```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-w
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 34 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 53.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e9
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('SparkML').getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.2.1

Master
local[*]

AppName
SparkML

```
df = spark.read.csv('a.csv',header=True,inferSchema=True)
```

```
df.show()
```

```
+----+----+-----------+-------+-------------+-------------+-------+----+------
|S.no|City|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|Diabet
+----+----+-----------+-------+-------------+-------------+-------+----+------
|   1|  MZ|          6|    148|           72|           35|      0|33.6|
|   2|  NY|          1|     85|           66|           29|      0|26.6|
|   3|  AG|          8|    183|           64|            0|      0|23.3|
|   4|  PU|          1|     89|           66|           23|     94|28.1|
|   5|  AR|          0|    137|           40|           35|    168|43.1|
|   6|  PY|          5|    116|           74|            0|      0|25.6|
|   7|  MY|          3|     78|           50|           32|     88|31.0|
|   8|  NZ|         10|    115|            0|            0|      0|35.3|
|   9|  AY|          2|    197|           70|           45|    543|30.5|
|  10|  BA|          8|    125|           96|            0|      0| 0.0|
|  11|  MZ|          4|    110|           92|            0|      0|37.6|
|  12|  NY|         10|    168|           74|            0|      0|38.0|
|  13|  AG|         10|    139|           80|            0|      0|27.1|
```

```
|  14|  PU|          1|    189|           60|           23|    846|30.1|
|  15|  AR|          5|    166|           72|           19|    175|25.8|
|  16|  PY|          7|    100|            0|            0|      0|30.0|
|  17|  MY|          0|    118|           84|           47|    230|45.8|
|  18|  NZ|          7|    107|           74|            0|      0|29.6|
|  19|  AY|          1|    103|           30|           38|     83|43.3|
|  20|  BA|          1|    115|           70|           30|     96|34.6|
+----+----+-----------+-------+-------------+-------------+-------+----+------
only showing top 20 rows
```

```
df=df.withColumnRenamed('Pregnancies','pg')
df=df.withColumnRenamed('Glucose','g')
df=df.withColumnRenamed('BloodPressure','bp')
df=df.withColumnRenamed('SkinThickness','st')
df=df.withColumnRenamed('Insulin','I')
df=df.withColumnRenamed('Diabetes PedigreeFunction','dbf')
```

```
df.head(3)
```

```
[Row(S.no=1, City='MZ', pg=6, g=148, bp=72, st=35, I=0, BMI=33.6, dbf=0.627, A
 Row(S.no=2, City='NY', pg=1, g=85, bp=66, st=29, I=0, BMI=26.6, dbf=0.351, Ag
 Row(S.no=3, City='AG', pg=8, g=183, bp=64, st=0, I=0, BMI=23.3, dbf=0.672, Ag
```

```
from pyspark.ml.feature import VectorAssembler
fa=VectorAssembler(inputCols=['pg','g','bp','st','I','BMI','dbf','Age'],outputCol='
```

```
df1=fa.transform(df)
```

```
df1.show()
```

```
+----+----+---+---+---+---+---+----+-----+---+-------+--------------------+
|S.no|City| pg|  g| bp| st|  I| BMI|  dbf|Age|Outcome|               newic|
+----+----+---+---+---+---+---+----+-----+---+-------+--------------------+
|   1|  MZ|  6|148| 72| 35|  0|33.6|0.627| 50|      1|[6.0,148.0,72.0,3...|
|   2|  NY|  1| 85| 66| 29|  0|26.6|0.351| 31|      0|[1.0,85.0,66.0,29...|
|   3|  AG|  8|183| 64|  0|  0|23.3|0.672| 32|      1|[8.0,183.0,64.0,0...|
|   4|  PU|  1| 89| 66| 23| 94|28.1|0.167| 21|      0|[1.0,89.0,66.0,23...|
|   5|  AR|  0|137| 40| 35|168|43.1|2.288| 33|      1|[0.0,137.0,40.0,3...|
|   6|  PY|  5|116| 74|  0|  0|25.6|0.201| 30|      0|[5.0,116.0,74.0,0...|
|   7|  MY|  3| 78| 50| 32| 88|31.0|0.248| 26|      1|[3.0,78.0,50.0,32...|
|   8|  NZ| 10|115|  0|  0|  0|35.3|0.134| 29|      0|[10.0,115.0,0.0,0...|
|   9|  AY|  2|197| 70| 45|543|30.5|0.158| 53|      1|[2.0,197.0,70.0,4...|
|  10|  BA|  8|125| 96|  0|  0| 0.0|0.232| 54|      1|[8.0,125.0,96.0,0...|
|  11|  MZ|  4|110| 92|  0|  0|37.6|0.191| 30|      0|[4.0,110.0,92.0,0...|
|  12|  NY| 10|168| 74|  0|  0|38.0|0.537| 34|      1|[10.0,168.0,74.0,...|
|  13|  AG| 10|139| 80|  0|  0|27.1|1.441| 57|      0|[10.0,139.0,80.0,...|
|  14|  PU|  1|189| 60| 23|846|30.1|0.398| 59|      1|[1.0,189.0,60.0,2...|
|  15|  AR|  5|166| 72| 19|175|25.8|0.587| 51|      1|[5.0,166.0,72.0,1...|
|  16|  PY|  7|100|  0|  0|  0|30.0|0.484| 32|      1|[7.0,100.0,0.0,0....|
|  17|  MY|  0|118| 84| 47|230|45.8|0.551| 31|      1|[0.0,118.0,84.0,4...|
|  18|  NZ|  7|107| 74|  0|  0|29.6|0.254| 31|      1|[7.0,107.0,74.0,0...|
|  19|  AY|  1|103| 30| 38| 83|43.3|0.183| 33|      0|[1.0,103.0,30.0,3...|
|  20|  BA|  1|115| 70| 30| 96|34.6|0.529| 32|      1|[1.0,115.0,70.0,3...|
+----+----+---+---+---+---+---+----+-----+---+-------+--------------------+
```

```
only showing top 20 rows
```

```
df2=df1.select("newic","Outcome")
```

```
op="Outcome"
```

```
df2.show()
```

```
+--------------------+-------+
|               newic|Outcome|
+--------------------+-------+
|[6.0,148.0,72.0,3...|      1|
|[1.0,85.0,66.0,29...|      0|
|[8.0,183.0,64.0,0...|      1|
|[1.0,89.0,66.0,23...|      0|
|[0.0,137.0,40.0,3...|      1|
|[5.0,116.0,74.0,0...|      0|
|[3.0,78.0,50.0,32...|      1|
|[10.0,115.0,0.0,0...|      0|
|[2.0,197.0,70.0,4...|      1|
|[8.0,125.0,96.0,0...|      1|
|[4.0,110.0,92.0,0...|      0|
|[10.0,168.0,74.0,...|      1|
|[10.0,139.0,80.0,...|      0|
|[1.0,189.0,60.0,2...|      1|
|[5.0,166.0,72.0,1...|      1|
|[7.0,100.0,0.0,0....|      1|
|[0.0,118.0,84.0,4...|      1|
|[7.0,107.0,74.0,0...|      1|
|[1.0,103.0,30.0,3...|      0|
|[1.0,115.0,70.0,3...|      1|
+--------------------+-------+
only showing top 20 rows
```

```
df.take(3)
```

```
[Row(S.no=1, City='MZ', pg=6, g=148, bp=72, st=35, I=0, BMI=33.6, dbf=0.627, A
 Row(S.no=2, City='NY', pg=1, g=85, bp=66, st=29, I=0, BMI=26.6, dbf=0.351, Ag
 Row(S.no=3, City='AG', pg=8, g=183, bp=64, st=0, I=0, BMI=23.3, dbf=0.672, Ag
```

```
from pyspark.sql.functions import *
print(df.stat.corr('pg','Outcome'))
print(df.stat.corr('g','Outcome'))
print(df.stat.corr('bp','Outcome'))
print(df.stat.corr('st','Outcome'))
print(df.stat.corr('i','Outcome'))
print(df.stat.corr('BMI','Outcome'))
print(df.stat.corr('dbf','Outcome'))
print(df.stat.corr('Age','Outcome'))
print(df.stat.corr('Outcome','Outcome'))
```

```
0.22189815303398636
0.4665813983068737
0.06506835955033274
0.07475223191831945
0.13054795488404794
0.2926946626444454
0.17384406565296
0.23835598302719757
1.0
```

```
from pyspark.ml.classification import LogisticRegression
train_data,test_data=df2.randomSplit([0.75,0.25])
applyml=LogisticRegression(featuresCol='newic', labelCol='Outcome')
applyml=applyml.fit(train_data)
```

```
predict=applyml.evaluate(test_data)
```

```
predict.predictions.show()
```

```
/usr/local/lib/python3.7/dist-packages/pyspark/sql/context.py:127: FutureWarni
  FutureWarning
+--------------------+-------+--------------------+--------------------+------
|               newic|Outcome|       rawPrediction|         probability|predic
+--------------------+-------+--------------------+--------------------+------
|(8,[0,1,6,7],[6.0...|      0|[3.48049530112385...|[0.97012767734942...|
|(8,[0,1,6,7],[7.0...|      0|[3.58638078896978...|[0.97304812803675...|
|(8,[0,1,6,7],[10....|      1|[2.86938058451299...|[0.94631188685415...|
|(8,[1,5,6,7],[99....|      0|[2.44409267267169...|[0.92012838474531...|
|(8,[1,5,6,7],[131...|      1|[-0.3302959557643...|[0.41816861430464...|
|(8,[1,5,6,7],[167...|      1|[-1.2887017112668...|[0.21607264072567...|
|[0.0,67.0,76.0,0....|      0|[2.06199046376390...|[0.88715359303877...|
|[0.0,91.0,68.0,32...|      0|[1.98163445655938...|[0.87885528830926...|
|[0.0,93.0,60.0,25...|      0|[2.55547677613581...|[0.92794058882718...|
|[0.0,102.0,52.0,0...|      0|[2.97652068185305...|[0.95150206705714...|
|[0.0,102.0,78.0,4...|      0|[2.10319124924338...|[0.89121296462128...|
|[0.0,102.0,86.0,1...|      0|[2.21280874237927...|[0.90139385779991...|
|[0.0,105.0,64.0,4...|      0|[1.45243750281390...|[0.81037328460339...|
|[0.0,105.0,68.0,2...|      0|[3.16331825750610...|[0.95943030269595...|
|[0.0,105.0,84.0,0...|      1|[1.49886932271716...|[0.81740577913167...|
|[0.0,106.0,70.0,3...|      0|[1.21124988848726...|[0.77052002769461...|
|[0.0,111.0,65.0,0...|      0|[2.02355612093751...|[0.88324821924414...|
|[0.0,114.0,80.0,3...|      0|[1.23201851076324...|[0.77417166558724...|
|[0.0,117.0,66.0,3...|      0|[1.77202452529322...|[0.85470925991321...|
|[0.0,117.0,80.0,3...|      0|[0.81348885382691...|[0.69285245919095...|
+--------------------+-------+--------------------+--------------------+------
only showing top 20 rows
```

```
predict.accuracy
```

```
0.7872340425531915
```

✓ 0s completed at 3:15 PM