Assignment # 1

Due Date: 17th October, 2021

Assignments are to be done individually. No late assignments are accepted. Submit your source code with pdf containing achieve results with graphs on Google Classroom.

Write your name and e-mail id in a comment line in on top of each source file. You are required to submit a single zip file containing all code files (.py and .ipynb) an archive of your documentation and ipython notebook on Google Classroom.

Data preprocessing is required for incomplete and redundant data. The data has been collected from different sources and has irrelevant or wrong information. The "derm.csv" dataset requires preprocessing for mining purposes. You have to apply following operation on it.

1. **Data Cleaning**
   Apply data cleaning by use of imputation and KNN. Show the achieve results by applying both mechanisms. Secondly explain which approach is better and how.

2. **Noise Removal**
   The data may contain an incorrect value which is known as noise. Apply any two smoothing methods which we have discussed in class and entropy based descritization and show achieve results on each operation.

3. **Data normalization**
   The normalization is also a very necessary step for applying any algorithm. Show the results by bringing any two attributes on same scale.

4. **Cosine Similarity**
   The "spam.csv" is attached on which you have to apply cosine similarity. Design the similarity and dissimilarity matrix of any 20 most frequently used words from the document.

Grading Policy
The assignment is given for learning purposes. No late assignments will be accepted. **In case of any plagiarism found you will get zero credit**. If you have any query contact at i181604@nu.edu.pk.

BEST OF LUCK