National University of Computer & Emerging Sciences, Karachi

Spring 2018 CS-Department

Midterm 2 (A)

2nd April 2018, 11:00 am – 12:30 pm

Course Code: CS481	**Course Name:** Data Science
Instructor Name: Dr Muhammad Atif Tahir	
Student Roll No:	**Section No:**

Instructions:

- Return the question paper.
- You are allowed to use PCs but all programs should be written in the answer sheet
- Read each question completely before answering it. There are 2 **questions and 2 pages**
- In case of any ambiguity, you may make assumption. But your assumption should not contradict any statement in the question paper.
- Show all steps clearly.

Time: 90 minutes. **Max Marks**: 10 points

Question 1: Consider the following table [3 Points]

Name	Field	Age	Marks
	C		-90
Ali	E		60
Ahmed	E		-10
Nida	C		70
	C		75

Perform following data cleansing operation on the given data. Write down the whole program in your answer sheet.

i. Drop column **Age** as it does not contain any value [0.75 Point]

ii. All empty strings in the **Name** column should be replaced by "---" [0.75 Point]

iii. In the **Field** column replace "C" with 0 and "E" with 1. The column must contain only numeric values after this operation [0.75 Point]

iv. Negative values are not permitted in **Marks** column. The invalid value in **Marks** column should be replaced with the average of all valid values in the same column [0.75 Point]

Question 2: Complete the following program [7 Points]

```
import numpy as np
from sklearn import datasets
from sklearn.cross_validation import KFold
# import only necessary classes. Any extra import (-0.5 points)

# load digits dataset
```

digits = datasets.load_digits()

print the number of samples and number of attributes [0.5 Points]

assign all data excluding target class to variable X [0.25 Points]
assign target class to variable Y [0.25 Points]

Now Divide Data into 4 Folds. For each fold, train and test the following models [0.5 Points]

For Fold 1; Decision Tree Classifier [1 Point]
For Fold 2: Support Vector Machine with linear kernel and value of C is set to 1 [1 Point]
For Fold 3: Naïve Bayes Classifier [1 Point]

For Fold 4; run kmeans with cluster size of 4. [1 Point] # Use cluster information as new feature of train / test data of that fold. [1 Point] # Afterwards apply knn classifier with k = 3 [0.5]

Screen Shot of the desired output of the program is shown below

```
The number of instances are:  1797
The number of attributes are:  64
Fold1: Accuracy using Decstion Tree 0.824444444444
Fold2: Accuracy using SVM: 0.937639198218
Fold3:  Accuracy using Naive Bayes: 0.824053452116

Fold 4
Training Data after adding cluster output as Feature:(1348, 65)
Testing Data after adding cluster output as Feature:(449, 65)
Fold4:  Accuracy using kNN classifier: 0.971046770601
```

[Hint] Look at the functions concatenation or column_stack. They may be needed for problem related to Fold 4. Also if needed, use help(KFold) in the program to get help about KFold class.

Appendix:

sklearn.cross_validation.KFold

class `sklearn.cross_validation.` **`KFold`** (*n*, *n_folds=3*, *shuffle=False*, *random_state=None*) [source]

K-Folds cross validation iterator.

Provides train/test indices to split data in train test sets. Split dataset into k consecutive folds (without shuffling by default).

Each fold is then used a validation set once while the k - 1 remaining fold form the training set.

Read more in the User Guide.

Parameters:

n : int

Total number of elements.

n_folds : int, default=3

Number of folds. Must be at least 2.

shuffle : boolean, optional

Whether to shuffle the data before splitting into batches.

random_state : None, int or RandomState

When shuffle=True, pseudo-random number generator state used for shuffling. If None, use default numpy RNG for shuffling.

BEST OF LUCK!