# National University of Computer & Emerging Sciences, Karachi

**Spring 2017/18 CS-Department**

**Midterm 1**

**26th May 2018, 10:30 am – 11:30 am**

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

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are 2 **questions and 1 page**
- 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**: 60 minutes. **Max Marks**: 10 points

**Question 1: Briefly answer the following questions. Each question should be answered in 3 – 4 lines including articles. Otherwise, answer will not be checked. [4 Points]**

a) What is data science?
b) Briefly discuss any two issues to consider during data integration.
c) What are the main steps of Data transformation
d) Discuss 2 ways to handle missing data

**Question 2**: You are given the following training examples. Each example has only one attribute, and the classification into positive / negative **[6 Points]**

| *Index* | *X* | *Label* |
|---|---|---|
| *1* | *1.0* | *Positive* |
| *2* | *2.0* | *Negative* |
| *3* | *4.0* | *Positive* |
| *4* | *5.0* | *Positive* |
| *5* | *6.0* | *Negative* |
| *6* | *7.0* | *Negative* |

Your main task is to evaluate the following algorithm that use a set *S* of training examples to classify the example with attribute value of *x.*

Algorithm:

Let $S_p, S_n$ be the sets of positive and negative examples in $S$.

If $S_p$ is empty classify $x$ as negative. If $S_n$ is empty classify $x$ as positive.

Otherwise, compute $u_p$, the mean of the $x$ values in $S_p$, and $u_n$, the mean of the $x$ values in $S_n$.

If $x$ value is closer to $u_p$ than it is to $u_n$ then classify $x$ as positive. Otherwise classify $x$ as negative.

**Example:** Using all the training examples above we have: $u_p = 3.33$, $u_n = 5$. Therefore, an example with $x = 2.5$ is classified as positive.

In case of Tie, randomly assign any class.

(a) Use leave-one-out cross validation to estimate the errors of Algorithm above [3 Points]
(b) Use 3 Fold CV to estimate the errors of Algorithm above. Run cross-validation using following permutations of the data points (1, 3, 4, 2 ,5 ,6) [3 Points]

***BEST OF LUCK!***