Data Science
Assignment 1
Prepared By
Dr Muhammad Atif Tahir

Deadline: 26th Feb 2021

Instructions:

- Only soft copy for 1 & 2. Hardcopy for Question # 3.
- You need to work as a Group of 2. But Question # 3 is individual.
- Filename Format: studentid1_studentid2.pdf
- Only one of you need to upload on Google Drive.

1. Download the following paper from

   https://arxiv.org/abs/1712.08971

   Human-Centric Data Cleaning

   Write a one-page summary of the above paper. Clearly mention the problems associated with Data Cleaning. [10 Points]

2. According to authors from the paper "Ziawasch Abedjan et al "Detecting Data Errors: Where are we and what needs to be done?" Proceedings of the VLDB Endowment, Vol. 9, No. 12, 2016", some existing open source tools are not good enough to correct different types of data errors. You need to evaluate these different type of open source tools (2 tools) mentioned in the paper along with Python (You can also look at other sources if these tools are not free) using data sets provided data.zip plus one missing values dataset downloaded from UCI website. Is it possible to clean these data sets using open source tools? If No then why and if Yes then provide the main steps (Two page maximum). There will be demo as well of selected groups [20 Points]

3. Use 3 Fold CV using kNN classifier to find the accuracy, precision, recall and F-measure of the following data points. Use city block distance metric i.e. *abs (a - b) and k = 1* and 3 [20 Points]

Attribute 1	Attribute 2	Attribute 3	Label
1	4th digit of student ID	1	A

	(only numeric characters e.g. K171234 will become 171234 and 2 is the 4th digit)		
2	3	5	A
1	1	1	A
1	5th digit of student ID (only numeric characters)	4	B
1	2	3	B
4	3rd digit of student ID (only numeric characters)	5	B