```
!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

Saved successfully!

v3.2.1

Master

local[*]

AppName

SparkML

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

```
df.printSchema()
```

```
root
 |-- sepal_length: double (nullable = true)
 |-- sepal_width: double (nullable = true)
 |-- petal_length: double (nullable = true)
 |-- petal_width: double (nullable = true)
 |-- species: string (nullable = true)
```

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

```
+------------+-----------+------------+-----------+-----------+
|sepal_length|sepal_width|petal_length|petal_width|    species|
+------------+-----------+------------+-----------+-----------+
|         5.1|        3.5|         1.4|        0.2|Iris-setosa|
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|
```

```
|          4.7|          3.2|          1.3|          0.2|Iris-setosa|
|          4.6|          3.1|          1.5|          0.2|Iris-setosa|
|          5.0|          3.6|          1.4|          0.2|Iris-setosa|
+-------------+-------------+-------------+-------------+-----------+
only showing top 5 rows
```

```
#Preprocessing steps
from pyspark.ml.feature import StringIndexer, OneHotEncoder

# create object of StringIndexer class and specify input and output column
SI_species = StringIndexer(inputCol='species',outputCol='species_Index')
```

```
# transform the data
df = SI_species.fit(df).transform(df)
```

```
df.tail(5)
```

```
[Row(sepal_length=6.7, sepal_width=3.0, petal_length=5.2, petal_width=2.3, spe
 Row(sepal_length=6.3, sepal_width=2.5, petal_length=5.0, petal_width=1.9, spe
 Row(sepal_length=6.5, sepal_width=3.0, petal_length=5.2, petal_width=2.0, spe
 Row(sepal_length=6.2, sepal_width=3.4, petal_length=5.4, petal_width=2.3, spe
 Row(sepal_length=5.9, sepal_width=3.0, petal_length=5.1, petal_width=1.8, spe
```

Saved successfully!

```
df.printSchema()
```

```
root
 |-- sepal_length: double (nullable = true)
 |-- sepal_width: double (nullable = true)
 |-- petal_length: double (nullable = true)
 |-- petal_width: double (nullable = true)
 |-- species: string (nullable = true)
 |-- species_Index: double (nullable = false)
```

```
from pyspark.ml.feature import VectorAssembler
fa1=VectorAssembler(inputCols=['sepal_length','sepal_width','petal_length','petal_w
```

```
df1=fa1.transform(df)
```

```
df1.show()
```

```
+------------+-----------+------------+-----------+-----------+-------------+-
|sepal_length|sepal_width|petal_length|petal_width|    species|species_Index|
+------------+-----------+------------+-----------+-----------+-------------+-
|         5.1|        3.5|         1.4|        0.2|Iris-setosa|          0.0|[
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|          0.0|[
|         4.7|        3.2|         1.3|        0.2|Iris-setosa|          0.0|[
|         4.6|        3.1|         1.5|        0.2|Iris-setosa|          0.0|[
|         5.0|        3.6|         1.4|        0.2|Iris-setosa|          0.0|[
```

```
|          5.4|          3.9|          1.7|          0.4|Iris-setosa|           0.0|[
|          4.6|          3.4|          1.4|          0.3|Iris-setosa|           0.0|[
|          5.0|          3.4|          1.5|          0.2|Iris-setosa|           0.0|[
|          4.4|          2.9|          1.4|          0.2|Iris-setosa|           0.0|[
|          4.9|          3.1|          1.5|          0.1|Iris-setosa|           0.0|[
|          5.4|          3.7|          1.5|          0.2|Iris-setosa|           0.0|[
|          4.8|          3.4|          1.6|          0.2|Iris-setosa|           0.0|[
|          4.8|          3.0|          1.4|          0.1|Iris-setosa|           0.0|[
|          4.3|          3.0|          1.1|          0.1|Iris-setosa|           0.0|[
|          5.8|          4.0|          1.2|          0.2|Iris-setosa|           0.0|[
|          5.7|          4.4|          1.5|          0.4|Iris-setosa|           0.0|[
|          5.4|          3.9|          1.3|          0.4|Iris-setosa|           0.0|[
|          5.1|          3.5|          1.4|          0.3|Iris-setosa|           0.0|[
|          5.7|          3.8|          1.7|          0.3|Iris-setosa|           0.0|[
|          5.1|          3.8|          1.5|          0.3|Iris-setosa|           0.0|[
+-----------+-----------+-----------+-----------+-----------+--------------+-
only showing top 20 rows
```

```
df2=df1.select("newic1","species_Index")
```

```
df2.show()
```

```
+-----------------+-------------+
|           newic1|species_Index|
+-----------------+-------------+
|[5.1,3.5,1.4,0.2]|          0.0|
|[4.9,3.0,1.4,0.2]|          0.0|
|[4.7,3.2,1.3,0.2]|          0.0|
|[4.6,3.1,1.5,0.2]|          0.0|
|[5.0,3.6,1.4,0.2]|          0.0|
|[5.4,3.9,1.7,0.4]|          0.0|
|[4.6,3.4,1.4,0.3]|          0.0|
|[5.0,3.4,1.5,0.2]|          0.0|
|[4.4,2.9,1.4,0.2]|          0.0|
|[4.9,3.1,1.5,0.1]|          0.0|
|[5.4,3.7,1.5,0.2]|          0.0|
|[4.8,3.4,1.6,0.2]|          0.0|
|[4.8,3.0,1.4,0.1]|          0.0|
|[4.3,3.0,1.1,0.1]|          0.0|
|[5.8,4.0,1.2,0.2]|          0.0|
|[5.7,4.4,1.5,0.4]|          0.0|
|[5.4,3.9,1.3,0.4]|          0.0|
|[5.1,3.5,1.4,0.3]|          0.0|
|[5.7,3.8,1.7,0.3]|          0.0|
|[5.1,3.8,1.5,0.3]|          0.0|
+-----------------+-------------+
only showing top 20 rows
```

Saved successfully!

```
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'))
```

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

```
from pyspark.ml.classification import RandomForestClassifier
train_data,test_data=df2.randomSplit([0.75,0.25])
applyml=RandomForestClassifier(featuresCol='newic1', labelCol='species_Index')
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
+-----------------+-------------+--------------+-----------------+----------+
|           newic1|species_Index| rawPrediction|      probability|prediction|
+-----------------+-------------+--------------+-----------------+----------+
|[4.4,3.0,1.3,0.2]|          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|                 |          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|                 |          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|                 |          1.0|[0.0,20.0,0.0]|    [0.0,1.0,0.0]|       1.0|
|[5.0,3.2,1.2,0.2]|          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|[5.0,3.5,1.6,0.6]|          0.0|[17.0,3.0,0.0]|  [0.85,0.15,0.0]|       0.0|
|[5.1,3.8,1.5,0.3]|          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|[5.2,4.1,1.5,0.1]|          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|[5.3,3.7,1.5,0.2]|          0.0|[20.0,0.0,0.0]|    [1.0,0.0,0.0]|       0.0|
|[5.4,3.4,1.5,0.4]|          0.0|[18.0,2.0,0.0]|    [0.9,0.1,0.0]|       0.0|
|[5.4,3.9,1.3,0.4]|          0.0|[18.0,2.0,0.0]|    [0.9,0.1,0.0]|       0.0|
|[5.5,3.5,1.3,0.2]|          0.0|[19.0,1.0,0.0]|  [0.95,0.05,0.0]|       0.0|
|[5.6,3.0,4.1,1.3]|          1.0|[0.0,19.0,1.0]|  [0.0,0.95,0.05]|       1.0|
|[5.7,2.5,5.0,2.0]|          2.0|[0.0,3.5,16.5]|[0.0,0.175,0.825]|       2.0|
|[5.7,3.0,4.2,1.2]|          1.0|[0.0,20.0,0.0]|    [0.0,1.0,0.0]|       1.0|
|[5.7,3.8,1.7,0.3]|          0.0|[12.0,8.0,0.0]|    [0.6,0.4,0.0]|       0.0|
|[5.7,4.4,1.5,0.4]|          0.0|[12.0,8.0,0.0]|    [0.6,0.4,0.0]|       0.0|
|[5.8,2.6,4.0,1.2]|          1.0|[0.0,20.0,0.0]|    [0.0,1.0,0.0]|       1.0|
|[5.8,2.8,5.1,2.4]|          2.0|[0.0,1.5,18.5]|[0.0,0.075,0.925]|       2.0|
|[5.9,3.0,4.2,1.5]|          1.0|[0.0,20.0,0.0]|    [0.0,1.0,0.0]|       1.0|
+-----------------+-------------+--------------+-----------------+----------+
only showing top 20 rows
```



```
predict.accuracy
```

```
0.9347826086956522
```

0s completed at 4:56 PM

Saved successfully!