



Edureka India

## Assessment report

### Machine Learning Assignment - 1 (B-14)

Report Access URL: <https://p.hck.re/TZSM>

Charan Kanaparthi

RANK<sup>\*</sup>

1 / 6

TOTAL SCORE

✓ 50/50

ATTEMPTED

3 of 3 questions

### Test time analysis

TEST INVITE TIME

Aug 05, 2020 11:45:51  
PM IST

TEST START TIME

Aug 07, 2020 06:27:53  
PM IST

TEST END TIME

Aug 13, 2020 12:34:35  
PM IST

TEST DURATION

5 day 18 hr 6 min 42  
sec of 7 day used.

### About Charan Kanaparthi



✉ Email ID

kvenkatcharan@gmail.com

### Detailed submission report

Python Project Questions

Questions attempted: 3 of 3

#	Questions (3)	No. of attempts	Result	Score (50/50)
1	Q1	1	Correct	16.67
2	Q2	1	Correct	16.67
3	Q3	1	Correct	16.67

## 1 Question 1

This assignment is a scenario-based assignment which uses Titanic Dataset and consists of 3 different questions. Read and understand the requirements and answer the questions carefully.

**Dataset:** Titanic disaster.

**Data Dictionary:**

**Variable | Definition | Key**

- survival | Survival | 0 = No, 1 = Yes
- pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd
- sex | Sex | M or F
- Age | Age in years
- sibsp | # of siblings / spouses aboard the Titanic
- parch | # of parents / children aboard the Titanic
- ticket | Ticket number
- fare | Passenger fare
- cabin | Cabin number
- embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton

**Variable Notes:**

- pclass: A proxy for socio-economic status (SES)
- 1st = Upper
- 2nd = Middle
- 3rd = Lower
- age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- sibsp: The dataset defines family relations in this way...
- Sibling = brother, sister, stepbrother, stepsister
- Spouse = husband, wife (mistresses and fiancés were ignored)
- parch: The dataset defines family relations in this way
- Parent = mother, father
- Child = daughter, son, stepdaughter, stepson. Some children travelled only with a nanny, therefore parch=0 for them.

**Dataset Path:**

The dataset **Titanic\_train.csv** is present at the location

location

res/Titanic\_train.csv

The dataset **Titanic\_test.csv** is present at the location

res/Titanic\_test.csv

### Problem Statement:

You are provided with the datasets about people from the Titanic disaster. Use the dataset resolve the following issues:

**Q1:** Find the relation of the following columns (having discrete values) with the "Survived" columns and answer the below questions:

- Pclass
- Sex
- Embarked

1. Find the total number of survivors from the 3rd PClass (Titanic\_train.csv)

Example:

If Total number of survivor from Pclass1

Output: 100

2. Find the total number of male who died in the accident (Titanic\_train.csv)

3. Find the total number of the survivor who embarked the ship from "Southampton" (Titanic\_train.csv)

### Hint:

Pclass relation with Survived Column:

Group | Total | Survived: 1 | 189 | 116

Sex relation with Survived Column:

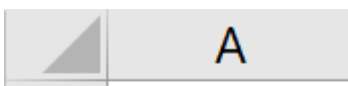
Group | Total | Survived: female | 262

Embarked relation with Survived Column:

Group | Total | Survived: C | 146 | 78

**\*\*\*Note:** Write the code only in solution() function and do not pass any arguments to the function. For predefined stub refer stub.py\*\*\*

Final Output Sample:




1	100
2	200
3	300

**NOTE:** Here, 100, 200 and 300 are the answer of 1st, 2nd and 3rd question respectively.

**Output Format:**


- Perform the above operations and write (written above as **print**) your output to a file named **output.csv**, which should be present at the location **output/output.csv**
- **output.csv** should contain the answer to **each question** on consecutive rows.

 [Download source code](#)

1



15/15

Partial scoring  
enabled 

2 Question 2

**Dataset: Titanic disaster**

**Q:** Some of the values in the "Age" column are missing. Use Linear Regression model to fill the missing values in the dataset.

(**Hint:** Dependent Variable(Age)) to fill(predict) the missing values.

1. Print the total number of cells having missing values in the Age column.

**Example:**

If Total number of cells with missing value is:  
100

**Output: 100**

2. Print the sum of the index number of all the cells with missing values.

**Example:**

If the Index Number of cells with missing value is: (4,6,20,40)

**Output: 70**

3. Print the mean of all the new values filled using linear regression. [For this first divide the training dataset into two halves, first half will contain only those rows which have missing values in 'Age' Column(let us say this dataframe

(df1), and the second half will contain the rows where you have valid numbers in 'Age' column(let us say this dataframe (df2)). Now we will train our model with df2 and predict the ages on the dataframe df1. Whatever age value we got for the df1 we will calculate the mean of it.]

**\*\*\*NOTE: Please use the features for predicting Age**  
**['Pclass','Survived','GenderLabel']**

**Example:**

If the new filled values are: (25.0,30.0, 30.0,35.0)

**Output: 30.0**

**Steps to be followed:**

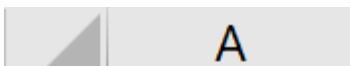
1. Load the Titanic\_train.csv file.
2. Calculate the missing values and count the occurrence. [Hint: You can use the isnull() with sum()]
3. Calculate the sum of the index where missing values are present. [Hint: You can use the is null() and pass the index to a list. Then you can sum the index of the list.]
4. Segregate the rows from the data having missing values(say in dataframe A) and rows from the dataframe having valid age values (say in dataframe B).
5. Convert the encode the string columns. So here we will encode the Sex column to "GenderLabel" columns
6. Now use the datarframe A from step 4 and fit into Linear Regression. [Hint: Use 'Pclass', 'GenderLabel,' 'Survived' as independent features.]
7. Now use the Linear regression model from step 5 and use it to predict the 'age' in dataframe B.
8. Once you get the predicted age from step 6, you can use the values to fit into the 'age' column of Dataframe B.
9. Calculate the mean for the Dataframe B having the age column and write the integer part of the mean. This will be the answer for part 3

**\*\*\*Note: Do not split the data into train\_test split\*\*\***

**Input Dataset path:**

res/Titanic\_train.csv

**Final Output Sample:**



1	100
2	200
3	300

**NOTE:** Here, 100, 200, and 300 are the answer of 1st, 2nd, and 3rd question respectively.

**Output Format:**

- Perform the above operations and write (written above as **print**) your output to a file named **output.csv**, which should be present at the location **output/output.csv**
- **output.csv** should contain the answer to each question on consecutive rows.

**\*\*\*Note:** Write the code only in solution() function and do not pass any arguments to the function. For predefined stub refer stub.py\*\*\*

[Download source code](#)

3



15/15

Partial scoring  
enabled ⓘ

### 3 Question 3

**Dataset:** Titanic disaster.

**Data Dictionary:**

**Variable | Definition | Key**

- survival | Survival | 0 = No, 1 = Yes
- pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd
- sex | Sex | M or F
- Age | Age in years
- sibsp | # of siblings / spouses aboard the Titanic
- parch | # of parents / children aboard the Titanic
- ticket | Ticket number
- fare | Passenger fare
- cabin | Cabin number
- embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton

After performing the analysis from the previous question, derive a new column called

"AdultOrChild" having categorical values as

"AdultOrChild" having categorical values as "Adult" or "Child" derived from Age column

**Hint:** A person having Age  $\geq 18$  is an "Adult" and the one having Age  $< 18$  is a "Child".

1. Find its relation with the "Survived" Column and print the total number of survivors.

**Example:**

If Total survived children: 100, Total survived adults: 200

**Output: 300**

2. Consider below features to create a Classification model and predict the survived category

- Pclass
- Age
- Sex (Encode values using LabelEncoder)

For the above prediction create a Confusion matrix for the model built by you and print the sum of all the elements of a matrix

**\*\*\*NOTE: 1. You should create the confusion matrix for the test data, not the training data.**

**2. Write the solution only in solution() function and do not pass any arguments to the function. For predefined stub refer stub.py\*\*\***

**Training Data: 'res/Titanic\_train.csv'**

**Testing Data: 'res/Titanic\_test.csv'**

**Example:** If the Confusion Matrix is [2 2

2 2]

(2+2+2+2)

**Output: 8**

**Hint:** Use Logistic Regression as the classification model

3. Use confusion matrix to print the accuracy of the model

Example:  $(2+2)/8 \times 100$

**Output: 50**

**\*\*\*NOTE: You should check the accuracy for the test data not the training data.**

**Steps to be followed:**

**Step 1:** In this question, you are supposed to read the CSV file using pandas.

**Step 2:** Print the total number of cells having missing values in the Age column. **Hint:** Using

.isnull().sum()

**Step 3:** Find the sum of all the index numbers of the missing values.

**Step 4:** Derive a new column called "AdultOrChild" having categorical values as "Adult" or "Child" derived from Age column. **Hint:** A person having Age  $\geq 18$  is an "Adult" and the one having Age  $< 18$  is a "Child".

**Step 5:** Find its relation with the "Survived" Column and print the total number of survivors. Obtain the complete dataset by combining it with the target attribute.

**Step 6:** Consider mentioned features to create a Classification model and predict the survived category. For the above prediction create a Confusion matrix for the model built by you and print the sum of all the elements of a matrix. **Hint:** Use `confusion_matrix(Y_train, Y_pred)`

**Step 7:** Use logistic regression on the `titanic_test.csv` data calculate accuracy score using:  
`round(accuracy_score(Y_train, Y_pred)*100,2)`

**Step 8:** Finally create a dataframe of the final output and write the output to `output.csv` which is present at `output/output.csv`

**Final Output Sample:**

	A
1	100
2	200
3	300

**NOTE:** Here, 100, 200 and 300 are the answer of 1st, 2nd and 3rd question respectively.

**Output Format:**

- Perform the above operations and write (written above as **print**) your output to a file named **output.csv**, which should be present at the location **output/output.csv**
- **output.csv** should contain the answer to each question on consecutive rows.

**\*\*\*NOTE:** For all the questions the numerical values saved in `output.csv` file should be in integer format with no decimals.

📄 [Download source code](#)



3



20/20  
Partial scoring  
enabled ⓘ

\* - This information is based on the data that is available untill Aug 13, 2020 12:34 PM IST. Your rank may be updated once all the candidates have successfully completed their tests.

Thank you for taking this test. All the best for further processes.

Note: This report is generated by HackerEarth. To know more, visit [www.hackerearth.com](https://www.hackerearth.com)