Titanic Data - (Test 2)

bookmark_border

subject Machine Learning / AI casino 15 points

DESCRIPTION

A data set containing records of the RMS Titanic ship is provided in input file **titanic.csv**

Data set contains 891 observations containing 11 variables as follows:

- PassengerId: ID of the passenger (integer)
- Survived: Survived or Not (1 or 0)
- Pclass: Class of Travel (1, 2 or 3)
- Name: Name of Passenger
- Sex: Gender (male or female)
- Age: Age of Passenger
- SibSp: Number of Sibling/Spouse aboard (integer)
- Parch: Number of Parent/Child aboard (integer)
- Ticket: Ticket number (Random)
- Fare: Amount in dollars
- Cabin: Number of the Cabin allotted if any
- Embarked: The port in which a passenger has embarked. C - Cherbourg, S - Southampton, Q = Queenstown

Here's a preview of the data under consideration:

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leon	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisa	female	27	0	2	347742	11.1333		S

Based on this data set, write Python programs to perform the following operations:

1. Load the data set from the input file **titanic.csv**

2. Compute the **mean** and **standard deviation** for its quantitative data columns

- Compute the above-mentioned statistics for the columns **Age, SibSp, Parch** and **Fare**

- ***Print*** these values in a file named **output1.csv**

3. What percentage of total population survived?

- Calculate and ***print*** the percentage of people who survived based on the data provided in the data set
- Write this value in a file named **output2.csv**

4. Calculate conditional probability

- Calculate the probability that a passenger survived given that she is a female
- Calculate the probability that a passenger survived given that he is a male
- ***Print*** these two values in the same file **output2.csv** below the existing row created in Step 3

5. Analyse if there is a significant difference in the mean Age between the passengers of class 1 and 3 (**Pclass** column from the data set**)**

- ***Print*** the difference as **<mean Age of Class 1 passengers> - <mean Age of Class 3 passengers>** in the same file named **output2.csv** below the existing rows created in Steps 3 & 4

Input Format:

Read data from a file named **titanic.csv** present at the location **/data/training/titanic.csv**

Output Format:

- You have to create 2 files named **output1.csv** and **output2.csv** at the location **/code/output/outputn.csv**
- **output1.csv** should contain the values on 8 separate rows as follows:
    - Write the mean & standard deviation of **Age** column on first & second row respectively with values rounded to **3 decimal places**
    - Write the mean & standard deviation of **SibSp** column on third & fourth row respectively with values rounded to **3 decimal places**
    - Write the mean & standard deviation of **Parch** column on fifth & sixth row respectively with values rounded to **3 decimal places**

    - Write the mean & standard deviation of **Fare** column on seventh & eighth row respectively with values rounded to **3 decimal places**
- **output2.csv** should values on 4 separate rows as follows:
    - Value of percentage of total population survived calculated in Step 3 and rounded to **3 decimal places** in the form such as **35.623** on the first row
    - Values of two probabilities calculated in Step 4 rounded to **3 decimal places** on the second and the third row
    - Value of difference in mean Age calculated in Step 5 and rounded to **3 decimal places** on the fourth row

Sample Output:

Example: output1.csv will have data looking like this:

	A
1	25.567
2	12.123
3	0.456
4	1.302
5	0.213
6	0.756
7	30.234
8	50.678
9	

Example: output2.csv will have data looking like this:

	A
1	35.623
2	0.564
3	0.124
4	12.2
5	
6	

DATASETS

- Training dataset