Example of Pyspark ML

In [1]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('Missing').getOrCreate()
```

In [2]:
```
### Read the dataset
training = spark.read.csv('test3.csv', header=True, inferSchema=True)
```

In [3]:
```
training.show()
```

```
+-------+---+----------+------+
|   name|age|experience|salary|
+-------+---+----------+------+
|Ajinkya| 32|        10| 30000|
|  Anish| 30|         8| 25000|
| Nikhil| 29|         4| 20000|
| Hitesh| 24|         3| 20000|
|  Onkar| 21|         1| 15000|
|  Ketan| 23|         2| 18000|
+-------+---+----------+------+
```

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

```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- experience: integer (nullable = true)
 |-- salary: integer (nullable = true)
```

In [5]:
```
training.columns
```

Out[5]:
```
['name', 'age', 'experience', 'salary']
```

In [ ]:
```
[age, experince]------> new feature -----------> independent feature
```

In [72]:
```
from pyspark.ml.feature import VectorAssembler

featureassembler = VectorAssembler(inputCols=['age', 'experience'], outputCol='independent featu
```

In [73]:
```
output = featureassembler.transform(training)
```

In [74]:
```
output.show()
```

```
+-------+---+----------+------+-------------------+
|   name|age|experience|salary|independent features|
+-------+---+----------+------+-------------------+
|Ajinkya| 32|        10| 30000|         [32.0,10.0]|
|  Anish| 30|         8| 25000|          [30.0,8.0]|
| Nikhil| 29|         4| 20000|          [29.0,4.0]|
| Hitesh| 24|         3| 20000|          [24.0,3.0]|
|  Onkar| 21|         1| 15000|          [21.0,1.0]|
|  Ketan| 23|         2| 18000|          [23.0,2.0]|
+-------+---+----------+------+-------------------+
```

In [75]:
```
output.columns
```

Out[75]:
```
['name', 'age', 'experience', 'salary', 'independent features']
```

In [76]:
```
finalized_data = output.select('independent features', 'salary')
```

In [77]:
```
finalized_data.show()
```

```
+-------------------+------+
|independent features|salary|
+-------------------+------+
|         [32.0,10.0]| 30000|
|          [30.0,8.0]| 25000|
|          [29.0,4.0]| 20000|
|          [24.0,3.0]| 20000|
|          [21.0,1.0]| 15000|
|          [23.0,2.0]| 18000|
+-------------------+------+
```

In [114…
```
from pyspark.ml.regression import LinearRegression
##train test split
train_data,test_data = finalized_data.randomSplit([0.75, 0.25])
regressor = LinearRegression(featuresCol='independent features', labelCol='salary')
regressor = regressor.fit(train_data)
```

In [115…
```
### Coefficients
regressor.coefficients
```

Out[115]:
```
DenseVector([-66.2894, 1682.2959])
```

In [116…
```
### Intercepts
regressor.intercept
```

Out[116]:
```
15340.339531123198
```

In [117…
```
### Prediction
pred_results = regressor.evaluate(test_data)
```

In [118…
```
pred_results.predictions.show()
```

```
+-------------------+------+------------------+
|independent features|salary|        prediction|
+-------------------+------+------------------+
|          [24.0,3.0]| 20000| 18796.28132578818|
|          [30.0,8.0]| 25000|26810.024252223106|
+-------------------+------+------------------+
```

In [119…
```
pred_results.meanAbsoluteError, pred_results.meanSquaredError
```

Out[119]:
```
(1506.8714632174633, 2362563.2201410383)
```

In [ ]:
```

```