Enjoy Programming!

I hope that all of you are familiarized basic syntax and semantics of Python language.

Note that the best way to learn how to code and become an expert **is to code yourself, make mistakes and fix them.** There are no other short-cuts.

The following problem is well discussed by different ML practitioners in the past. So, please avoid copy paste from any other source.

If you have any questions about any of the Python commands, please check the manuals and documentation included with the package first.

PROGRAMMING ASSIGNMENT (PART II)

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

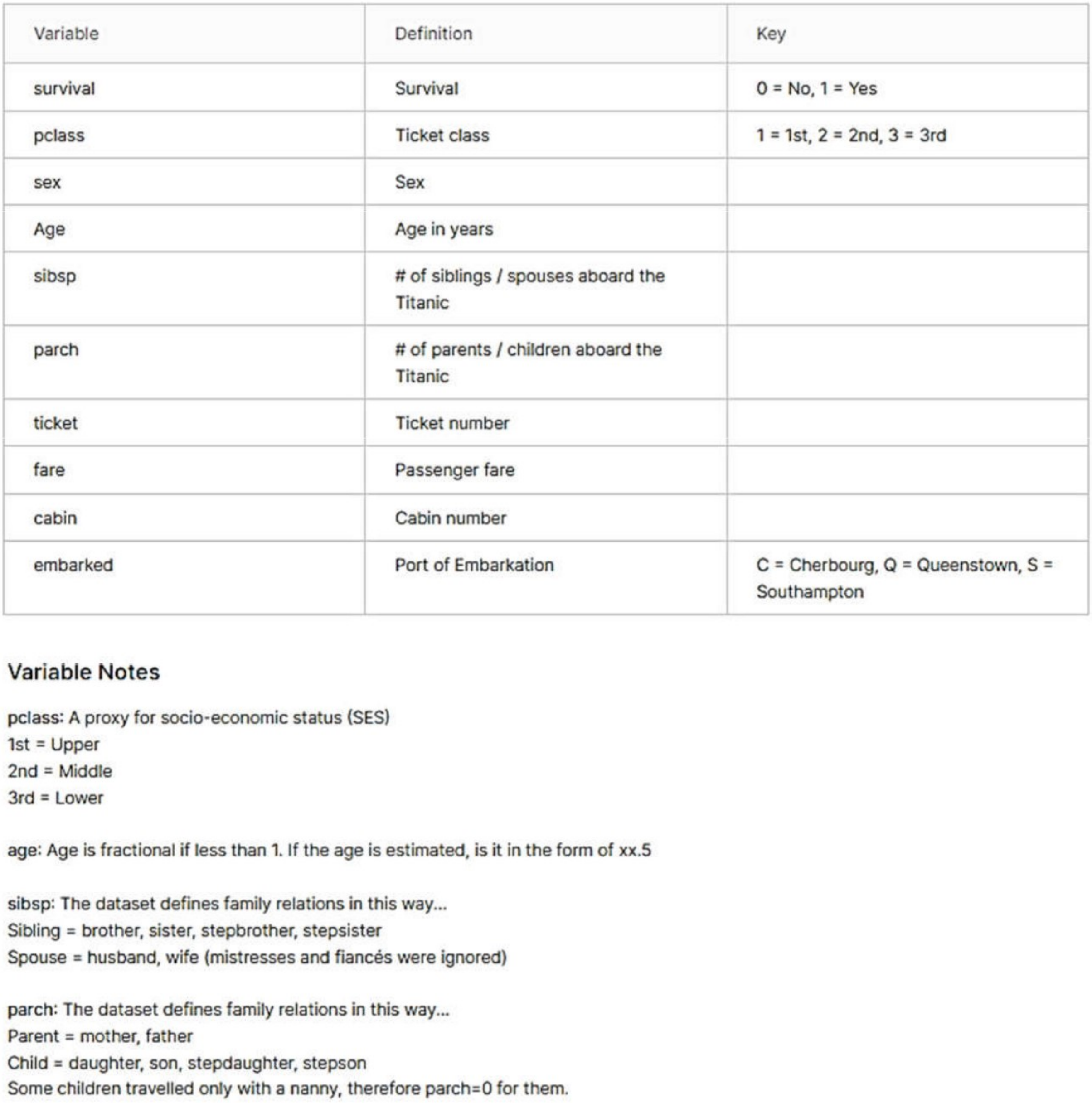
While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this assignment, you should build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data.

I. DATA VISUALIZATION

Titanic.csv data provided will contain the details of a subset of the passengers on board (891 to be exact).

Data details:



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
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.

Write and execute Python scripts to do the followings:

(i) Read CSV file & display information on the dataframe.

Hints: read_csv(), info() method

(ii) Display first 10 rows of the data.

(iii) Display first 5 rows of the data having the given columns only.

‘PassengerID’, ‘Name’, ‘Age’, ‘Sex’

II. DATA ANALYSIS

For data visualization, the popular packages are Matplotlib and Seaborn. More advanced functionality is available with Seaborn.

Write and execute Python scripts to do the followings:

(i) Plot the count of survived passengers.

(ii) Plot histogram of ‘Age’ column

Hints: hist() method

III. DATAWRANGLING & FEATURE SELECTION

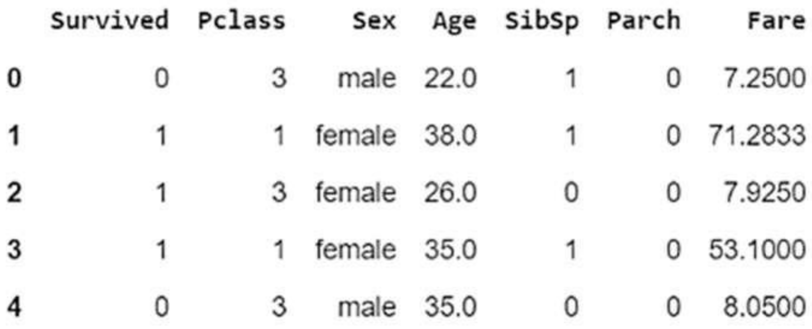
You can easily understand that all the columns (features) in the dataset are not significant for a binary classification problem to classify ‘survived’ or ‘not’. Also, you can see NaN values in the dataset. So, data pre-processing is required here.

Write and execute Python scripts to do the followings:

(i) Drop the following unnecessary columns.
‘PassengerID’,’Name’, ‘Ticket’, ‘Cabin’, ‘Embarked’

Hints: drop([…],axis=1,inplace=True) method

A sample for expected output:



```
   Survived  Pclass     Sex   Age  SibSp  Parch     Fare
0         0       3    male  22.0      1      0   7.2500
1         1       1  female  38.0      1      0  71.2833
2         1       3  female  26.0      0      0   7.9250
3         1       1  female  35.0      1      0  53.1000
4         0       3    male  35.0      0      0   8.0500
```

(ii) How many ‘NaN’ entries in ‘Age’ column? Replace all ‘NaN’ values in the ‘Age’ column with mean value of the ‘Age’ column vector. (Mean value replacement is a popular choice. It will not make a considerable damage to the data distribution in the column vector!). Please round off the mean value to two decimals.

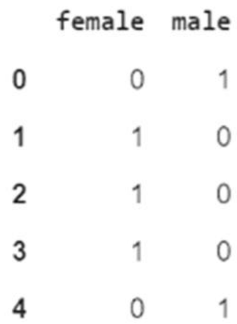
Hints: mean(), round(), fillna() methods

(iii) The entries in ‘Sex’ column are ‘Male’ or ‘Female’. ‘Pclass’ can have ‘1st’, ‘2nd’, or ‘3rd’. We should convert them to numerical values.

Hints:get_dummies() method

A sample for expected output:

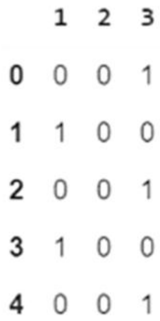
(a) 'Sex':



	female	male
0	0	1
1	1	0
2	1	0
3	1	0
4	0	1

The result has two separate column for 'Female' and 'Male'. It is obvious that the '0' value in 'Female' column means '1' in Male and vice versa (based on the given data). So, we need any one column only in the pre-processed dataset. (I hope that it is clear to you!! ☺).

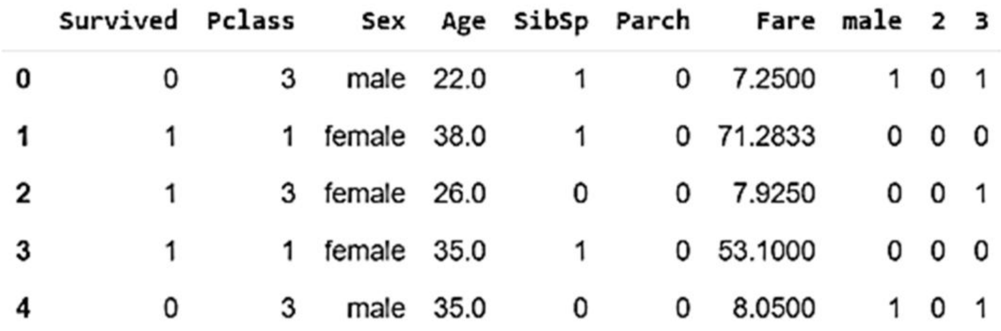
(b) 'Pclass':



	1	2	3
0	0	0	1
1	1	0	0
2	0	0	1
3	1	0	0
4	0	0	1

Here, 3 'Pclass' category and hence, we need any two columns in the results.

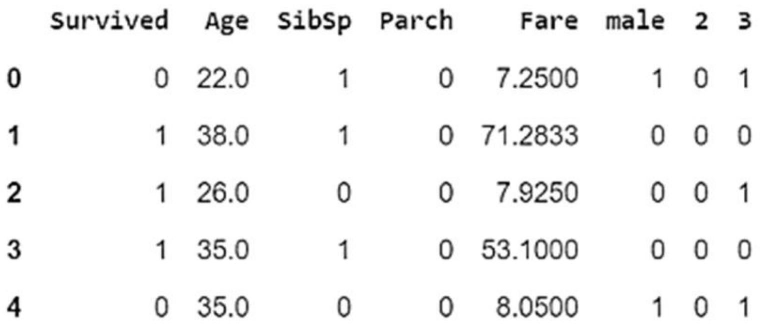
(iv) Concatenate the results of 'Sex' and 'Pclass' from previous step to get the following pre-processed dataset.



	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	male	2	3
0	0	3	male	22.0	1	0	7.2500	1	0	1
1	1	1	female	38.0	1	0	71.2833	0	0	0
2	1	3	female	26.0	0	0	7.9250	0	0	1
3	1	1	female	35.0	1	0	53.1000	0	0	0
4	0	3	male	35.0	0	0	8.0500	1	0	1

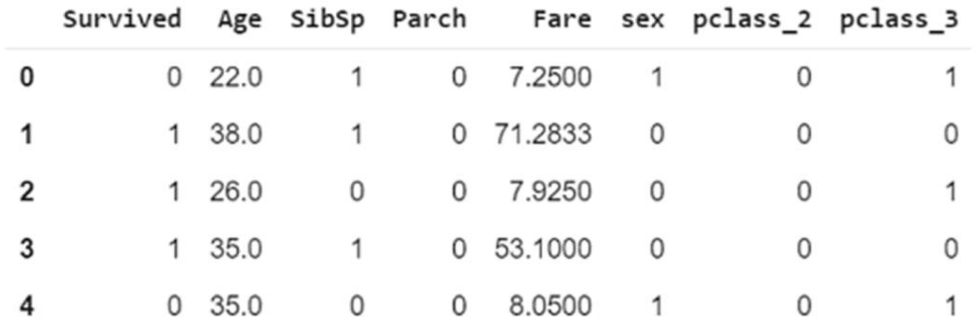
Hints:concat([],axis=1) method

(v) Next, drop ‘Pclass’ and ‘Sex’ from the data frame to obtain the following:



	Survived	Age	SibSp	Parch	Fare	male	2	3
0	0	22.0	1	0	7.2500	1	0	1
1	1	38.0	1	0	71.2833	0	0	0
2	1	26.0	0	0	7.9250	0	0	1
3	1	35.0	1	0	53.1000	0	0	0
4	0	35.0	0	0	8.0500	1	0	1

(vi) We can rename the column names as shown below (for convenience):



	Survived	Age	SibSp	Parch	Fare	sex	pclass_2	pclass_3
0	0	22.0	1	0	7.2500	1	0	1
1	1	38.0	1	0	71.2833	0	0	0
2	1	26.0	0	0	7.9250	0	0	1
3	1	35.0	1	0	53.1000	0	0	0
4	0	35.0	0	0	8.0500	1	0	1

(vii) Apply Z-score scaling with StandardScalar if mean and standard deviation are 0 and 1, respectively (optional in this assignment)

IV. TRAINING & TESTING

Write and execute Python scripts to do the followings:

(i) Make a ratio of 30% and 70% for test and train dataset.

(ii) Apply the following models:
(a) **Logistic regression**
(b) Neural Networks classifier

V. PERFORMANCE STUDY

Write and execute Python scripts to do the followings:

(i) Plot confusion matrix.

(ii) Find Precision, Recall, F1score, and Accuracy.

ALL THE BEST!