Madhav Institute of Technology & Science, Gwalior

Department of Information Technology
B. Tech. , Semester – IV
Machine Learning and Optimization
Assignment: II
Date of issue 6/03/2023 Date of Submission 11/03/2023
Maximum Marks 10

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.

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)
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)

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*****

Question 3

Dataset: Titanic disaster.

After performing the analysis from the previous question, derive a new column called "AdultOrChild" having categorical values as "Adult" or "Child" derived from Age column

Hint: A person having Age >=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.***