```
!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 32 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 51.4 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()
```

```
df = spark.read.csv('new1.csv',header=True,inferSchema=True)
```

```
df.show()
```

```
+-----+----+---+------+------+
|empid|name|age|deptid|salary|
+-----+----+---+------+------+
|  110| aaa| 35|     1|  2000|
|  111| bbb| 45|     2|  3000|
|  112| aaa| 50|     1|  2500|
|  113| bbb| 35|     2|  3000|
|  104| aaa| 40|     1|  2000|
|  105| bbb| 40|     2|  3500|
|  106| aaa| 35|     1|  2000|
|  107| bbb| 45|     2|  4500|
|  108| aaa| 50|     1|  2000|
|  109| bbb| 35|     2|  3000|
+-----+----+---+------+------+
```

```
df.printSchema()
```

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

```
df.columns
```

```
['empid', 'name', 'age', 'deptid', 'salary']
```

```
from pyspark.ml.feature import VectorAssembler
fa=VectorAssembler(inputCols=["age","deptid"],outputCol="newinputcolumn")
```

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

```
df1.show()
```

```
+-----+----+---+------+------+--------------+
|empid|name|age|deptid|salary|newinputcolumn|
+-----+----+---+------+------+--------------+
|  110| aaa| 35|     1|  2000|    [35.0,1.0]|
|  111| bbb| 45|     2|  3000|    [45.0,2.0]|
|  112| aaa| 50|     1|  2500|    [50.0,1.0]|
|  113| bbb| 35|     2|  3000|    [35.0,2.0]|
|  104| aaa| 40|     1|  2000|    [40.0,1.0]|
|  105| bbb| 40|     2|  3500|    [40.0,2.0]|
|  106| aaa| 35|     1|  2000|    [35.0,1.0]|
|  107| bbb| 45|     2|  4500|    [45.0,2.0]|
|  108| aaa| 50|     1|  2000|    [50.0,1.0]|
|  109| bbb| 35|     2|  3000|    [35.0,2.0]|
+-----+----+---+------+------+--------------+
```

```
df2=df1.select("newinputcolumn","salary")
```

```
df2.show()
```

```
+--------------+------+
|newinputcolumn|salary|
+--------------+------+
|    [35.0,1.0]|  2000|
|    [45.0,2.0]|  3000|
|    [50.0,1.0]|  2500|
|    [35.0,2.0]|  3000|
|    [40.0,1.0]|  2000|
|    [40.0,2.0]|  3500|
|    [35.0,1.0]|  2000|
|    [45.0,2.0]|  4500|
|    [50.0,1.0]|  2000|
|    [35.0,2.0]|  3000|
+--------------+------+
```

```
from pyspark.ml.regression import LinearRegression
train_data,test_data=df2.randomSplit([0.75,0.25])
applyml=LinearRegression(featuresCol='newinputcolumn', labelCol='salary')
applyml=applyml.fit(train_data)
```

```
applyml.coefficients
```

```
DenseVector([50.0, 1275.0])
```

```
applyml.intercept
```

```
-1150.0000000000182
```

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

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

```
/usr/local/lib/python3.7/dist-packages/pyspark/sql/context.py:127: FutureWarni
  FutureWarning
+--------------+------+------------------+
|newinputcolumn|salary|        prediction|
+--------------+------+------------------+
|    [50.0,1.0]|  2000|2625.0000000000064|
+--------------+------+------------------+
```

```
predict.meanAbsoluteError,predict.meanSquaredError
```

```
(625.0000000000064, 390625.000000008)
```

✓ 0s completed at 8:43 PM