

```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /opt/conda/lib/python3.7/site-packages (3.3.1)
Requirement already satisfied: py4j==0.10.9.5 in /opt/conda/lib/python3.7/site-packages (from pyspark) (0.10.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the s
```

```
from pyspark.sql import SparkSession
from pyspark.ml import Pipeline
from pyspark.sql.functions import mean,col,split, col, regexp_extract, when, lit
from pyspark.ml.feature import StringIndexer
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.feature import QuantileDiscretizer
```

```
spark = SparkSession \
    .builder \
    .appName("Spark ML for titanic data ") \
    .getOrCreate()
```

```
df_train = spark.read.csv('../input/titanic/train.csv', header = True, inferSchema=True)
df_test = spark.read.csv('../input/titanic/train.csv', header = True, inferSchema=True)
```

```
titanic_train = df_train.alias("titanic_train")
```

```
df_train.show()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	null	
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	null	
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	null	
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	null	

7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	null
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	null
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	null
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103
13	0	3	Saundercock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	null
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	null
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	null
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	null
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	null
18	1	2	Williams, Mr. Cha...	male	null	0	0	244373	13.0	null
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	null
20	1	3	Masselmani, Mrs. ...	female	null	0	0	2649	7.225	null

only showing top 20 rows

```
df_train.select("Survived","Pclass","Embarked").show()
```

Survived	Pclass	Embarked
0	3	S
1	1	C
1	3	S
1	1	S
0	3	S
0	3	Q
0	1	S
0	3	S
1	3	S
1	2	C
1	3	S
1	1	S
0	3	S
0	3	S
0	3	S
1	2	S
0	3	Q
1	2	S

```

|      0 |      3 |      S |
|      1 |      3 |      C |
+-----+-----+-----+
only showing top 20 rows

```

```
df_train.groupby("Survived").count().show()
```

```

+-----+-----+
|Survived|count|
+-----+-----+
|      1 |   342|
|      0 |   549|
+-----+-----+

```

```
grpby_output = df_train.groupby("Survived").count()
```

```
grpby_output.show()
```

```

+-----+-----+
|Survived|count|
+-----+-----+
|      1 |   342|
|      0 |   549|
+-----+-----+

```

```
df_train.groupby("Sex","Survived").count().show()
```

```

+-----+-----+-----+
|  Sex |Survived|count|
+-----+-----+-----+
| male |      0 |   468|
| female|     1 |   233|
| female|     0 |    81|
| male |     1 |   109|

```

```

+-----+-----+-----+
# combined_dt = df_train.join(df_test,['PassengerId'],how='inner')
combined_dt = df_train.union(df_test)
combined_dt.show()
#df_test.printSchema()
#df_test.collect()

```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Emba |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 0 | 3 | Braund, Mr. Owen ... | male | 22.0 | 1 | 0 | A/5 21171 | 7.25 | null | |
| 2 | 1 | 1 | Cumings, Mrs. Joh... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | |
| 3 | 1 | 3 | Heikkinen, Miss. ... | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.925 | null | |
| 4 | 1 | 1 | Futrelle, Mrs. Ja... | female | 35.0 | 1 | 0 | 113803 | 53.1 | C123 | |
| 5 | 0 | 3 | Allen, Mr. Willia... | male | 35.0 | 0 | 0 | 373450 | 8.05 | null | |
| 6 | 0 | 3 | Moran, Mr. James | male | null | 0 | 0 | 330877 | 8.4583 | null | |
| 7 | 0 | 1 | McCarthy, Mr. Tim... | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E46 | |
| 8 | 0 | 3 | Palsson, Master. ... | male | 2.0 | 3 | 1 | 349909 | 21.075 | null | |
| 9 | 1 | 3 | Johnson, Mrs. Osc... | female | 27.0 | 0 | 2 | 347742 | 11.1333 | null | |
| 10 | 1 | 2 | Nasser, Mrs. Nich... | female | 14.0 | 1 | 0 | 237736 | 30.0708 | null | |
| 11 | 1 | 3 | Sandstrom, Miss. ... | female | 4.0 | 1 | 1 | PP 9549 | 16.7 | G6 | |
| 12 | 1 | 1 | Bonnell, Miss. El... | female | 58.0 | 0 | 0 | 113783 | 26.55 | C103 | |
| 13 | 0 | 3 | Saundercock, Mr. ... | male | 20.0 | 0 | 0 | A/5. 2151 | 8.05 | null | |
| 14 | 0 | 3 | Andersson, Mr. An... | male | 39.0 | 1 | 5 | 347082 | 31.275 | null | |
| 15 | 0 | 3 | Vestrom, Miss. Hu... | female | 14.0 | 0 | 0 | 350406 | 7.8542 | null | |
| 16 | 1 | 2 | Hewlett, Mrs. (Ma... | female | 55.0 | 0 | 0 | 248706 | 16.0 | null | |
| 17 | 0 | 3 | Rice, Master. Eugene | male | 2.0 | 4 | 1 | 382652 | 29.125 | null | |
| 18 | 1 | 2 | Williams, Mr. Cha... | male | null | 0 | 0 | 244373 | 13.0 | null | |
| 19 | 0 | 3 | Vander Planke, Mr... | female | 31.0 | 1 | 0 | 345763 | 18.0 | null | |
| 20 | 1 | 3 | Masselmani, Mrs. ... | female | null | 0 | 0 | 2649 | 7.225 | null | |

```

only showing top 20 rows

```
#combined_dt.write.csv('../titanic/output.csv')
```

```

# Checking null values and null count
def null_value_count(df):
    null_columns_counts = []

```

```

numRows = df.count()
for k in df.columns:
    nullRows = df.where(col(k).isNull()).count()
    if(nullRows > 0):
        temp = k,nullRows
        null_columns_counts.append(temp)
return(null_columns_counts)

```

```

null_cols_counts = null_value_count(combined_dt)

```

```

combined_dt.where(combined_dt['Age'].isNull()).show()

```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	null	
18	1	2	Williams, Mr. Cha...	male	null	0	0	244373	13.0	null	
20	1	3	Masselmani, Mrs. ...	female	null	0	0	2649	7.225	null	
27	0	3	Emir, Mr. Farred ...	male	null	0	0	2631	7.225	null	
29	1	3	"O'Dwyer, Miss. E...	female	null	0	0	330959	7.8792	null	
30	0	3	Todoroff, Mr. Lalio	male	null	0	0	349216	7.8958	null	
32	1	1	Spencer, Mrs. Wil...	female	null	1	0	PC 17569	146.5208	B78	
33	1	3	Glynn, Miss. Mary...	female	null	0	0	335677	7.75	null	
37	1	3	Mamee, Mr. Hanna	male	null	0	0	2677	7.2292	null	
43	0	3	Kraeff, Mr. Theodor	male	null	0	0	349253	7.8958	null	
46	0	3	Rogers, Mr. Willi...	male	null	0	0	S.C./A.4. 23567	8.05	null	
47	0	3	Lennon, Mr. Denis	male	null	1	0	370371	15.5	null	
48	1	3	O'Driscoll, Miss....	female	null	0	0	14311	7.75	null	
49	0	3	Samaan, Mr. Youssef	male	null	2	0	2662	21.6792	null	
56	1	1	Woolner, Mr. Hugh	male	null	0	0	19947	35.5	C52	
65	0	1	Stewart, Mr. Albe...	male	null	0	0	PC 17605	27.7208	null	
66	1	3	Moubarek, Master....	male	null	1	1	2661	15.2458	null	
77	0	3	Staneff, Mr. Ivan	male	null	0	0	349208	7.8958	null	
78	0	3	Moutal, Mr. Raham...	male	null	0	0	374746	8.05	null	
83	1	3	McDermott, Miss. ...	female	null	0	0	330932	7.7875	null	

only showing top 20 rows

```
combined_dt.where(combined_dt['Fare'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
combined_dt.where(combined_dt['Age'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 6 | 0 | 3 | Moran, Mr. James | male | null | 0 | 0 | 330877 | 8.4583 | null | null |
| 18 | 1 | 2 | Williams, Mr. Cha... | male | null | 0 | 0 | 244373 | 13.0 | null | null |
| 20 | 1 | 3 | Masselmani, Mrs. ... | female | null | 0 | 0 | 2649 | 7.225 | null | null |
| 27 | 0 | 3 | Emir, Mr. Farred ... | male | null | 0 | 0 | 2631 | 7.225 | null | null |
| 29 | 1 | 3 | O'Dwyer, Miss. E... | female | null | 0 | 0 | 330959 | 7.8792 | null | null |
| 30 | 0 | 3 | Todoroff, Mr. Lalio | male | null | 0 | 0 | 349216 | 7.8958 | null | null |
| 32 | 1 | 1 | Spencer, Mrs. Wil... | female | null | 1 | 0 | PC 17569 | 146.5208 | B78 | null |
| 33 | 1 | 3 | Glynn, Miss. Mary... | female | null | 0 | 0 | 335677 | 7.75 | null | null |
| 37 | 1 | 3 | Mamee, Mr. Hanna | male | null | 0 | 0 | 2677 | 7.2292 | null | null |
| 43 | 0 | 3 | Kraeff, Mr. Theodor | male | null | 0 | 0 | 349253 | 7.8958 | null | null |
| 46 | 0 | 3 | Rogers, Mr. Willi... | male | null | 0 | 0 | S.C./A.4. 23567 | 8.05 | null | null |
| 47 | 0 | 3 | Lennon, Mr. Denis | male | null | 1 | 0 | 370371 | 15.5 | null | null |
| 48 | 1 | 3 | O'Driscoll, Miss.... | female | null | 0 | 0 | 14311 | 7.75 | null | null |
| 49 | 0 | 3 | Samaan, Mr. Youssef | male | null | 2 | 0 | 2662 | 21.6792 | null | null |
| 56 | 1 | 1 | Woolner, Mr. Hugh | male | null | 0 | 0 | 19947 | 35.5 | C52 | null |
| 65 | 0 | 1 | Stewart, Mr. Albe... | male | null | 0 | 0 | PC 17605 | 27.7208 | null | null |
| 66 | 1 | 3 | Moubarek, Master.... | male | null | 1 | 1 | 2661 | 15.2458 | null | null |
| 77 | 0 | 3 | Staneff, Mr. Ivan | male | null | 0 | 0 | 349208 | 7.8958 | null | null |
| 78 | 0 | 3 | Moutal, Mr. Raham... | male | null | 0 | 0 | 374746 | 8.05 | null | null |
| 83 | 1 | 3 | McDermott, Miss. ... | female | null | 0 | 0 | 330932 | 7.7875 | null | null |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

only showing top 20 rows

```
combined_dt.where(combined_dt['Fare'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```

|PassengerId|Survived|Pclass|Name|Sex|Age|SibSp|Parch|Ticket|Fare|Cabin|Embarked|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

```
combined_dt.where(combined_dt['Cabin'].isNull()).show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|PassengerId|Survived|Pclass|Name|Sex|Age|SibSp|Parch|Ticket|Fare|Cabin|Emba
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|1|0|3|Braund, Mr. Owen ...|male|22.0|1|0|A/5 21171|7.25|null|
|3|1|3|Heikkinen, Miss. ...|female|26.0|0|0|STON/O2. 3101282|7.925|null|
|5|0|3|Allen, Mr. Willia...|male|35.0|0|0|373450|8.05|null|
|6|0|3|Moran, Mr. James|male|null|0|0|330877|8.4583|null|
|8|0|3|Palsson, Master. ...|male|2.0|3|1|349909|21.075|null|
|9|1|3|Johnson, Mrs. Osc...|female|27.0|0|2|347742|11.1333|null|
|10|1|2|Nasser, Mrs. Nich...|female|14.0|1|0|237736|30.0708|null|
|13|0|3|Saunderscock, Mr. ...|male|20.0|0|0|A/5. 2151|8.05|null|
|14|0|3|Andersson, Mr. An...|male|39.0|1|5|347082|31.275|null|
|15|0|3|Vestrom, Miss. Hu...|female|14.0|0|0|350406|7.8542|null|
|16|1|2|Hewlett, Mrs. (Ma...|female|55.0|0|0|248706|16.0|null|
|17|0|3|Rice, Master. Eugene|male|2.0|4|1|382652|29.125|null|
|18|1|2|Williams, Mr. Cha...|male|null|0|0|244373|13.0|null|
|19|0|3|Vander Planke, Mr...|female|31.0|1|0|345763|18.0|null|
|20|1|3|Masselmani, Mrs. ...|female|null|0|0|2649|7.225|null|
|21|0|2|Fynney, Mr. Joseph J|male|35.0|0|0|239865|26.0|null|
|23|1|3|"McGowan, Miss. A...|female|15.0|0|0|330923|8.0292|null|
|25|0|3|Palsson, Miss. To...|female|8.0|3|1|349909|21.075|null|
|26|1|3|Asplund, Mrs. Car...|female|38.0|1|5|347077|31.3875|null|
|27|0|3|Emir, Mr. Farred ...|male|null|0|0|2631|7.225|null|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

only showing top 20 rows

```
combined_dt = combined_dt.na.fill('N', subset=['Cabin'])
#df_test = df_test.na.fill('N', subset=['Cabin'])
```

```
missing_value = combined_dt.filter(
    (combined_dt['Pclass'] == 3) &
```

```
(combined_dt.Embarked == 'S') &
(combined_dt.Sex == "male")
)
## filling in the null value in the fare column using Fare mean.
combined_dt = combined_dt.na.fill(
    missing_value.select(mean('Fare')).collect()[0][0],
    subset=['Fare']
)
```

```
combined_dt.where(combined_dt['Fare'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
combined_dt = combined_dt.na.fill('C', subset=['Embarked'])
```

```
combined_dt.where(combined_dt['Embarked'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
#df_test.where(df_test.Embarked.isNull()).show()
```

```
combined_dt.where(combined_dt['Cabin'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```



```
combined_dt.where(combined_dt['Cabin'].isNull()).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
combined_dt.select("Name").count()
```

```
1782
```

```
combined_dt.select("Name").show()
```

```
+-----+
| Name |
+-----+
| Braund, Mr. Owen ... |
| Cumings, Mrs. Joh... |
| Heikkinen, Miss. ... |
| Futrelle, Mrs. Ja... |
| Allen, Mr. Willia... |
| Moran, Mr. James |
| McCarthy, Mr. Tim... |
| Palsson, Master. ... |
| Johnson, Mrs. Osc... |
| Nasser, Mrs. Nich... |
| Sandstrom, Miss. ... |
| Bonnell, Miss. El... |
| Saundercock, Mr. ... |
| Andersson, Mr. An... |
| Vestrom, Miss. Hu... |
| Hewlett, Mrs. (Ma... |
| Rice, Master. Eugene |
| Williams, Mr. Cha... |
| Vander Planke, Mr... |
| Masselmani, Mrs. ... |
+-----+
```

```
only showing top 20 rows
```

```
combined_dt.filter(combined_dt.Age==46).select("Name").show()
```

```
+-----+
|           Name |
+-----+
| Chaffee, Mr. Herb... |
| McKane, Mr. Peter... |
| Guggenheim, Mr. B... |
| Chaffee, Mr. Herb... |
| McKane, Mr. Peter... |
| Guggenheim, Mr. B... |
+-----+
```

```
combined_dt.select("Age").show()
```

```
+-----+
| Age |
+-----+
| 22.0 |
| 38.0 |
| 26.0 |
| 35.0 |
| 35.0 |
| null |
| 54.0 |
| 2.0 |
| 27.0 |
| 14.0 |
| 4.0 |
| 58.0 |
| 20.0 |
| 39.0 |
| 14.0 |
| 55.0 |
| 2.0 |
| null |
| 31.0 |
| null |
+-----+
```

only showing top 20 rows

```
#Adding family size
combined_dt = combined_dt.withColumn("Family_Size",col('SibSp')+col('Parch'))
```

```
combined_dt.groupBy("Family_Size").count().show()
```

```
+-----+-----+
|Family_Size|count|
+-----+-----+
|          1|   322|
|          6|    24|
|          3|    58|
|          5|    44|
|          4|    30|
|          7|    12|
|         10|    14|
|          2|   204|
|          0|  1074|
+-----+-----+
```

```
#Adding alone
combined_dt = combined_dt.withColumn('Alone',lit(0))
```

```
combined_dt = combined_dt.withColumn("Alone",when(combined_dt["Family_Size"] == 0, 1).otherwise(combined_dt["Alone"]
```

```
combined_dt.columns
```

```
['PassengerId',
 'Survived',
 'Pclass',
 'Name',
 'Sex',
 'Age',
```

```
'SibSp',
'Parch',
'Ticket',
'Fare',
'Cabin',
'Embarked',
'Family_Size',
'Alone']
```

```
combined_dt.show()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	N	
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	N	
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	N	
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	N	
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	N	
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	N	
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	N	
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	N	
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	N	
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	N	
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	N	
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	N	
18	1	2	Williams, Mr. Cha...	male	null	0	0	244373	13.0	N	
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	N	
20	1	3	Masselmani, Mrs. ...	female	null	0	0	2649	7.225	N	

only showing top 20 rows

```
combined_dt.where(combined_dt['Age'].isnull()).show()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	N	
18	1	2	Williams, Mr. Cha...	male	null	0	0	244373	13.0	N	
20	1	3	Masselmani, Mrs. ...	female	null	0	0	2649	7.225	N	
27	0	3	Emir, Mr. Farred ...	male	null	0	0	2631	7.225	N	
29	1	3	"O'Dwyer, Miss. E...	female	null	0	0	330959	7.8792	N	
30	0	3	Todoroff, Mr. Lallo	male	null	0	0	349216	7.8958	N	
32	1	1	Spencer, Mrs. Wil...	female	null	1	0	PC 17569	146.5208	B78	
33	1	3	Glynn, Miss. Mary...	female	null	0	0	335677	7.75	N	
37	1	3	Mamee, Mr. Hanna	male	null	0	0	2677	7.2292	N	
43	0	3	Kraeff, Mr. Theodor	male	null	0	0	349253	7.8958	N	
46	0	3	Rogers, Mr. Willi...	male	null	0	0	S.C./A.4. 23567	8.05	N	
47	0	3	Lennon, Mr. Denis	male	null	1	0	370371	15.5	N	
48	1	3	O'Driscoll, Miss....	female	null	0	0	14311	7.75	N	
49	0	3	Samaan, Mr. Youssef	male	null	2	0	2662	21.6792	N	
56	1	1	Woolner, Mr. Hugh	male	null	0	0	19947	35.5	C52	
65	0	1	Stewart, Mr. Albe...	male	null	0	0	PC 17605	27.7208	N	
66	1	3	Moubarek, Master....	male	null	1	1	2661	15.2458	N	
77	0	3	Staneff, Mr. Ivan	male	null	0	0	349208	7.8958	N	
78	0	3	Moutal, Mr. Raham...	male	null	0	0	374746	8.05	N	
83	1	3	McDermott, Miss. ...	female	null	0	0	330932	7.7875	N	

only showing top 20 rows

```
combined_dt.where(combined_dt['Cabin'].isNull()).show()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Family_Size	Alone
-------------	----------	--------	------	-----	-----	-------	-------	--------	------	-------	----------	-------------	-------

```
combined_dt = combined_dt.withColumn("Initial", regexp_extract(col("Name"), "([A-Za-z]+)\.", 1))
```

```
combined_dt.show()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	N	
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	N	
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	N	
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	N	
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	N	
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	N	
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	N	
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	N	
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	N	
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	N	
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	N	
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	N	
18	1	2	Williams, Mr. Cha...	male	null	0	0	244373	13.0	N	
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	N	
20	1	3	Masselmani, Mrs. ...	female	null	0	0	2649	7.225	N	

only showing top 20 rows

```
combined_dt.select("Initial").distinct().show()
```

```
+-----+
| Initial|
+-----+
| Don    |
| Miss   |
| Countess|
| Col    |
| Rev    |
| Lady   |
| Master |
| Mme    |
```

```

|      Capt |
|      Mr  |
|      Dr  |
|      Mrs |
|      Sir |
|Jonkheer |
|      Mlle|
|      Major|
|      Ms  |
+-----+

```

```
combined_dt = combined_dt.replace(['Mlle','Mme', 'Ms', 'Dr','Major','Lady','Countess','Jonkheer','Col','Rev','Capt',
                                   ['Miss','Miss','Miss','Mr','Mr', 'Mrs', 'Mrs', 'Other', 'Other','Other','Mr','Mr','Mr']])
```

```
combined_dt.select("Initial").distinct().show()
```

```

+-----+
|Initial|
+-----+
|  Miss |
|  Other|
| Master|
|    Mr |
|    Mrs|
+-----+

```

```
combined_dt.groupby('Initial').avg('Age').collect()
```

```

[Row(Initial='Miss', avg(Age)=21.86),
 Row(Initial='Other', avg(Age)=45.888888888888886),
 Row(Initial='Master', avg(Age)=4.574166666666667),
 Row(Initial='Mr', avg(Age)=32.73960880195599),
 Row(Initial='Mrs', avg(Age)=35.981818181818184)]

```

```
combined_dt = combined_dt.withColumn("Age",when((combined_dt["Initial"] == "Miss") & (combined_dt["Age"].isNull()))
combined_dt = combined_dt.withColumn("Age",when((combined_dt["Initial"] == "Other") & (combined_dt["Age"].isNull()))
```

```
combined_dt = combined_dt.withColumn("Age",when((combined_dt["Initial"] == "Master") & (combined_dt["Age"].isNull()),
combined_dt = combined_dt.withColumn("Age",when((combined_dt["Initial"] == "Mr") & (combined_dt["Age"].isNull()),
combined_dt = combined_dt.withColumn("Age",when((combined_dt["Initial"] == "Mrs") & (combined_dt["Age"].isNull()),
```

```
combined_dt.select("Age").show()
```

```
+----+
| Age|
+----+
|22.0|
|38.0|
|26.0|
|35.0|
|35.0|
|33.0|
|54.0|
| 2.0|
|27.0|
|14.0|
| 4.0|
|58.0|
|20.0|
|39.0|
|14.0|
|55.0|
| 2.0|
|33.0|
|31.0|
|36.0|
+----+
```

only showing top 20 rows

```
combined_dt.printSchema()
```

```
root
|-- PassengerId: integer (nullable = true)
|-- Survived: integer (nullable = true)
|-- Pclass: integer (nullable = true)
|-- Name: string (nullable = true)
```



```

|-- Sex: string (nullable = true)
|-- Age: double (nullable = true)
|-- SibSp: integer (nullable = true)
|-- Parch: integer (nullable = true)
|-- Ticket: string (nullable = true)
|-- Fare: double (nullable = false)
|-- Cabin: string (nullable = false)
|-- Embarked: string (nullable = false)
|-- Family_Size: integer (nullable = true)
|-- Alone: integer (nullable = false)
|-- Initial: string (nullable = true)

```

```
combined_dt.where(combined_dt['Embarked'].isNull()).show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Family_Size | Alone | Initial |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

```

indexers = [StringIndexer(inputCol=column, outputCol=column+"_index").fit(combined_dt) for column in ["Sex", "Embar"]
pipeline = Pipeline(stages=indexers)
combined_dt = pipeline.fit(combined_dt).transform(combined_dt)

```

```
combined_dt.show()
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Family_Size | Alone | Initial |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 0 | 3 | Braund, Mr. Owen ... | male | 22.0 | 1 | 0 | A/5 21171 | 7.25 | N |  |  |  |
| 2 | 1 | 1 | Cumings, Mrs. Joh... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 |  |  |  |
| 3 | 1 | 3 | Heikkinen, Miss. ... | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.925 | N |  |  |  |
| 4 | 1 | 1 | Futrelle, Mrs. Ja... | female | 35.0 | 1 | 0 | 113803 | 53.1 | C123 |  |  |  |
| 5 | 0 | 3 | Allen, Mr. Willia... | male | 35.0 | 0 | 0 | 373450 | 8.05 | N |  |  |  |
| 6 | 0 | 3 | Moran, Mr. James | male | 33.0 | 0 | 0 | 330877 | 8.4583 | N |  |  |  |
| 7 | 0 | 1 | McCarthy, Mr. Tim... | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E46 |  |  |  |
| 8 | 0 | 3 | Palsson, Master. ... | male | 2.0 | 3 | 1 | 349909 | 21.075 | N |  |  |  |
| 9 | 1 | 3 | Johnson, Mrs. Osc... | female | 27.0 | 0 | 2 | 347742 | 11.1333 | N |  |  |  |

```

10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	N
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103
13	0	3	Saundercock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	N
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	N
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	N
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	N
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	N
18	1	2	Williams, Mr. Cha...	male	33.0	0	0	244373	13.0	N
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	N
20	1	3	Masselmani, Mrs. ...	female	36.0	0	0	2649	7.225	N

only showing top 20 rows

```
combined_dt.printSchema()
```

```
root
|-- PassengerId: integer (nullable = true)
|-- Survived: integer (nullable = true)
|-- Pclass: integer (nullable = true)
|-- Name: string (nullable = true)
|-- Sex: string (nullable = true)
|-- Age: double (nullable = true)
|-- SibSp: integer (nullable = true)
|-- Parch: integer (nullable = true)
|-- Ticket: string (nullable = true)
|-- Fare: double (nullable = false)
|-- Cabin: string (nullable = false)
|-- Embarked: string (nullable = false)
|-- Family_Size: integer (nullable = true)
|-- Alone: integer (nullable = false)
|-- Initial: string (nullable = true)
|-- Sex_index: double (nullable = false)
|-- Embarked_index: double (nullable = false)
|-- Initial_index: double (nullable = false)
```

```
combined_dt = combined_dt.drop("PassengerId", "Name", "Ticket", "Cabin", "Embarked", "Sex", "Initial")
```

```
combined_pd = combined_dt.toPandas()
```

```
combined_pd
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Family_Size	Alone	Sex_index	Embarked_index	Initial_index
0	0	3	22.0	1	0	7.2500	1	0	0.0	0.0	0.0
1	1	1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0
2	1	3	26.0	0	0	7.9250	0	1	1.0	0.0	1.0
3	1	1	35.0	1	0	53.1000	1	0	1.0	0.0	2.0
4	0	3	35.0	0	0	8.0500	0	1	0.0	0.0	0.0
...
1777	0	2	27.0	0	0	13.0000	0	1	0.0	0.0	4.0
1778	1	1	19.0	0	0	30.0000	0	1	1.0	0.0	1.0
1779	0	3	22.0	1	2	23.4500	3	0	1.0	0.0	1.0
1780	1	1	26.0	0	0	30.0000	0	1	0.0	1.0	0.0
1781	0	3	32.0	0	0	7.7500	0	1	0.0	2.0	0.0

1782 rows × 11 columns

```
feature = VectorAssembler(inputCols=combined_dt.columns[1:],outputCol="features")
feature_vector= feature.transform(combined_dt)
```

```
feature_vector.show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Survived|Pclass| Age|SibSp|Parch|   Fare|Family_Size|Alone|Sex_index|Embarked_index|Initial_index|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|         0|        3|22.0|    1|    0|   7.25|          1|    0|         0.0|          0.0|          0.0|(10,[0,1,2
```

1	1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0	[1.0, 38.0,
1	3	26.0	0	0	7.925	0	1	1.0	0.0	1.0	[3.0, 26.0,
1	1	35.0	1	0	53.1	1	0	1.0	0.0	2.0	[1.0, 35.0,
0	3	35.0	0	0	8.05	0	1	0.0	0.0	0.0	(10, [0, 1, 4
0	3	33.0	0	0	8.4583	0	1	0.0	2.0	0.0	(10, [0, 1, 4
0	1	54.0	0	0	51.8625	0	1	0.0	0.0	0.0	(10, [0, 1, 4
0	3	2.0	3	1	21.075	4	0	0.0	0.0	3.0	[3.0, 2.0, 3
1	3	27.0	0	2	11.1333	2	0	1.0	0.0	2.0	[3.0, 27.0,
1	2	14.0	1	0	30.0708	1	0	1.0	1.0	2.0	[2.0, 14.0,
1	3	4.0	1	1	16.7	2	0	1.0	0.0	1.0	[3.0, 4.0, 1
1	1	58.0	0	0	26.55	0	1	1.0	0.0	1.0	[1.0, 58.0,
0	3	20.0	0	0	8.05	0	1	0.0	0.0	0.0	(10, [0, 1, 4
0	3	39.0	1	5	31.275	6	0	0.0	0.0	0.0	[3.0, 39.0,
0	3	14.0	0	0	7.8542	0	1	1.0	0.0	1.0	[3.0, 14.0,
1	2	55.0	0	0	16.0	0	1	1.0	0.0	2.0	[2.0, 55.0,
0	3	2.0	4	1	29.125	5	0	0.0	2.0	3.0	[3.0, 2.0, 4
1	2	33.0	0	0	13.0	0	1	0.0	0.0	0.0	(10, [0, 1, 4
0	3	31.0	1	0	18.0	1	0	1.0	0.0	2.0	[3.0, 31.0,
1	3	36.0	0	0	7.225	0	1	1.0	1.0	2.0	[3.0, 36.0,

only showing top 20 rows

```
train_pd = combined_pd[:df_train.count()]
test_pd = combined_pd[df_train.count():]
```

```
train_pd
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Family_Size	Alone	Sex_index	Embarked_index	Initial_index
0	0	3	22.0	1	0	7.2500	1	0	0.0	0.0	0.0
1	1	1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0
2	1	3	26.0	0	0	7.9250	0	1	1.0	0.0	1.0
3	1	1	35.0	1	0	53.1000	1	0	1.0	0.0	2.0
4	0	3	35.0	0	0	8.0500	0	1	0.0	0.0	0.0

```
df_train = spark.createDataFrame(train_pd)
df_test = spark.createDataFrame(test_pd)
df_test = df_test.drop('Survived')
#df_test.show()
```

```
220      1      1 26.0      0      0 30.0000      0      1      0.0      1.0      0.0
```

```
assembler = VectorAssembler(inputCols=df_train.columns[1:],outputCol="features")
train_assembler_vector = assembler.transform(df_train)
train_assembler_vector.show()
```

Survived	Pclass	Age	SibSp	Parch	Fare	Family_Size	Alone	Sex_index	Embarked_index	Initial_index
0	3	22.0	1	0	7.25	1	0	0.0	0.0	0.0
1	1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0
1	3	26.0	0	0	7.925	0	1	1.0	0.0	1.0
1	1	35.0	1	0	53.1	1	0	1.0	0.0	2.0
0	3	35.0	0	0	8.05	0	1	0.0	0.0	0.0
0	3	33.0	0	0	8.4583	0	1	0.0	2.0	0.0
0	1	54.0	0	0	51.8625	0	1	0.0	0.0	0.0
0	3	2.0	3	1	21.075	4	0	0.0	0.0	3.0
1	3	27.0	0	2	11.1333	2	0	1.0	0.0	2.0
1	2	14.0	1	0	30.0708	1	0	1.0	1.0	2.0
1	3	4.0	1	1	16.7	2	0	1.0	0.0	1.0
1	1	58.0	0	0	26.55	0	1	1.0	0.0	1.0
0	3	20.0	0	0	8.05	0	1	0.0	0.0	0.0
0	3	39.0	1	5	31.275	6	0	0.0	0.0	0.0
0	3	14.0	0	0	7.8542	0	1	1.0	0.0	1.0

1	2	55.0	0	0	16.0	0	1	1.0	0.0	2.0	[2.0,55.0,
0	3	2.0	4	1	29.125	5	0	0.0	2.0	3.0	[3.0,2.0,4
1	2	33.0	0	0	13.0	0	1	0.0	0.0	0.0	(10,[0,1,4
0	3	31.0	1	0	18.0	1	0	1.0	0.0	2.0	[3.0,31.0,
1	3	36.0	0	0	7.225	0	1	1.0	1.0	2.0	[3.0,36.0,

only showing top 20 rows

```
test_assembler = VectorAssembler(inputCols=df_test.columns,outputCol="features")
test_assembler_vector = test_assembler.transform(df_test)
test_assembler_vector.show()
```

Pclass	Age	SibSp	Parch	Fare	Family_Size	Alone	Sex_index	Embarked_index	Initial_index	feature
3	22.0	1	0	7.25	1	0	0.0	0.0	0.0	(10,[0,1,2,4,5],[..
1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0	[1.0,38.0,1.0,0.0..
3	26.0	0	0	7.925	0	1	1.0	0.0	1.0	[3.0,26.0,0.0,0.0..
1	35.0	1	0	53.1	1	0	1.0	0.0	2.0	[1.0,35.0,1.0,0.0..
3	35.0	0	0	8.05	0	1	0.0	0.0	0.0	(10,[0,1,4,6],[3...
3	33.0	0	0	8.4583	0	1	0.0	2.0	0.0	(10,[0,1,4,6,8],[..
1	54.0	0	0	51.8625	0	1	0.0	0.0	0.0	(10,[0,1,4,6],[1...
3	2.0	3	1	21.075	4	0	0.0	0.0	3.0	[3.0,2.0,3.0,1.0,..
3	27.0	0	2	11.1333	2	0	1.0	0.0	2.0	[3.0,27.0,0.0,2.0..
2	14.0	1	0	30.0708	1	0	1.0	1.0	2.0	[2.0,14.0,1.0,0.0..
3	4.0	1	1	16.7	2	0	1.0	0.0	1.0	[3.0,4.0,1.0,1.0,..
1	58.0	0	0	26.55	0	1	1.0	0.0	1.0	[1.0,58.0,0.0,0.0..
3	20.0	0	0	8.05	0	1	0.0	0.0	0.0	(10,[0,1,4,6],[3...
3	39.0	1	5	31.275	6	0	0.0	0.0	0.0	[3.0,39.0,1.0,5.0..
3	14.0	0	0	7.8542	0	1	1.0	0.0	1.0	[3.0,14.0,0.0,0.0..
2	55.0	0	0	16.0	0	1	1.0	0.0	2.0	[2.0,55.0,0.0,0.0..
3	2.0	4	1	29.125	5	0	0.0	2.0	3.0	[3.0,2.0,4.0,1.0,..
2	33.0	0	0	13.0	0	1	0.0	0.0	0.0	(10,[0,1,4,6],[2...
3	31.0	1	0	18.0	1	0	1.0	0.0	2.0	[3.0,31.0,1.0,0.0..
3	36.0	0	0	7.225	0	1	1.0	1.0	2.0	[3.0,36.0,0.0,0.0..

only showing top 20 rows

```
(trainData, testData) = train_assembler_vector.randomSplit([0.8, 0.2], seed = 11)
```

```
from pyspark.ml.classification import LinearSVC
svm = LinearSVC(labelCol="Survived", featuresCol="features")
svm_model = svm.fit(trainData)
svm_prediction = svm_model.transform(testData)
svm_prediction.select("prediction", "Survived", "features").show()
```

```
+-----+-----+-----+
|prediction|Survived|          features|
+-----+-----+-----+
|         0.0|         0|[1.0,28.0,1.0,0.0...|
|         0.0|         0|(10,[0,1,4,6],[1....|
|         0.0|         0|(10,[0,1,2,4,5],[...|
|         0.0|         0|(10,[0,1,2,4,5],[...|
|         0.0|         0|(10,[0,1,4,6],[1....|
|         0.0|         0|[1.0,51.0,0.0,1.0...|
|         0.0|         0|(10,[0,1,4,6],[1....|
|         0.0|         0|[1.0,65.0,0.0,1.0...|
|         0.0|         0|(10,[0,1,2,4,5],[...|
|         1.0|         0|[2.0,24.0,0.0,0.0...|
|         1.0|         0|[2.0,27.0,1.0,0.0...|
|         0.0|         0|(10,[0,1,4,6],[2....|
|         0.0|         0|(10,[0,1,4,6,8],[...|
|         0.0|         0|[3.0,1.0,4.0,1.0,...|
|         0.0|         0|[3.0,2.0,4.0,2.0,...|
|         1.0|         0|[3.0,17.0,0.0,0.0...|
|         0.0|         0|(10,[0,1,4,6,8],[...|
|         0.0|         0|(10,[0,1,4,6],[3....|
|         0.0|         0|(10,[0,1,4,6],[3....|
|         0.0|         0|(10,[0,1,4,6],[3....|
```

only showing top 20 rows

```
svm_prediction.filter(df_train.Survived==1).show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Survived|Pclass| Age|SibSp|Parch|      Fare|Family_Size|Alone|Sex_index|Embarked_index|Initial_index|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|         1|         1|28.0|         0|         0|      35.5|         0|         1|         0.0|         0.0|         0.0|(10,[0,1,
```

1	1	36.0	0	1	55.0	1	0	1.0	0.0	2.0	[1.0,36.0
1	1	36.0	1	0	146.5208	1	0	1.0	1.0	2.0	[1.0,36.0
1	1	40.0	0	0	31.0	0	1	0.0	1.0	0.0	(10,[0,1,
1	1	49.0	1	0	76.7292	1	0	1.0	1.0	2.0	[1.0,49.0
1	2	3.0	1	2	41.5792	3	0	1.0	1.0	1.0	[2.0,3.0,
1	2	21.0	0	0	10.5	0	1	1.0	0.0	1.0	[2.0,21.0
1	2	29.0	0	0	10.5	0	1	1.0	0.0	2.0	[2.0,29.0
1	2	29.0	1	0	26.0	1	0	1.0	0.0	2.0	[2.0,29.0
1	2	29.0	1	0	26.0	1	0	1.0	0.0	2.0	[2.0,29.0
1	3	16.0	0	0	8.05	0	1	0.0	0.0	0.0	(10,[0,1,
1	3	22.0	0	0	7.75	0	1	1.0	2.0	1.0	[3.0,22.0
1	3	24.0	1	0	15.85	1	0	1.0	0.0	2.0	[3.0,24.0
1	3	30.0	0	0	12.475	0	1	1.0	0.0	1.0	[3.0,30.0
1	3	36.0	0	0	7.225	0	1	1.0	1.0	2.0	[3.0,36.0
1	1	18.0	2	2	262.375	4	0	1.0	1.0	1.0	[1.0,18.0
1	1	35.0	0	0	135.6333	0	1	1.0	0.0	1.0	[1.0,35.0
1	1	36.0	0	0	135.6333	0	1	1.0	1.0	1.0	[1.0,36.0
1	1	37.0	1	1	52.5542	2	0	0.0	0.0	0.0	[1.0,37.0
1	1	41.0	0	0	134.5	0	1	1.0	1.0	1.0	[1.0,41.0

only showing top 20 rows

```
svm_prediction.filter(df_train.Survived==1).count()
```

72

```
evaluator = MulticlassClassificationEvaluator(
    labelCol="Survived", predictionCol="prediction", metricName="accuracy")
svm_accuracy = evaluator.evaluate(svm_prediction)
print("Accuracy of Support Vector Machine is = %g"% (svm_accuracy))
print("Test Error of Support Vector Machine = %g " % (1.0 - svm_accuracy))
```

Accuracy of Support Vector Machine is = 0.802198

Test Error of Support Vector Machine = 0.197802

```
submission = spark.read.csv('../input/titanic/gender_submission.csv', header = True, inferSchema=True)
submission.printSchema()
```



```

root
|-- PassengerId: integer (nullable = true)
|-- Survived: integer (nullable = true)

```

```
df_test = df_test.drop('Survived')
```

```

f_predictions = svm_model.transform(test_assembler_vector)
f_predictions = f_predictions.toPandas()

```

```
f_predictions
```

	Pclass	Age	SibSp	Parch	Fare	Family_Size	Alone	Sex_index	Embarked_index	Initial_index	features
0	3	22.0	1	0	7.2500	1	0	0.0	0.0	0.0	(3.0, 22.0, 1.0, 0.0, 7.25, 1.0, 0.0, 0.0, 0.0...
1	1	38.0	1	0	71.2833	1	0	1.0	1.0	2.0	[1.0, 38.0, 1.0, 0.0, 71.2833, 1.0, 0.0, 1.0, ...
2	3	26.0	0	0	7.9250	0	1	1.0	0.0	1.0	[3.0, 26.0, 0.0, 0.0, 7.925, 0.0, 1.0, 1.0, 0....
3	1	35.0	1	0	53.1000	1	0	1.0	0.0	2.0	[1.0, 35.0, 1.0, 0.0, 53.1, 1.0, 0.0, 1.0, ^ ^

```

submission = submission.toPandas()
submission['Survived'] = f_predictions['prediction']

```

```
submission
```

	PassengerId	Survived
0	892	0.0
1	893	1.0
2	894	1.0
3	895	1.0
4	896	0.0
...
413	1305	0.0
414	1306	0.0
415	1307	1.0
416	1308	1.0
417	1309	1.0

418 rows x 2 columns

```
submission['Survived'] = submission['Survived'].astype(int)
```

```
submission.to_csv("submission.csv",index=False)
```

[Colab paid products](#) - [Cancel contracts here](#)

