92|6.05|111.0|91.0|  74|0=Blood Donor|
0.0|   0.0|
| 32|  m|46.3|41.3|17.5|17.8| 8.5| 7.01|4.79| 70.0|16.9|74.5|0=Blood Donor|
0.0|   0.0|
| 32|  m|42.2|41.9|35.8|31.1|16.1| 5.82| 4.6|109.0|21.5|67.1|0=Blood Donor|
0.0|   0.0|
| 32|  m|50.9|65.5|23.2|21.2| 6.9| 8.69| 4.1| 83.0|13.7|71.3|0=Blood Donor|
0.0|   0.0|
| 32|  m|42.4|86.3|20.3|20.0|35.2| 5.46|4.45| 81.0|15.9|69.9|0=Blood Donor|
0.0|   0.0|
| 32|  m|44.3|52.3|21.7|22.4|17.2| 4.15|3.57| 78.0|24.1|75.4|0=Blood Donor|
0.0|   0.0|
| 33|  m|46.4|68.2|10.3|20.0| 5.7| 7.36| 4.3| 79.0|18.7|68.6|0=Blood Donor|
0.0|   0.0|
| 33|  m|36.3|78.6|23.6|22.0| 7.0| 8.56|5.38| 78.0|19.4|68.7|0=Blood Donor|
0.0|   0.0|
| 33|  m|  39|51.7|15.9|24.0| 6.8| 6.46|3.38| 65.0| 7.0|70.4|0=Blood Donor|
0.0|   0.0|
| 33|  m|38.7|39.8|22.5|23.0| 4.1| 4.63|4.97| 63.0|15.2|71.9|0=Blood Donor|
0.0|   0.0|
| 33|  m|41.8|  65|33.1|38.0| 6.6| 8.83|4.43| 71.0|24.0|72.7|0=Blood Donor|
0.0|   0.0|
| 33|  m|40.9|  73|17.2|22.9|10.0| 6.98|5.22| 90.0|14.7|72.4|0=Blood Donor|
0.0|   0.0|
| 33|  m|45.2|88.3|32.4|31.2|10.1| 9.78|5.51|102.0|48.5|76.5|0=Blood Donor|
0.0|   0.0|
| 33|  m|36.6|57.1|38.9|40.3|24.9| 9.62| 5.5|112.0|27.6|69.3|0=Blood Donor|
0.0|   0.0|
| 33|  m|  42|63.1|32.6|34.9|11.2| 7.01|4.05|105.0|19.1|68.1|0=Blood Donor|
0.0|   0.0|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+-----
-+------+
only showing top 20 rows
```

[40]:
```
print(df.columns)
```

```
['Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA', 'GGT',
'PROT', 'Category', 'Gender', 'Target']
```

[45]:
```
df.dtypes
```

[45]:
```
[('Age', 'int'),
 ('Sex', 'string'),
 ('ALB', 'string'),
 ('ALP', 'string'),
 ('ALT', 'string'),
 ('AST', 'double'),
 ('BIL', 'double'),
 ('CHE', 'double'),
 ('CHOL', 'string'),
 ('CREA', 'double'),
 ('GGT', 'double'),
 ('PROT', 'string'),
 ('Category', 'string'),
 ('Gender', 'double'),
 ('Target', 'double')]
```

[47]:
```
df2 = df.select('Age','Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE',␣
↪'CHOL', 'CREA', 'GGT', 'PROT', 'Target')
```

[48]:
```
df2.printSchema()
```

```
root
 |-- Age: integer (nullable = true)
 |-- Gender: double (nullable = false)
 |-- ALB: string (nullable = true)
 |-- ALP: string (nullable = true)
 |-- ALT: string (nullable = true)
 |-- AST: double (nullable = true)
 |-- BIL: double (nullable = true)
 |-- CHE: double (nullable = true)
 |-- CHOL: string (nullable = true)
 |-- CREA: double (nullable = true)
 |-- GGT: double (nullable = true)
 |-- PROT: string (nullable = true)
 |-- Target: double (nullable = false)
```

[ ]:
```
# df2.fillna(0,subset=['col1'])
```

[73]:
```
df2 = df2.toPandas().replace('NA',0).astype(float)
```

[74]:
```
type(df2)
```

[74]:
```
pandas.core.frame.DataFrame
```

[75]:
```
type(df)
```

[75]:
```
pyspark.sql.dataframe.DataFrame
```

[76]:
```
# Convert To PySpark Dataframe
new_df = spark.createDataFrame(df2)
```

[77]:
```
new_df.show()
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+
| Age|Gender| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+
|32.0|   0.0|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|69.0|   0.0|
|32.0|   0.0|38.5|70.3|18.0|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|   0.0|
|32.0|   0.0|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|   0.0|
|32.0|   0.0|43.2|52.0|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|   0.0|
|32.0|   0.0|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|   0.0|
|32.0|   0.0|41.6|43.3|18.5|19.7|12.3| 9.92|6.05|111.0|91.0|74.0|   0.0|
|32.0|   0.0|46.3|41.3|17.5|17.8| 8.5| 7.01|4.79| 70.0|16.9|74.5|   0.0|
|32.0|   0.0|42.2|41.9|35.8|31.1|16.1| 5.82| 4.6|109.0|21.5|67.1|   0.0|
|32.0|   0.0|50.9|65.5|23.2|21.2| 6.9| 8.69| 4.1| 83.0|13.7|71.3|   0.0|
|32.0|   0.0|42.4|86.3|20.3|20.0|35.2| 5.46|4.45| 81.0|15.9|69.9|   0.0|
|32.0|   0.0|44.3|52.3|21.7|22.4|17.2| 4.15|3.57| 78.0|24.1|75.4|   0.0|
|33.0|   0.0|46.4|68.2|10.3|20.0| 5.7| 7.36| 4.3| 79.0|18.7|68.6|   0.0|
|33.0|   0.0|36.3|78.6|23.6|22.0| 7.0| 8.56|5.38| 78.0|19.4|68.7|   0.0|
|33.0|   0.0|39.0|51.7|15.9|24.0| 6.8| 6.46|3.38| 65.0| 7.0|70.4|   0.0|
|33.0|   0.0|38.7|39.8|22.5|23.0| 4.1| 4.63|4.97| 63.0|15.2|71.9|   0.0|
|33.0|   0.0|41.8|65.0|33.1|38.0| 6.6| 8.83|4.43| 71.0|24.0|72.7|   0.0|
|33.0|   0.0|40.9|73.0|17.2|22.9|10.0| 6.98|5.22| 90.0|14.7|72.4|   0.0|
|33.0|   0.0|45.2|88.3|32.4|31.2|10.1| 9.78|5.51|102.0|48.5|76.5|   0.0|
|33.0|   0.0|36.6|57.1|38.9|40.3|24.9| 9.62| 5.5|112.0|27.6|69.3|   0.0|
|33.0|   0.0|42.0|63.1|32.6|34.9|11.2| 7.01|4.05|105.0|19.1|68.1|   0.0|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+
only showing top 20 rows
```

[78]:
```
# Check For DTYpes and Schema
new_df.printSchema()
```

```
root
 |-- Age: double (nullable = true)
 |-- Gender: double (nullable = true)
 |-- ALB: double (nullable = true)
 |-- ALP: double (nullable = true)
 |-- ALT: double (nullable = true)
 |-- AST: double (nullable = true)
 |-- BIL: double (nullable = true)
 |-- CHE: double (nullable = true)
 |-- CHOL: double (nullable = true)
 |-- CREA: double (nullable = true)
```

```
 |-- GGT: double (nullable = true)
 |-- PROT: double (nullable = true)
 |-- Target: double (nullable = true)
```

[79]:
```
required_features = ['Age','Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE',␣
 ↪'CHOL', 'CREA', 'GGT', 'PROT', 'Target']
```

[80]:
```
# VectorAsm
vec_assembler =␣
 ↪VectorAssembler(inputCols=required_features,outputCol='features')
```

[81]:
```
vec_df = vec_assembler.transform(new_df)
```

[82]:
```
vec_df.show(5)
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
| Age|Gender| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|
features|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
|32.0|   0.0|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|69.0|
0.0|[32.0,0.0,38.5,52…|
|32.0|   0.0|38.5|70.3|18.0|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|
0.0|[32.0,0.0,38.5,70…|
|32.0|   0.0|46.9|74.7|36.2|52.6| 6.1| 8.84| 5.2| 86.0|33.2|79.3|
0.0|[32.0,0.0,46.9,74…|
|32.0|   0.0|43.2|52.0|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|
0.0|[32.0,0.0,43.2,52…|
|32.0|   0.0|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|
0.0|[32.0,0.0,39.2,74…|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
only showing top 5 rows
```

1.0.1 Train,Test Split

[83]:
```
train_df,test_df = vec_df.randomSplit([0.7,0.3])
```

[84]:
```
train_df.count()
```

[84]: 439

[89]:
```
train_df.show(4)
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
| Age|Gender| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|
features|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
|32.0|   0.0|38.5|52.5| 7.7|22.1| 7.5| 6.93|3.23|106.0|12.1|69.0|
0.0|[32.0,0.0,38.5,52…|
|32.0|   0.0|38.5|70.3|18.0|24.7| 3.9|11.17| 4.8| 74.0|15.6|76.5|
0.0|[32.0,0.0,38.5,70…|
|32.0|   0.0|39.2|74.1|32.6|24.8| 9.6| 9.15|4.32| 76.0|29.9|68.7|
0.0|[32.0,0.0,39.2,74…|
|32.0|   0.0|41.6|43.3|18.5|19.7|12.3| 9.92|6.05|111.0|91.0|74.0|
0.0|[32.0,0.0,41.6,43…|
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+--------
------------+
only showing top 4 rows
```

[ ]:
```
#### Model Building
+ Pyspark.ml: DataFrame
+ Pyspark.mllib: RDD /Legacy
```

[85]:
```
from pyspark.ml.classification import LogisticRegression,DecisionTreeClassifier
```

[87]:
```
# Logist Model
lr = LogisticRegression(featuresCol='features',labelCol='Target')
```

[88]:
```
lr_model = lr.fit(train_df)
```

[90]:
```
y_pred = lr_model.transform(test_df)
```

[91]:
```
y_pred.show()
```

```
+----+------+----+-----+----+----+----+-----+----+-----+----+----+------+-------
------------+--------------------+--------------------+----------+
| Age|Gender| ALB|  ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|
features|       rawPrediction|         probability|prediction|
+----+------+----+-----+----+----+----+-----+----+-----+----+----+------+-------
------------+--------------------+--------------------+----------+
|32.0|   0.0|42.4| 86.3|20.3|20.0|35.2| 5.46|4.45| 81.0|15.9|69.9|
0.0|[32.0,0.0,42.4,86…|[286.954869615973…|[1.0,3.5781823002…|       0.0|
|32.0|   0.0|43.2| 52.0|30.6|22.6|18.9| 7.33|4.74| 80.0|33.8|75.7|
0.0|[32.0,0.0,43.2,52…|[281.614235265908…|[1.0,1.0244873107…|       0.0|
|32.0|   0.0|46.3| 41.3|17.5|17.8| 8.5| 7.01|4.79| 70.0|16.9|74.5|
0.0|[32.0,0.0,46.3,41…|[329.637070801162…|[1.0,2.3941933511…|       0.0|
|32.0|   0.0|50.9| 65.5|23.2|21.2| 6.9| 8.69| 4.1| 83.0|13.7|71.3|
0.0|[32.0,0.0,50.9,65…|[375.018142931757…|[1.0,3.8219147599…|       0.0|
```

```
|33.0|   0.0|36.6| 57.1|38.9|40.3|24.9| 9.62| 5.5|112.0|27.6|69.3|
0.0|[33.0,0.0,36.6,57…|[266.834562998572…|[1.0,6.7251788183…|       0.0|
|33.0|   0.0|38.7| 39.8|22.5|23.0| 4.1| 4.63|4.97| 63.0|15.2|71.9|
0.0|[33.0,0.0,38.7,39…|[282.416699150824…|[1.0,3.1777824373…|       0.0|
|33.0|   0.0|40.9| 73.0|17.2|22.9|10.0| 6.98|5.22| 90.0|14.7|72.4|
0.0|[33.0,0.0,40.9,73…|[340.661989788631…|[1.0,1.8429339582…|       0.0|
|33.0|   0.0|42.0| 63.1|32.6|34.9|11.2| 7.01|4.05|105.0|19.1|68.1|
0.0|[33.0,0.0,42.0,63…|[287.262205429983…|[1.0,9.1668901248…|       0.0|
|33.0|   0.0|44.3| 49.8|32.1|21.6|13.1| 7.44|5.59|103.0|30.2|74.0|
0.0|[33.0,0.0,44.3,49…|[319.289845055447…|[1.0,1.9214316512…|       0.0|
|34.0|   0.0|40.5| 32.4|29.6|27.1| 5.8| 10.5|4.56| 91.0|26.6|72.0|
0.0|[34.0,0.0,40.5,32…|[298.285285431123…|[1.0,2.6943487118…|       0.0|
|34.0|   0.0|46.1| 70.6|35.8|30.0| 7.6|  7.7| 4.2| 93.0|14.3|78.7|
0.0|[34.0,0.0,46.1,70…|[332.472498609766…|[1.0,1.1235126500…|       0.0|
|35.0|   0.0|47.3| 92.2|30.7|25.7| 6.6|11.58| 5.9| 82.0|36.9|77.8|
0.0|[35.0,0.0,47.3,92…|[417.652369137353…|[1.0,5.1200365022…|       0.0|
|35.0|   0.0|48.7| 72.7|24.1|31.0|45.1|  9.4| 3.8| 90.0|20.0|75.8|
0.0|[35.0,0.0,48.7,72…|[280.898330170141…|[1.0,1.9821944238…|       0.0|
|36.0|   0.0|45.9| 58.8|29.7|27.7|11.7|  5.6|4.89| 93.0|23.1|70.8|
0.0|[36.0,0.0,45.9,58…|[312.341131137088…|[1.0,1.4393983397…|       0.0|
|36.0|   0.0|48.9| 82.8|16.9|24.4| 8.9| 8.91| 5.1| 97.0|14.8|79.9|
0.0|[36.0,0.0,48.9,82…|[397.198685530572…|[1.0,6.3206555501…|       0.0|
|37.0|   0.0|31.4|106.0|16.6|17.0| 2.4| 5.95| 5.3| 68.0|22.9|72.3|
0.0|[37.0,0.0,31.4,10…|[337.464456568178…|[1.0,3.9155153043…|       0.0|
|37.0|   0.0|41.5| 64.6|23.7|29.9| 9.3| 5.49|3.97|100.0|10.4|69.3|
0.0|[37.0,0.0,41.5,64…|[292.470031650232…|[1.0,3.1666875311…|       0.0|
|37.0|   0.0|46.1| 44.3|42.7|26.5| 6.4|10.86|5.05| 74.0|22.2|73.1|
0.0|[37.0,0.0,46.1,44…|[349.673660599920…|[1.0,4.8755600788…|       0.0|
|37.0|   0.0|46.4| 53.3|20.2|24.9| 8.7| 8.63| 5.9| 86.0|23.3|78.9|
0.0|[37.0,0.0,46.4,53…|[363.382047038556…|[1.0,2.1054673980…|       0.0|
|37.0|   0.0|47.9| 68.8|40.3|46.9| 6.0| 9.76|6.42| 81.0|22.7|80.6|
0.0|[37.0,0.0,47.9,68…|[382.596627540151…|[1.0,3.6134337666…|       0.0|
+----+------+----+-----+----+----+----+-----+----+-----+----+----+------+-------
-------------+--------------------+--------------------+----------+
only showing top 20 rows
```

[93]:
```
print(y_pred.columns)
```

```
['Age', 'Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA',
'GGT', 'PROT', 'Target', 'features', 'rawPrediction', 'probability',
'prediction']
```

[95]:
```
y_pred.select('target','rawPrediction', 'probability', 'prediction').show()
```

```
+------+--------------------+--------------------+----------+
|target|       rawPrediction|         probability|prediction|
+------+--------------------+--------------------+----------+
```

```
|   0.0|[286.954869615973…|[1.0,3.5781823002…|       0.0|
|   0.0|[281.614235265908…|[1.0,1.0244873107…|       0.0|
|   0.0|[329.637070801162…|[1.0,2.3941933511…|       0.0|
|   0.0|[375.018142931757…|[1.0,3.8219147599…|       0.0|
|   0.0|[266.834562998572…|[1.0,6.7251788183…|       0.0|
|   0.0|[282.416699150824…|[1.0,3.1777824373…|       0.0|
|   0.0|[340.661989788631…|[1.0,1.8429339582…|       0.0|
|   0.0|[287.262205429983…|[1.0,9.1668901248…|       0.0|
|   0.0|[319.289845055447…|[1.0,1.9214316512…|       0.0|
|   0.0|[298.285285431123…|[1.0,2.6943487118…|       0.0|
|   0.0|[332.472498609766…|[1.0,1.1235126500…|       0.0|
|   0.0|[417.652369137353…|[1.0,5.1200365022…|       0.0|
|   0.0|[280.898330170141…|[1.0,1.9821944238…|       0.0|
|   0.0|[312.341131137088…|[1.0,1.4393983397…|       0.0|
|   0.0|[397.198685530572…|[1.0,6.3206555501…|       0.0|
|   0.0|[337.464456568178…|[1.0,3.9155153043…|       0.0|
|   0.0|[292.470031650232…|[1.0,3.1666875311…|       0.0|
|   0.0|[349.673660599920…|[1.0,4.8755600788…|       0.0|
|   0.0|[363.382047038556…|[1.0,2.1054673980…|       0.0|
|   0.0|[382.596627540151…|[1.0,3.6134337666…|       0.0|
+------+------------------+------------------+----------+
only showing top 20 rows
```

Model Evaluation

[96]:
```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

[97]:
```
# How to Check For Accuracy
multi_evaluator =␣
 ↪MulticlassClassificationEvaluator(labelCol='Target',metricName='accuracy')
```

[98]:
```
multi_evaluator.evaluate(y_pred)
```

[98]:
```
0.9659090909090909
```

[ ]:
```
# Precision,F1 Score,Recall : Classification Report
```

[100]:
```
from pyspark.mllib.evaluation import MulticlassMetrics
```

[101]:
```
lr_metric = MulticlassMetrics(y_pred['target', 'prediction'].rdd)
```

[102]:
```
dir(lr_metric)
```

[102]:
```
['__class__',
 '__del__',
 '__delattr__',
 '__dict__',
```

```
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__lt__',
 '__module__',
 '__ne__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
 '__subclasshook__',
 '__weakref__',
 '_java_model',
 '_sc',
 'accuracy',
 'call',
 'confusionMatrix',
 'fMeasure',
 'falsePositiveRate',
 'logLoss',
 'precision',
 'recall',
 'truePositiveRate',
 'weightedFMeasure',
 'weightedFalsePositiveRate',
 'weightedPrecision',
 'weightedRecall',
 'weightedTruePositiveRate']
```

[103]:
```
print("Accuracy",lr_metric.accuracy)
```

```
Accuracy 0.9659090909090909
```

[105]:
```
print("Precision",lr_metric.precision(1.0))
print("Recall",lr_metric.recall(1.0))
print("F1Score",lr_metric.fMeasure(1.0))
```

```
Precision 0.8
Recall 0.8888888888888888
F1Score 0.8421052631578948
```

[106]:
```
dir(lr_model)
```

[106]:
```
['__class__',
 '__del__',
 '__delattr__',
 '__dict__',
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__lt__',
 '__metaclass__',
 '__module__',
 '__ne__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
 '__subclasshook__',
 '__weakref__',
 '_call_java',
 '_checkThresholdConsistency',
 '_copyValues',
 '_copy_params',
 '_create_from_java_class',
 '_create_params_from_java',
 '_defaultParamMap',
 '_dummy',
 '_empty_java_param_map',
 '_from_java',
 '_java_obj',
 '_make_java_param_pair',
 '_new_java_array',
```

```
'_new_java_obj',
'_paramMap',
'_params',
'_randomUID',
'_resetUid',
'_resolveParam',
'_set',
'_setDefault',
'_shouldOwn',
'_to_java',
'_transfer_param_map_from_java',
'_transfer_param_map_to_java',
'_transfer_params_from_java',
'_transfer_params_to_java',
'_transform',
'aggregationDepth',
'clear',
'coefficientMatrix',
'coefficients',
'copy',
'elasticNetParam',
'evaluate',
'explainParam',
'explainParams',
'extractParamMap',
'family',
'featuresCol',
'fitIntercept',
'getAggregationDepth',
'getElasticNetParam',
'getFamily',
'getFeaturesCol',
'getFitIntercept',
'getLabelCol',
'getLowerBoundsOnCoefficients',
'getLowerBoundsOnIntercepts',
'getMaxIter',
'getOrDefault',
'getParam',
'getPredictionCol',
'getProbabilityCol',
'getRawPredictionCol',
'getRegParam',
'getStandardization',
'getThreshold',
'getThresholds',
'getTol',
```

```
'getUpperBoundsOnCoefficients',
'getUpperBoundsOnIntercepts',
'getWeightCol',
'hasDefault',
'hasParam',
'hasSummary',
'intercept',
'interceptVector',
'isDefined',
'isSet',
'labelCol',
'load',
'lowerBoundsOnCoefficients',
'lowerBoundsOnIntercepts',
'maxIter',
'numClasses',
'numFeatures',
'params',
'predict',
'predictProbability',
'predictRaw',
'predictionCol',
'probabilityCol',
'rawPredictionCol',
'read',
'regParam',
'save',
'set',
'setFeaturesCol',
'setPredictionCol',
'setProbabilityCol',
'setRawPredictionCol',
'setThreshold',
'setThresholds',
'standardization',
'summary',
'threshold',
'thresholds',
'tol',
'transform',
'uid',
'upperBoundsOnCoefficients',
'upperBoundsOnIntercepts',
'weightCol',
'write']
```

[ ]:
```
# Saving Model
lr_model.save("lr_model_30")

lr_model.write().save("mylr_model")
```

[111]:
```
# Thanks For Watching
# Jesus Saves @JCharisTech
# By Jesse E.Agbe(JCharis)
```

[ ]:
```

```