National University of Computer & Emerging Sciences, Karachi
Spring 2018/19 CS-Department
Midterm 1
26th February 2019, 11:00 am – 12:00 pm

Course Code: CS481	**Course Name:** Data Science
Instructor Name: Dr Muhammad Atif Tahir	
Student Roll No:	**Section No: GR3**

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are 3 **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: 60 minutes. **Max Marks**: 12.5 points

Question 2: As a data scientist, you got a project from the Traffic Police to know the factors of theft in certain areas. What are the main steps you will do to accomplish the above mentioned task? Explain **[3 Points**]

Ans: Student basically have to discuss data science process according to above task including Setting the Research Goal, Retirving Data, Data Preparation, Data Exploration, Data Modelling and Presentation / Automation

Question 3: Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "-". Half of the data set is used for training while the remaining half is used for testing **[4 Points]**

(a) Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data? Show error using clear formula.
50% i.e 1 – 0.5x1.0 + 0.5x0.0 = 0.5

(b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2. Show error using clear formula.
50% i.e.1 - 0.8 x 0.5 + 0.2 x 0.5 = 0.5

(c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive? Show error using clear formula.
33.3% = 1 - 2/3 x 1.0 + 1/3 x 0 = 0.33

(d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3. Show error using clear formula.
=1 – 2/3 x 2/3 + 1/3 x 1/3 = 44.4

BEST OF LUCK!