Assignment 2: K-NN algorithm

1 Instructions

- Use python programming language for your implementation.
- Use appropriate approach if you find some attribute is missing in your data.
- Report must contain step-wise description of your implementation and analysis of results. Since data analysis is a crucial task for any machine learning algorithm, report should demonstrate detailed analysis of results and conclusion. It should also clearly mention the steps to run your code.
- Do not use any direct in-built functions of libraries to implement K-nn algorithm, otherwise -10 will be deducted.

2 Dataset

- Spam mail dataset: https://www.kaggle.com/venky73/spam-mails-dataset. Download the dataset from the given link which also has description about the dataset.

3 Problem statement: K-nn algorithm

1. Implement a k-NN classifier and measure the classification accuracy on the test instances. Classification accuracy is defined as the percentage of the total number of correctly classified instances to the total number of test instances. Use a train-to-test split ratio of $80:20$. **25**
2. Vary the value of k (depending on the number of classes) with three different similarity/distance measures such as a) cosine similarity, b) Euclidean distance, and c) Manhattan distance and evaluate the performance of your classifier on each of them independently. Compare their performances and analyse the results. **10+10+10+5**
3. Plot your results in different graphs (x-axis: k, y-axis: accuracy) for all the three metrics. What trends can be observed from the graphs? **10+10**
4. A brief report explaining the procedure and the results. **20**