In [1]:
```
import findspark
findspark.init()
```

In [2]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("irisApp").getOrCreate()
```

In [3]:
```
df=spark.read.option("header","true").csv('D:\iris.csv',inferSchema=True)
df.show()
df.count()
```

```
+-------------+------------+-------------+------------+-------+
|SepalLength1|SepalWidth1|PetalLength1|PetalWidth1|Species|
+-------------+------------+-------------+------------+-------+
|          5.1|         3.5|          1.4|         0.2| Setosa|
|          4.9|         3.0|          1.4|         0.2| Setosa|
|          4.7|         3.2|          1.3|         0.2| Setosa|
|          4.6|         3.1|          1.5|         0.2| Setosa|
|          5.0|         3.6|          1.4|         0.2| Setosa|
|          5.4|         3.9|          1.7|         0.4| Setosa|
|          4.6|         3.4|          1.4|         0.3| Setosa|
|          5.0|         3.4|          1.5|         0.2| Setosa|
|          4.4|         2.9|          1.4|         0.2| Setosa|
|          4.9|         3.1|          1.5|         0.1| Setosa|
|          5.4|         3.7|          1.5|         0.2| Setosa|
|          4.8|         3.4|          1.6|         0.2| Setosa|
|          4.8|         3.0|          1.4|         0.1| Setosa|
|          4.3|         3.0|          1.1|         0.1| Setosa|
|          5.8|         4.0|          1.2|         0.2| Setosa|
|          5.7|         4.4|          1.5|         0.4| Setosa|
|          5.4|         3.9|          1.3|         0.4| Setosa|
|          5.1|         3.5|          1.4|         0.3| Setosa|
|          5.7|         3.8|          1.7|         0.3| Setosa|
|          5.1|         3.8|          1.5|         0.3| Setosa|
+-------------+------------+-------------+------------+-------+
only showing top 20 rows
```

Out[3]:
```
150
```

In [4]:
```
df.printSchema()
```

```
root
 |-- SepalLength: double (nullable = true)
 |-- SepalWidth: double (nullable = true)
 |-- PetalLength: double (nullable = true)
 |-- PetalWidth: double (nullable = true)
 |-- Species: string (nullable = true)
```

In [5]:
```
df.na.drop(how='any')
```

Out[5]:
```
DataFrame[SepalLength: double, SepalWidth: double, PetalLength: double, PetalWidth: double, Species: string]
```

In [6]:
```
df.count()
```

Out[6]:
```
150
```

In [7]:
```
df.select('Species').distinct().show()
```

```
+----------+
|   Species|
+----------+
| Virginica|
|    Setosa|
|Versicolor|
+----------+
```

In [8]:
```
df.columns[0:4]
```

Out[8]:
```
['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth']
```

In [9]:
```
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.feature import StringIndexer
```

In [10]:
```
vavVectorAssembler(inputCols=df.columns[:4],outputCol='Input Features')
indexer=StringIndexer(inputCol='Species',outputCol='Species_Data')
df3=indexer.fit(df).transform(df)
df2=va.transform(df3)
df2.show()
```

```
+-------------+------------+-------------+------------+----------+--------------------+-----------------+
|SepalLength1|SepalWidth1|PetalLength1|PetalWidth1|   Species|      Species_Data|   Input Features|
+-------------+------------+-------------+------------+----------+--------------------+-----------------+
|          5.1|         3.5|          1.4|         0.2|    Setosa|                 0.0|[5.1,3.5,1.4,0.2]|
|          4.9|         3.0|          1.4|         0.2|    Setosa|                 0.0|[4.9,3.0,1.4,0.2]|
|          4.7|         3.2|          1.3|         0.2|    Setosa|                 0.0|[4.7,3.2,1.3,0.2]|
|          4.6|         3.1|          1.5|         0.2|    Setosa|                 0.0|[4.6,3.1,1.5,0.2]|
|          5.0|         3.6|          1.4|         0.2|    Setosa|                 0.0|[5.0,3.6,1.4,0.2]|
|          5.4|         3.9|          1.7|         0.4|    Setosa|                 0.0|[5.4,3.9,1.7,0.4]|
|          4.6|         3.4|          1.4|         0.3|    Setosa|                 0.0|[4.6,3.4,1.4,0.3]|
|          5.0|         3.4|          1.5|         0.2|    Setosa|                 0.0|[5.0,3.4,1.5,0.2]|
|          4.4|         2.9|          1.4|         0.2|    Setosa|                 0.0|[4.4,2.9,1.4,0.2]|
|          4.9|         3.1|          1.5|         0.1|    Setosa|                 0.0|[4.9,3.1,1.5,0.1]|
|          5.4|         3.7|          1.5|         0.2|    Setosa|                 0.0|[5.4,3.7,1.5,0.2]|
|          4.8|         3.4|          1.6|         0.2|    Setosa|                 0.0|[4.8,3.4,1.6,0.2]|
|          4.8|         3.0|          1.4|         0.1|    Setosa|                 0.0|[4.8,3.0,1.4,0.1]|
|          4.3|         3.0|          1.1|         0.1|    Setosa|                 0.0|[4.3,3.0,1.1,0.1]|
|          5.8|         4.0|          1.2|         0.2|    Setosa|                 0.0|[5.8,4.0,1.2,0.2]|
|          5.7|         4.4|          1.5|         0.4|    Setosa|                 0.0|[5.7,4.4,1.5,0.4]|
|          5.4|         3.9|          1.3|         0.4|    Setosa|                 0.0|[5.4,3.9,1.3,0.4]|
|          5.1|         3.5|          1.4|         0.3|    Setosa|                 0.0|[5.1,3.5,1.4,0.3]|
|          5.7|         3.8|          1.7|         0.3|    Setosa|                 0.0|[5.7,3.8,1.7,0.3]|
|          5.1|         3.8|          1.5|         0.3|    Setosa|                 0.0|[5.1,3.8,1.5,0.3]|
+-------------+------------+-------------+------------+----------+--------------------+-----------------+
only showing top 20 rows
```

In [11]:
```
finaldata=df2.select('Input Features','Species_Data')
finaldata.show()
```

```
+-----------------+------------+
|  Input Features|Species_Data|
+-----------------+------------+
|[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
```

In [12]:
```
train,test=finaldata.randomSplit([0.70,0.30])
```

In [13]:
```
# Decision tree classification
from pyspark.ml.classification import DecisionTreeClassifier
```

In [14]:
```
dtcmodel=DecisionTreeClassifier(labelCol='Species_Data',featuresCol='Input Features')
model=dtcmodel.fit(train)
```

In [15]:
```
model
```

Out[15]:
```
DecisionTreeClassificationModel: uid=DecisionTreeClassifier_45189a09aad7, depth=5, numNodes=15, numClasses=3, numFeatures=4
```

In [25]:
```
prediction_res=model.transform(test)
prediction_res.show()
prediction_res.printSchema()
```

```
+-----------------+------------+--------------+-------------+----------+
|  Input Features|Species_Data| rawPrediction|  probability|prediction|
+-----------------+------------+--------------+-------------+----------+
|[4.4,3.0,1.3,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[4.6,3.1,1.5,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[4.9,2.5,4.5,1.7]|         2.0| [0.0,1.0,0.0]|[0.0,1.0,0.0]|       1.0|
|[4.9,3.1,1.5,0.1]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[4.9,3.1,1.5,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[4.9,3.6,1.4,0.1]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.0,3.0,1.6,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.0,3.4,1.6,0.4]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.0,3.5,1.6,0.6]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.1,3.7,1.5,0.4]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.2,3.4,1.4,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.2,3.4,1.4,0.3]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.3,3.7,1.5,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.4,3.4,1.7,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.4,3.9,1.7,0.4]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.5,2.4,3.8,1.1]|         1.0| [0.0,32.0,0.0]|[0.0,1.0,0.0]|       1.0|
|[5.5,2.4,3.8,1.1]|         1.0| [0.0,32.0,0.0]|[0.0,1.0,0.0]|       1.0|
|[5.5,2.6,4.4,1.2]|         1.0| [0.0,32.0,0.0]|[0.0,1.0,0.0]|       1.0|
|[5.5,3.5,1.3,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.5,4.2,1.4,0.2]|         0.0|[35.0,0.0,0.0]|[1.0,0.0,0.0]|       0.0|
|[5.6,2.5,3.9,1.1]|         1.0| [0.0,32.0,0.0]|[0.0,1.0,0.0]|       1.0|
|[5.6,2.9,3.6,1.3]|         1.0| [0.0,32.0,0.0]|[0.0,1.0,0.0]|       1.0|
+-----------------+------------+--------------+-------------+----------+
only showing top 20 rows

root
 |-- Input Features: vector (nullable = true)
 |-- Species_Data: double (nullable = false)
 |-- rawPrediction: vector (nullable = true)
 |-- probability: vector (nullable = true)
 |-- prediction: double (nullable = false)
```

In [26]:
```
prediction_res.select('Input Features','Species_Data','prediction').show()
```

```
+-----------------+------------+----------+
|  Input Features|Species_Data|prediction|
+-----------------+------------+----------+
|[4.4,3.0,1.3,0.2]|         0.0|       0.0|
|[4.6,3.1,1.5,0.2]|         0.0|       0.0|
|[4.9,2.5,4.5,1.7]|         2.0|       1.0|
|[4.9,3.1,1.5,0.1]|         0.0|       0.0|
|[4.9,3.1,1.5,0.2]|         0.0|       0.0|
|[4.9,3.6,1.4,0.1]|         0.0|       0.0|
|[5.0,3.0,1.6,0.2]|         0.0|       0.0|
|[5.0,3.4,1.6,0.4]|         0.0|       0.0|
|[5.0,3.5,1.6,0.6]|         0.0|       0.0|
|[5.1,3.7,1.5,0.4]|         0.0|       0.0|
|[5.2,3.4,1.4,0.2]|         0.0|       0.0|
|[5.2,3.4,1.4,0.3]|         0.0|       0.0|
|[5.3,3.7,1.5,0.2]|         0.0|       0.0|
|[5.4,3.4,1.7,0.2]|         0.0|       0.0|
|[5.4,3.9,1.7,0.4]|         0.0|       0.0|
|[5.5,2.4,3.8,1.1]|         1.0|       1.0|
|[5.5,2.4,3.8,1.1]|         1.0|       1.0|
|[5.5,2.6,4.4,1.2]|         1.0|       1.0|
|[5.5,3.5,1.3,0.2]|         0.0|       0.0|
|[5.5,4.2,1.4,0.2]|         0.0|       0.0|
|[5.6,2.5,3.9,1.1]|         1.0|       1.0|
|[5.6,2.9,3.6,1.3]|         1.0|       1.0|
+-----------------+------------+----------+
only showing top 20 rows
```

In [27]:
```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

In [28]:
```
evaluator=MulticlassClassificationEvaluator(labelCol='Species_Data',predictionCol='prediction')
accuracy=evaluator.evaluate(prediction_res)
print("Accuracy of model ",accuracy)
print("Error of model ",(1-accuracy))
```

```
Accuracy of model  0.9767738869583296
Error of model  0.023226113041670438
```

In [43]:
```
from pyspark.ml.feature import IndexToString
f=prediction_res.select('Input Features','Species_Data','prediction')
itos=IndexToString(inputCol='Species_Data',outputCol='SpeciesCategory')
c=itos.transform(f)
c.show()
```

```
+-----------------+------------+----------+---------------+
|  Input Features|Species_Data|prediction|SpeciesCategory|
+-----------------+------------+----------+---------------+
|[4.4,3.0,1.3,0.2]|         0.0|       0.0|         Setosa|
|[4.6,3.1,1.5,0.2]|         0.0|       0.0|         Setosa|
|[4.9,2.5,4.5,1.7]|         2.0|       1.0|     Virginica|
|[4.9,3.1,1.5,0.1]|         0.0|       0.0|         Setosa|
|[4.9,3.1,1.5,0.2]|         0.0|       0.0|         Setosa|
|[4.9,3.6,1.4,0.1]|         0.0|       0.0|         Setosa|
|[5.0,3.0,1.6,0.2]|         0.0|       0.0|         Setosa|
|[5.0,3.4,1.6,0.4]|         0.0|       0.0|         Setosa|
|[5.0,3.5,1.6,0.6]|         0.0|       0.0|         Setosa|
|[5.1,3.7,1.5,0.4]|         0.0|       0.0|         Setosa|
|[5.2,3.4,1.4,0.2]|         0.0|       0.0|         Setosa|
|[5.2,3.4,1.4,0.3]|         0.0|       0.0|         Setosa|
|[5.3,3.7,1.5,0.2]|         0.0|       0.0|         Setosa|
|[5.4,3.4,1.7,0.2]|         0.0|       0.0|         Setosa|
|[5.4,3.9,1.7,0.4]|         0.0|       0.0|         Setosa|
|[5.5,2.4,3.8,1.1]|         1.0|       1.0|    Versicolor|
|[5.5,2.4,3.8,1.1]|         1.0|       1.0|    Versicolor|
|[5.5,2.6,4.4,1.2]|         1.0|       1.0|    Versicolor|
|[5.5,3.5,1.3,0.2]|         0.0|       0.0|         Setosa|
|[5.5,4.2,1.4,0.2]|         0.0|       0.0|         Setosa|
|[5.6,2.5,3.9,1.1]|         1.0|       1.0|    Versicolor|
|[5.6,2.9,3.6,1.3]|         1.0|       1.0|    Versicolor|
+-----------------+------------+----------+---------------+
only showing top 20 rows
```

In [44]:
```
itos=IndexToString(inputCol='Species_Data',outputCol='SpeciesCategory')
c=itos.transform(df1)
c.select('Species_Data','SpeciesCategory').distinct().show()
```

```
+------------+---------------+
|Species_Data|SpeciesCategory|
+------------+---------------+
|         1.0|     Versicolor|
|         0.0|         Setosa|
|         2.0|      Virginica|
+------------+---------------+
```

In [45]:
```
from pyspark.ml.classification import RandomForestClassifier
```

In [46]:
```
dtcmodel=RandomForestClassifier(labelCol='Species_Data',featuresCol='Input Features')
model=dtcmodel.fit(train)
```

In [47]:
```
prediction_result=model.transform(test)
prediction_result.show()
```

```
+-----------------+------------+--------------------+--------------------+----------+
|  Input Features|Species_Data|       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|
|[4.6,3.1,1.5,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[4.9,2.5,4.5,1.7]|         2.0|[1.0,16.921568027...|[0.05,0.8407841...|       1.0|
|[4.9,3.1,1.5,0.1]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[4.9,3.1,1.5,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[4.9,3.6,1.4,0.1]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.0,3.0,1.6,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.0,3.4,1.6,0.4]|         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|[14.0,5.921568027...|[0.7,0.2960784313...|       0.0|
|[5.1,3.7,1.5,0.4]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.1,3.8,1.9,0.4]|         0.0|      [18.0,2.0,0.0]|      [0.9,0.1,0.0]|       0.0|
|[5.4,3.7,1.5,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.4,3.9,1.3,0.4]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.5,2.3,4.0,1.3]|         1.0|[0.0,9.1921568027...|[0.0,0.9919784313...|       1.0|
|[5.5,2.4,3.7,1.0]|         1.0|[0.0,9.1921568027...|[0.0,0.9919784313...|       1.0|
|[5.5,2.4,3.8,1.1]|         1.0|[0.0,9.1921568027...|[0.0,0.9919784313...|       1.0|
|[5.5,3.5,1.3,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.5,4.2,1.4,0.2]|         0.0|      [20.0,0.0,0.0]|      [1.0,0.0,0.0]|       0.0|
|[5.6,2.5,3.9,1.1]|         1.0|[0.0,9.1921568027...|[0.0,0.9919784313...|       1.0|
|[5.6,2.9,3.6,1.3]|         1.0|[0.0,9.1921568027...|[0.0,0.9919784313...|       1.0|
+-----------------+------------+--------------------+--------------------+----------+
only showing top 20 rows
```

In [48]:
```
evaluator=MulticlassClassificationEvaluator(labelCol='Species_Data',predictionCol='prediction')
accuracy=evaluator.evaluate(prediction_result)
print("Accuracy of model ",accuracy)
print("Error of model ",(1-accuracy))
```

```
Accuracy of model  0.9767738869583296
Error of model  0.023226113041670438
```

In [ ]:
```

```