MACHINE LEARNING WITH SPARK

Dosen Pengampu : Grezio Arifiyan Primajaya S.Kom, M.Kom

EKY FERNANDA | 3322600025

ML with PySpark

- Classify/Predict

Datasource

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

```
# Load our Pkgs
from pyspark import SparkContext
```

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

```
# Spark UI
sc
```

SparkContext

Spark UI

Version
    v3.5.3
Master
    local[2]
AppName
    pyspark-shell

```
# Load Pkgs
from pyspark.sql import SparkSession
```

```
# Spark
spark = SparkSession.builder.appName("MLwithSpark").getOrCreate()
```

WorkFlow

- Data Prep
- Feature Engineering
- Build Model
- Evaluate

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.

```
# Load our dataset
df = spark.read.csv("hcvdata.csv",header=True,inferSchema=True)
```

```
# 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
```

```
# check for columns
print(df.columns)
```

```
['_c0', 'Category', 'Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', '
```

```
# Rearrange
df = df.select('Age', 'Sex', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA', 'G
```

```
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
```

```
# Check for datatypes
# Before InferSchema=True
df.dtypes
```

```
[('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')]
```

```
# After InferSchema
df.dtypes
```

```
[('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')]
```

```
# 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)
```

```
# Descriptive summary
print(df.describe().show())
```

```
+-------+------------------+----+-----------------+------------------+---------------
|summary|               Age| Sex|              ALB|               ALP|
+-------+------------------+----+-----------------+------------------+---------------
|  count|               615| 615|              615|               615|
|   mean| 47.40813008130081|NULL|41.62019543973941| 68.28391959798999| 28.45081433224
| stddev|10.055105445519239|NULL|5.780629404103076|26.028315300123676|25.469688813870
|    min|                19|   f|             14.9|             100.4|
|    max|                77|   m|               NA|                NA|
+-------+------------------+----+-----------------+------------------+---------------

None
```

```
# 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
```

```
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
```

```
import pyspark.ml
```

```
dir(pyspark.ml)
```

```
['Estimator',
 'Model',
 'Pipeline',
 'PipelineModel',
 'PredictionModel',
 'Predictor',
 'TorchDistributor',
 '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',
 'torch',
 'tree',
 'tuning',
 'util',
 'wrapper']
```

```
# Load ML Pkgs
from pyspark.ml.feature import VectorAssembler,StringIndexer
```

```
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
```

```
# Unique Values for Sex
df.select('Sex').distinct().show()
```

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

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

```
df = genderEncoder.transform(df)
```

```
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
```

```
# Encoding for Category
```

```
# Label Encoding
catEncoder = StringIndexer(inputCol='Category',outputCol='Target').fit(df)
df = catEncoder.transform(df)
```

```
df.show(5)
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|     Category|Gender|Tar
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
| 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
```

```
# Get the labels
catEncoder.labels
```

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

```
# IndexToString
from pyspark.ml.feature import IndexToString
```

```
converter = IndexToString(inputCol='Target',outputCol='orig_cat')
```

```
converted_df = converter.transform(df)
```

```
converted_df.show()
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+--
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|     Category|Gender|Tar
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
```

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

```
### Feature
df.show()
```

```
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
|Age|Sex| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|     Category|Gender|Tar
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
| 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|
+---+---+----+----+----+----+----+-----+----+-----+----+----+-------------+------+---
only showing top 20 rows
```

```
print(df.columns)
```

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

```
df.dtypes
```

```
[('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')]
```

```
df2 = df.select('Age','Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA',
```

```
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'])
```

```
df2 = df2.toPandas().replace('NA',0).astype(float)
```

```
type(df2)
```

```
type(df)
```

```
# Convert To PySpark Dataframe
new_df = spark.createDataFrame(df2)
```

```
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
```

```
# 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)
```

```
required_features = ['Age','Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'C
```

```
# VectorAsm
vec_assembler = VectorAssembler(inputCols=required_features,outputCol='features')
```

```
vec_df = vec_assembler.transform(new_df)
```

```
vec_df.show(5)
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+-------------
| Age|Gender| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|            f
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+-------------
|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.
|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.
|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.
|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.
|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.
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+-------------
only showing top 5 rows
```

Train,Test Split

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

```
train_df.count()
```

```
432
```

```
train_df.show(4)
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+-------------
| Age|Gender| ALB| ALP| ALT| AST| BIL|  CHE|CHOL| CREA| GGT|PROT|Target|            f
```

```
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+------------
|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.
|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.
|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|[32.0,0.0,42.
|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.
+----+------+----+----+----+----+----+-----+----+-----+----+----+------+------------
only showing top 4 rows
```

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

```
from pyspark.ml.classification import LogisticRegression,DecisionTreeClassifier
```

```
# Logist Model
lr = LogisticRegression(featuresCol='features',labelCol='Target')
```

```
lr_model = lr.fit(train_df)
```

```
y_pred = lr_model.transform(test_df)
```

```
y_pred.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
|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
|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|[32.0,0.0,44
|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
|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
|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
|34.0|   0.0|29.0| 41.6|29.1|16.1| 4.8| 6.82|4.03| 62.0|14.5|53.2|   0.0|[34.0,0.0,29
|34.0|   0.0|43.6| 58.9|47.1|31.1|18.5| 9.14|4.99| 95.0|22.2|69.3|   0.0|[34.0,0.0,43
|34.0|   0.0|44.8| 77.7|36.9|31.0|19.5|10.51|5.59| 80.0|23.7|78.9|   0.0|[34.0,0.0,44
|35.0|   0.0|44.5| 70.3|26.2|25.1| 5.1|10.12|4.69| 82.0|20.7|67.2|   0.0|[35.0,0.0,44
|36.0|   0.0|42.6| 65.3|35.8|27.1|15.7|10.66|4.38| 96.0|34.7|71.0|   0.0|[36.0,0.0,42
|36.0|   0.0|46.1| 58.5|26.8|25.3| 6.0| 6.61|5.07| 71.0|10.5|79.6|   0.0|[36.0,0.0,46
|37.0|   0.0|40.8|118.9|17.2|19.2| 3.2| 9.17|4.26| 88.0|13.5|72.0|   0.0|[37.0,0.0,40
|37.0|   0.0|43.6| 72.8|51.4|43.7|13.8| 8.16|4.88| 70.0|94.5|75.2|   0.0|[37.0,0.0,43
|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
|37.0|   0.0|48.7| 62.3|21.0|21.1|41.9| 9.71|4.02| 84.0|16.0|75.1|   0.0|[37.0,0.0,48
|38.0|   0.0|40.5| 61.7|18.6|24.7| 6.7| 8.47|6.05| 89.0|19.6|75.6|   0.0|[38.0,0.0,40
|38.0|   0.0|42.0| 42.7|34.8|42.2| 3.3|  6.1|4.74| 96.0|14.6|66.7|   0.0|[38.0,0.0,42
|38.0|   0.0|44.7| 69.4|47.4|35.1|16.7|  6.9|4.14| 67.0|17.3|70.1|   0.0|[38.0,0.0,44
|39.0|   0.0|38.8| 52.5|54.3|31.3|10.1|10.68|6.26| 81.0|31.5|77.2|   0.0|[39.0,0.0,38
+----+------+----+-----+----+----+----+-----+----+-----+----+----+------+-----------
only showing top 20 rows
```

```
print(y_pred.columns)
```

```
['Age', 'Gender', 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA', 'GGT', 'P
```

```
y_pred.select('target','rawPrediction', 'probability', 'prediction').show()
```

```
+------+--------------------+--------------------+----------+
|target|       rawPrediction|         probability|prediction|
+------+--------------------+--------------------+----------+
|   0.0|[101.642773965286...|[1.0,1.2996356446...|       0.0|
|   0.0|[100.685400347765...|[1.0,2.9875755900...|       0.0|
|   0.0|[95.5397494065700...|[1.0,5.0698301252...|       0.0|
|   0.0|[108.023431331312...|[1.0,2.6499082283...|       0.0|
|   0.0|[90.7458371264038...|[1.0,2.9392236820...|       0.0|
|   0.0|[107.125675388156...|[1.0,1.1367421106...|       0.0|
|   0.0|[89.8138911596561...|[0.99999999999992...|       0.0|
|   0.0|[111.212065814535...|[1.0,4.4675788873...|       0.0|
|   0.0|[99.9904476721623...|[1.0,1.4964236132...|       0.0|
|   0.0|[118.089253726244...|[1.0,3.1657793756...|       0.0|
|   0.0|[103.996663549684...|[1.0,1.0199994422...|       0.0|
|   0.0|[106.293920030025...|[1.0,4.5660193153...|       0.0|
|   0.0|[124.395758477991...|[1.0,1.9894080921...|       0.0|
|   0.0|[84.8910388865561...|[1.0,1.1625273040...|       0.0|
|   0.0|[102.744211158330...|[1.0,8.8429670519...|       0.0|
|   0.0|[93.9041136697406...|[0.99999985749169...|       0.0|
|   0.0|[99.1396325342617...|[1.0,9.1969669784...|       0.0|
|   0.0|[114.710840838580...|[1.0,3.3847603408...|       0.0|
|   0.0|[111.063146444288...|[1.0,2.5865412225...|       0.0|
|   0.0|[90.8409206745444...|[1.0,8.7340976197...|       0.0|
+------+--------------------+--------------------+----------+
only showing top 20 rows
```

Model Evaluation

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

```
# How to Check For Accuracy
multi_evaluator = MulticlassClassificationEvaluator(labelCol='Target',metricName='accurac
```

```
multi_evaluator.evaluate(y_pred)
```

```
0.9180327868852459
```

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

```
from pyspark.mllib.evaluation import MulticlassMetrics
```

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

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/context.py:158: FutureWarning: De
  warnings.warn(
```

```
dir(lr_metric)
```

```
['__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']
```

```
print("Accuracy",lr_metric.accuracy)
```

```
Accuracy 0.9180327868852459
```

```
print("Precision",lr_metric.precision(1.0))
print("Recall",lr_metric.recall(1.0))
print("F1Score",lr_metric.fMeasure(1.0))
```

```
Precision 0.5
Recall 0.7142857142857143
F1Score 0.588235294117647
```

```
dir(lr_model)
```

```
['__abstractmethods__',
 '_annotations_',
 '__class__',
 '__class_getitem__',
 '_del_',
 '__delattr__',
 '_dict_',
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '_ge_',
 '__getattribute__',
 '_gt_',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '_le_',
 '__lt__',
 '__module__',
 '_ne_',
 '__new__',
 '__orig_bases__',
 '__parameters__',
 '_reduce_',
 '__reduce_ex__',
 '_repr_',
 '__setattr__',
 '__sizeof__',
 '_slots_',
 '__str__',
 '__subclasshook__',
 '_weakref_',
 '_abc_impl',
 '_call_java',
 '_checkThresholdConsistency',
 '_copyValues',
 '_copy_params',
```

```
        '_create_from_java_class',
        '_create_params_from_java',
        '_defaultParamMap',
        '_dummy',
        '_empty_java_param_map',
        '_from_java',
        '_is_protocol',
        '_java_obj',
        '_make_java_param_pair',
        '_new_java_array',
        '_new_java_obj',
        '_paramMap',
        '_params',
        '_randomUID',
        '_resetUid',
        '_resolveParam',
        '_set',
        '_setDefault',
        '_shouldOwn',
        '_testOwnParam',
```

```
# Saving Model
lr_model.save("lr_model_30")

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