Harshil Shah - SEC01 (NUID 002780887)

Big Data System Engineering with Scala Spring 2023 Spark Assignment 2



- GitHub Repo URL - https://github.com/harshilshahneu/CSYE7200-Harshil-Shah/tree/Spring2023/assig nment-spark-2

- List of Tasks Implemented

Exploratory Data Analysis:-

- Show basic statistics for each column in train.csv
- Calculate the count of each categorical value in the Pclass column
- Calculate the count of each categorical value in the Embarked column
- Calculate Mean age of passengers by sex and class

Feature Engineering:-

 Create the columns - Title, Family Size, isAlone to get the the title, family size and whether the passenger was traveling alone

Prediction:-

Use the train.csv to train a Machine Learning model of your choice & test it on the test.csv. You are required to predict if the records in test.csv survived or not. Note(1 = Survived, 0 = Dead)

- Code & Results

1. Exploratory Data Analysis

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.{avg, col, udf, when}
import org.apache.spark.ml.feature.{OneHotEncoder, StringIndexer, VectorAssembler}
import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.{Pipeline, PipelineModel}
object Titanic extends App {
 val spark = SparkSession
   .builder()
   .appName("Titanic Dataset Analysis")
   .master("local[*]")
   .get0rCreate()
 import spark.implicits._
 val testDf = spark.read.option("header", "true").csv("assignment-spark-2/src/main/resources/test.csv")
 val trainDf = spark.read.option("header", "true").csv("assignment-spark-2/src/main/resources/train.csv")
 /** EDA */
 trainDf.show(10) // Display the first 10 rows
 trainDf.printSchema() // Print the schema of the dataset
 trainDf.describe().show() // Compute basic statistics for each column
```

```
|PassengerId|Survived|Pclass|
                                     Name| Sex| Age|SibSp|Parch|
                                                                      Ticket| Fare|Cabin|Embarked|
                      3|Braund, Mr. Owen ...| male| 22|
                                                                   A/5 21171| 7.25| null|
                      1|Cumings, Mrs. Joh...|female| 38| 1| 0|
                                                                   PC 17599|71.2833| C85|
                1| 3|Heikkinen, Miss. ...|female| 26| 0| 0|STON/02. 3101282| 7.925| null|
                                                                                             SI
                1| 1|Futrelle, Mrs. Ja...|female| 35| 1| 0| 113803| 53.1| C123|
                0| 3|Allen, Mr. Willia...| male| 35| 0| 0|
                                                                    373450| 8.05| null|
                0| 3| Moran, Mr. James| male|null| 0| 0|
                                                                    330877| 8.4583| null|
                      1|McCarthy, Mr. Tim...| male| 54|
                                                                      17463|51.8625| E46|
                                                                                             SI
                      3|Palsson, Master. ...| male| 2|
                                                                      349909| 21.075| null|
                                                                                             SI
                      3|Johnson, Mrs. Osc...|female| 27| 0| 2|
                                                                    237736|30.0708| null|
                      2|Nasser, Mrs. Nich...|female| 14| 1| 0|
only showing top 10 rows
root
|-- PassengerId: string (nullable = true)
|-- Survived: string (nullable = true)
|-- Pclass: string (nullable = true)
|-- Name: string (nullable = true)
|-- Sex: string (nullable = true)
|-- Age: string (nullable = true)
|-- SibSp: string (nullable = true)
|-- Parch: string (nullable = true)
|-- Ticket: string (nullable = true)
|-- Fare: string (nullable = true)
|-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

```
// Count of each categorical value in the Pclass column
trainDf.groupBy("Pclass").count().show()

// Count of each categorical value in the Embarked column
trainDf.groupBy("Embarked").count().show()

// Mean age of passengers by sex and class
trainDf.groupBy("Sex", "Pclass")
    .agg(avg("Age").alias("mean_age"))
    .show()
```

```
+----+
|Pclass|count|
+----+
| 3| 491|
| 1| 216|
| 2| 184|
+----+
```

```
+----+
|Embarked|count|
+----+
| Q| 77|
| null| 2|
| C| 168|
| S| 644|
+----+
```

```
+----+
| Sex|Pclass| mean_age|
+----+
|female| 3| 21.75|
| male| 2| 30.74070707070707|
|female| 1| 34.61176470588235|
| male| 1| 41.28138613861386|
| male| 3|26.507588932806325|
|female| 2|28.722972972972|
+----+
```

2. Feature Engineering

```
/** Feature Engineering */
// Extract the titles from the Name column and create a new column called Title
val titleRegex = """([\w]+)\. """.r
val getTitle = udf((name: String) => titleRegex.findFirstMatchIn(name).map(_.group(1)).getOrElse(""))
val trainDFWithTitle = trainDf.withColumn("Title", getTitle($"Name"))

// Create a new column called FamilySize by adding the SibSp and Parch columns
val trainDFWithFamilySize = trainDFWithTitle.withColumn("FamilySize", $"SibSp" + $"Parch" + 1)
val testDFWithFamilySize = testDFWithTitle.withColumn("FamilySize", $"SibSp" + $"Parch" + 1)

// Create a new column called IsAlone to indicate whether a passenger was traveling alone or with family
val trainDFWithIsAlone = trainDFWithFamilySize.withColumn("IsAlone", when($"FamilySize" === 1, 1).otherwise(0))
val testDFWithIsAlone.show()
trainDFWithIsAlone.show()
```

sengerId Sur		lass +	Name			ibSp P		Ticket					FamilySize	
1	 0	3 Braund, Mr. Owe				+- 1	0				s			 0
2	1	1 Cumings, Mrs. J	oh fe	male	38	1	0	PC 17599	71.2833	C85		Mrs	2.0	Θ
3	1	3 Heikkinen, Miss	fe	male	26	0	0 :	STON/02. 3101282	7.925	null		Miss	1.0	
41	1	1 Futrelle, Mrs.	Ja fe	male	35	1	0	113803	53.1	C123		Mrs	2.0	0
5	0	3 Allen, Mr. Will:	ia	male	35			373450	8.05	null		Mr	1.0	1
61	0	3 Moran, Mr.	James	male r	null			330877	8.4583	null		Mr	1.0	1
7	0	1 McCarthy, Mr. T	im	male	54			17463	51.8625	E46		Mr	1.0	1
8		3 Palsson, Master		male	2	3	1	349909	21.075	null		Master	5.0	
91	1	3 Johnson, Mrs. 0	sc fe	male	27		2	347742	11.1333	null		Mrs	3.0	
10	1	2 Nasser, Mrs. Ni	ch fe	male	14	1		237736	30.0708	null		Mrs	2.0	
11	1	3 Sandstrom, Miss	fe	male	4	1	1	PP 9549	16.7	G6		Miss	3.0	
12	1	1 Bonnell, Miss.	El fe	male	58			113783	26.55	C103		Miss	1.0	1
13		3 Saundercock, Mr		male	20			A/5. 2151	8.05	null		Mr	1.0	
14		3 Andersson, Mr.	An	male	39	1	5	347082	31.275	null		Mr	7.0	
15	0	3 Vestrom, Miss.	Hu fe	male	14		0	350406	7.8542	null		Miss	1.0	
16	1	2 Hewlett, Mrs. (Ma fe	male	55			248706	16	null		Mrs	1.0	
17		3 Rice, Master. E	ugene	male	2	4	1	382652	29.125	null		Master	6.0	
18	1	2 Williams, Mr. C	ha	male r	null			244373	13	null		Mr	1.0	
				malel	31 l	1	0	345763	18	null		Mrs	2.0	
19	0	3 Vander Planke, I	mr te											
20 + engerId Surv	1 + vived Pc	3 Masselmani, Mrs	fe + + Name \$	male r +- + Sex A	null +- + \ge Sib	+ Sp Par	0 + +		+ Fare Ca	+- + bin Emb	+ arked T	+ +- itle Fa	nilySize IsA	+ + Lone
20 + engerId Surv	1 + vived Pcl	3 Masselmani, Mrs	fe + + Name \$	male r +- +- Sex A	null +- + .ge Sib	+- Sp Par	0 +- +	Ticket	+ + Fare Ca	+- + bin Emb	 + arked T	itle Far	+ + nilySize IsA	+ lone +
20 + engerId Surv +	1 + vived Pc	3 Masselmani, Mrs	fe + Name \$ +	male r +- Sex A +-	null +- + age Sib 	+- Sp Par	0 + ch +	Ticket 	Fare Ca	+- + bin Emb + vll	t arked T + S	itle Far	+ + milySize IsA 	+ lone + 0
20 + engerId Surv 1 2	1 + vived Pc + 0 1	3 Masselmani, Mrs	fe + Name	male r +- Sex A + ale ale	null +- dge Sib + 22 38	+- Sp Par + 1 1	0 + ch + 0 0	Ticket +	Fare Ca + 7.25 n	+- + bin Emb + ull C85	 arked T + S C	itle Far	nilySize IsA 2.0	+ Lone + 0
20 + engerId Surv + 1 2 3	1 + vived Pc1 + 0 1 1	3 Masselmani, Mrs ass ' 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss.	fe + Name \$ + ma h fema	male r + Sex A + ale ale	null +- age Sib + 22 38 26	0Sp Par 1 1	0 + ch 0 0 0 ST	Ticket	Fare Ca + 7.25 n	+- + bin Emb + vll C85 vll		itle Far	nilySize IsA 2.0 2.0 1.0	+ lone + 0 0 1
20 + engerId Surv 	1 + vived Pc1 + 0 1 1 1	3 Masselmani, Mrs	fe	male r +- Sex A + ale ale ale	null + age Sib + 22 38 26 35	DSp Par	0 + ch 0 0 0 ST	Ticket	Fare Ca Fare Ca 7.25 n 2833 7.925 n	 bin Emb + ull C85 ull	arked T + S C S S	itle Far	2.0 2.0 2.0 2.0 2.0	+ lone + 0 0 1
20 	1 + vived Pc1 + 0 1 1	3 Masselmani, Mrs ass /- 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ja	fer	male r + Sex A + ale ale ale ale	null + age Sib + 22 38 26 35	0Sp Par 1 1	0 + ch 0 0 0 ST	Ticket	Fare Ca + 7.25 n 2833 7.925 n 53.1 C 8.05 n	 bin Emb + ull C85 ull 123		itle Far	nilySize IsA 2.0 2.0 1.0	+ lone + 0 0 1
20 	1 + vived Pc1+ 0 1 1 1 0	3 Masselmani, Mrs ass	fe + Name S+	male r + Sex A + ale ale ale ale	null +- age Sib + 22 38 26 35 35	DSp Par 1 1 0 1 0	0 + och 0 0 0 ST 0 0	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 8	Fare Ca + 7.25 n 2833 7.925 n 53.1 C 8.05 n	+- bin Emb + vll C85 vll 123 vll	arked T+- SI CI SI SI SI SI QI	itle Far	2.0 2.0 2.0 1.0 2.0 1.0	+ Lone + 0 1 0 1 1
20 	1 + vived Pc1 + 0 1 1 1 0	3 Masselmani, Mrs ass /- 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ja	fel	male r + Sex A + ale ale ale ale ale nu	null +- age Sib + 22 38 26 35 35	DSp Par 1 1 0 1	0 + ch + 0 0 0 ST 0	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 8	Fare Ca+- 7.25 n2833 7.925 n 53.1 C 8.05 n 3.4583 n	+ bin Emb + vll C85 vll 123 vll vll	sl Sl Sl Sl Sl Sl Sl	itle Far	2.0 2.0 2.0 2.0 2.0 1.0 2.0	+ lone + 0 0 1 0
20 	1 + vived Pc' 0 1 1 1 0 0	3 Masselmani, Mrs ass	fen	male r + Sex A + ale ale ale ale ale ale	null +- age Sib+ 22 38 26 35 35 51 2	1 1 0 0 0	0 + o 0 0 0 ST 0 0	A/5 21171 PC 17599 71 0N/02. 3101282 113803 373450 330877 8	Fare Ca 	+ bin Emb+ vll C85 vll 123 vll vll E46	sl Sl Sl Sl Sl Sl Sl	itle Far	2.0 2.0 2.0 1.0 2.0 1.0 1.0	+ lone + 0 0 1 0 1
20 + engerId Surv 1 2 3 4 5 6 7 8	1 +	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ja 3 Allen, Mr. Willia 3 Moran, Mr. Ja 1 McCarthy, Mr. Tin 3 Palsson, Master.	fen ma h fem a fem ames ma ma c fema	male r+ Sex A+ ale ale ale ale ale ale ale ale	null +- age Sib 22 38 26 35 35 35 35 51 21 21 27	DSp Par 1 0 1 0 0 0 0 3	0 + 0 0 0 ST 0 0 0	A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 8 17463 51	7.25 n2833 7.925 n 53.1 C 8.05 n 8.4583 n8625 11.075 n	+ bin Emb+ vll C85 vll 123 vll vll vll E46 vll		itle Far	2.0 2.0 2.0 1.0 2.0 1.0 1.0 1.0 5.0	+ lone + 0 1 0 1 1 1
20 + engerId Surv 1 2 3 4 5 6 7 8 9	1 +	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ja 3 Allen, Mr. Willia 3 Moran, Mr. Ja 1 McCarthy, Mr. Tir 3 Palsson, Master. 3 Johnson, Mrs. Osc	fer	male r+- Sex A+- ale ale ale ale ale ale ale ale	null +- age Sib 22 38 26 35 35 35 35 51 21 21 27	1 1 0 1 0 0 0 0 3	0 +	A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 8 17463 51 349909 2	7.25 n2833 7.925 n 53.1 C 8.05 n 8.4583 n8625 11.075 n	+ bin Emb+ vll C85 vll 123 vll vll vll E46 vll		itle Far	2.0 2.0 2.0 2.0 1.0 2.0 1.0 1.0 5.0 3.0	+ lone + 0 1 1 1 1 1 0
20 	1 + vived Pc'	3 Masselmani, Mrs ass 1 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ja 3 Allen, Mr. Willia 3 Moran, Mr. Ja 1 McCarthy, Mr. Tir 3 Palsson, Master. 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nicl	ferande femande femande femande mande mande mande femande fe	male r+- Sex A+- ale ale ale ale ale ale ale ale	null +- age Sib 22 38 26 35 35 51 52 27 14	DSp Par 1 1 0 1 0 0 0 3 0	0 + Cch	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 8 17463 53 349999 2 347742 11 237736 36	Fare Ca 7.25 n .2833 7.925 n 53.1 C 8.4583 n .8625 1.975 n .1333 n .0708 n			itle Far	2.0 2.0 2.0 1.0 2.0 1.0 1.0 1.0 5.0 3.0	Older Control
20 	1 + vived Pc'+ 0 1 1 0 0 0 1 1	3 Masselmani, Mrs ass ' 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Jal 3 Moran, Mr. Willia 3 Moran, Mr. Jal 1 McCarthy, Mr. Tir 3 Palsson, Master 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nicl 3 Sandstrom, Miss.	fee	male r+- Gex A+- ale ale ale ale ale ale ale ale	null +- age Sib 22 38 26 35 35 35 31 21 21 21 41	DSp Par 1 0 1 0 0 0 0 1 1 1	0 +	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 E 17463 53 349909 2 347742 13 237736 36	Fare Ca 7.25 n .2833 7.925 n 53.1 C 8.4583 n .8625 1.975 n .1333 n .0708 n			itle Far	2.0 2.0 2.0 1.0 2.0 1.0 1.0 5.0 3.0 2.0	+ lone + 0 0 1 1 1 0 0 0
20 	1	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Jol 3 Moran, Mr. Jal 1 McCarthy, Mr. Tir 3 Palsson, Master. 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nicl 3 Sandstrom, Miss.	fer	male r+ Sex A A+ sle sle sle sle sle sle sle sle	null +- age Sib 22 38 26 35 35 35 21 54 21 27 14 4 58 20	1 1 0 0 1 1 1 1 1 1	0 + o o o st o o	Ticket A/5 21171 PC 17599 71 0N/02. 3101282 113803 373450 330877 8 17463 51 349709 1 237736 36 PP 9549 113783	Fare Ca 7.25 n .2833 7.925 n 53.1 C 8.05 n 8.4583 n .8625 11.075 n .1333 n .0708 n 16.7 26.55 C 8.05 n	+ bin Emb+ vll C85 vll 123 vll vll vll vll vll vll ull vll	arked T S C S S S S Q S S S S S S S S S	itle Fai	2.0 2.0 2.0 1.0 1.0 1.0 5.0 3.0 2.0	+ Lone + 0 1 1 1 1 0 0 0
20	1	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Jol 3 Moran, Mr. Jal 1 McCarthy, Mr. Ti 3 Palsson, Master. 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nicl 3 Sandstrom, Miss. 1 Bonnell, Miss. El 3 Saundercock, Mr.	fer	male r+ sle sle sle sle sle sle sle sle	null	Sp Par 1 1 0 1 0 0 0 3 0 1 1 0	0 + 0 0 0 ST 0 0 0 1 2 0 1 0	A/5 21171 PC 17599 71 ON/02. 3101282 113803 373450 330877 8 17463 53 349709 2 347742 1 2237736 36 PP 9549 113783 A/5. 2151	Fare Ca 7.25 n .2833 7.925 n 53.1 C 8.05 n 8.4583 n .8625 11.075 n .1.333 n 0.0708 n 16.7 26.55 C 8.05 n 11.275 n	+ bin Emb+ vll C85 vll 123 vll vll c46 vll vll ull ull vll	sl Cl Sl Sl Sl Sl Sl Sl Sl Sl Sl Sl Sl Sl	itle Fai	2.0 2.0 2.0 1.0 2.0 1.0 1.0 5.0 3.0 2.0 3.0 1.0	+ Lone + 0 0 1 1 1 0 0 0 1
20 +	1	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Ji 3 Moran, Mr. Ji 1 McCarthy, Mr. Ti 3 Palsson, Master. 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nich 3 Sandstrom, Miss. El 3 Saundercock, Mr. 3 Andersson, Mr. An	fee fem fem ma ma ma ma fem fem fem fem fem fem ma ma . ma ma . ma ma . ma ma . ma	male r++ ale ale ale ale ale ale ale ale	null	DSp Par 1 1 0 1 0 0 1 1 1 1	0 0 0 0 0 0 0 0 1 2 0 1 0 5	A/5 21171 PC 1759 77 0N/02. 3101282 113803 373450 330877 8 17463 51 349909 2 347742 11 237736 36 PP 9549 113783 A/5. 2151 347082 3	Fare Ca 7.25 n .2833 7.925 n 53.1 C 8.05 n 8.4583 n .8625 11.075 n .1.333 n 0.0708 n 16.7 26.55 C 8.05 n 11.275 n	+ bin Emb+ vll C85 vll 123 vll vll vll vll vll vll vll vl	sl Cl Sl Sl Sl Sl Sl Sl Sl Sl Sl Sl Sl	itle Far	2.0 2.0 2.0 1.0 2.0 1.0 1.0 5.0 3.0 2.0 3.0 1.0 1.0 7.0	+ Lone + 0 0 1 1 1 0 0 1 1 0
20 + engerId Surv 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	1 + vived PC 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 0 0 0 1 1 0 0	3 Masselmani, Mrs .ass	fee	male r+- Sex A sle sle sle sle sle sle sle sle	null	SSp Par 1 1 0 0 0 1 1 0 1 1	0 0 0 0 0 0 0 1 2 0 1 0 5 0	A/5 21171 PC 1759 77 0N/02. 3101282 113803 373450 330877 8 17463 51 349909 2 347742 11 237736 36 PP 9549 113783 A/5. 2151 347082 3 350406 7	7.25 n	+ bin Emb+ vll C85 vll 123 vll vll vll vll vll vll vll vl		itle Far	2.0 2.0 2.0 1.0 2.0 1.0 5.0 3.0 2.0 3.0 2.0 3.0 1.0 7.0	+ lone + 0 1 1 1 1 0 0 1 1 1 1 1 1 1
20	1	3 Masselmani, Mrs ass	fee	male r	null	Sp Par 1 1 0 0 0 0 0 1 1 0 0 0	0 0 0 0 ST 0 0 0 1 2 0 1 0 5 0 0	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 E 17463 53 349909 2 347742 11 237736 36 PP 9549 113783 A/5. 2151 347082 3 359406 2	7.25 n	+ bin Emb+ vll C85 vll 123 vll vll E46 vll vll ull ull ull ull ull vll v		itle Fai	2.0 2.0 2.0 1.0 2.0 1.0 5.0 3.0 2.0 3.0 2.0 3.0 1.0 1.0	Lone
20	1	3 Masselmani, Mrs ass 3 Braund, Mr. Owen 1 Cumings, Mrs. Jol 3 Heikkinen, Miss. 1 Futrelle, Mrs. Jal 3 Allen, Mr. Willia 3 Moran, Mr. Jal 1 McCarthy, Mr. Tir 3 Palsson, Master 3 Johnson, Mrs. Osc 2 Nasser, Mrs. Nicl 3 Sandstrom, Miss. El 3 Saundercock, Mr. 3 Andersson, Mr. An 3 Vestrom, Miss. Hi 2 Hewlett, Mrs. (Ma	fee	male r Sex A S	null	0Sp Par 1	0 	Ticket A/5 21171 PC 17599 73 0N/02. 3101282 113803 373450 330877 E 17463 53 349909 2 347742 11 237736 36 PP 9549 113783 A/5. 2151 347082 3 350406 7 248706 382652 2	Fare Ca	+ bin Emb+ vll C85 vll 123 vll vll E46 vll vll vll vll vll vll vll vl	sl cl sl sl sl sl cl sl sl sl sl sl sl sl	itle Fai	2.0 2.0 2.0 1.0 1.0 1.0 5.0 3.0 2.0 3.0 1.0 1.0 7.0 1.0	Older Control Control

3. Prediction

```
/** Prediction */
// Select the required columns for training and testing
val selectedTrain = trainDf.select("Survived", "Pclass", "Sex", "Age", "Fare", "Embarked")
val selectedTest = testDf.select("PassengerId", "Pclass", "Sex", "Age", "Fare", "Embarked")
// Clean the data
val cleanTrain = selectedTrain.na.drop()
val cleanTest = selectedTest.na.drop()
// Cast the data to required types
val finalTrain = cleanTrain
  .withColumn("Age", col("Age").cast("Double"))
  .withColumn("Fare", col("Fare").cast("Double"))
  .withColumn("Pclass", col("Pclass").cast("Integer"))
  .withColumn("Survived", col("Survived").cast("Integer"))
val finalTest = cleanTest
  .withColumn("Age", col("Age").cast("Double"))
  .withColumn("Fare", col("Fare").cast("Double"))
  .withColumn("Pclass", col("Pclass").cast("Integer"))
  .withColumn("PassengerId", col("PassengerId").cast("Integer"))
// Fill missing values with the mean
val ageMeanTrain = finalTrain.agg(avg("Age")).first()(0).asInstanceOf[Double]
val fareMeanTrain = finalTrain.agg(avg("Fare")).first()(0).asInstanceOf[Double]
val trainFilled = finalTrain.na.fill(ageMeanTrain, Seq("Age")).na.fill(fareMeanTrain, Seq("Fare"))
val ageMeanTest = finalTest.agg(avg("Age")).first()(0).asInstanceOf[Double]
val fareMeanTest = finalTest.agg(avg("Fare")).first()(0).asInstanceOf[Double]
val testFilled = finalTest.na.fill(ageMeanTest, Seq("Age")).na.fill(fareMeanTest, Seq("Fare"))
```

```
// Preprocessing: StringIndexer and OneHotEncoder
val sexIndexer = new StringIndexer()
  .setInputCol("Sex")
  .setOutputCol("SexIndex")
val sexEncoder = new OneHotEncoder()
  .setInputCol("SexIndex")
  .setOutputCol("SexVec")
val embarkedIndexer = new StringIndexer()
  .setInputCol("Embarked")
  .setOutputCol("EmbarkedIndex")
// OneHotEncoder
val embarkedEncoder = new OneHotEncoder()
  .setInputCol("EmbarkedIndex")
  .setOutputCol("EmbarkedVec")
val assembler = new VectorAssembler()
  .setInputCols(Array("Pclass", "SexVec", "Age", "Fare", "EmbarkedVec"))
  .setOutputCol("features")
val logisticRegModel = new LogisticRegression()
  .setFeaturesCol("features")
  .setLabelCol("Survived")
val pipeline = new Pipeline()
  .setStages(Array(sexIndexer, embarkedIndexer, sexEncoder, embarkedEncoder,
    assembler, logisticRegModel))
```

```
val modelFit: PipelineModel = pipeline.fit(trainFilled)
val results = modelFit.transform(testFilled)
val predictionResults = results.select("PassengerId", "prediction")
predictionResults.write
    .format("csv")
    .option("header", "true")
    .option("header", "true")
    .option("delimiter", ",")
    .mode("overwrite")
    .save("/Users/harshilshah/Desktop/Scala assignments/Scala new assignment repo/CSYE7200-Harshil-Shah/assignment-spark-2/src/main/scala/results.csv")
```

+	+	+
Passeng	erId predic	tion
+	+	+
1	892	0.0
1	893	0.0
1	894	0.0
1	895	0.0
1	896	1.0
1	897	0.0
1	898	0.0
1	899	0.0
1	900	1.0
1	901	0.0
1	903	0.0
1	904	1.0
1	905	0.0
1	906	1.0
1	907	1.0
1	908	0.0
1	909	0.0
1	910	1.0
		4 61

Юl