Assignment 3

Objective

You are required to demonstrate an understanding of **classification using Nearest Neighbour & NearestCentroid**. You will demonstrate your knowledge by making use of an **ipynb** file.

Provide any references where applicable (such as websites, books, etc) that were used during the completion of this assignment

Submission: Linked through brightspace portal

Tasks (total points = 140):

1. Define Big data and big data analytics and give examples of each. Also, why loading big data in batches is an effective strategy, please provide an example of batch operation using dataframe. Finally, which of the 3V relates to batch operations? **(20 points)**
    - Definition 3 points + Example (3 points) = 6 * 2 (big data and big data analytics) = 12 points
    - Discussion on the batch (2 point) and link with dataframe (3 points), connection with 3Vs (3 points) = 8 points
2. Answer the following questions (**20 Points**):
    - What is the difference between clustering and classification and discuss examples (6 points)
    - How useful is big data analysis for increasing business revenue (4 points)
    - Discuss some of the challenges of ethics concerning data (5 points)
    - Discuss map, reduce, and filter from python language. Highlight the benefit of each. (5 points)
3. The table below reports the pairwise distances between a set of 8 labelled training examples and a new query example q. **(10 points)**

Example	Class	Distance to q
1	over	3.6
2	under	3.0
3	over	1.3
4	under	5.6
5	under	4.0
6	under	2.8
7	over	4.3
8	over	1.2

a) What class label would a 3-NN classifier assign to q? And provide reason (3 points)

b) What class label would a 4-NN classifier assign to q? And provide reason (3 points)
c) What class label would a weighted 4-NN classifier assign to q? And provide reason (4 points)

- **Answer Policy**: Following describes the policy:
    1. 2 Points for the correct label and 1 point for reasoning.
    2. 2 Points for the correct label and 1 point for reasoning.
    3. 2 Points for the correct label and 2 points for reasoning

4. What is the difference between overfitting and underfitting? And what is the correct approach in relation to overfitting and underfitting? Use visualisation and description to explain **(10 Points)**

- **Answer Policy**: Following describes the policy:
    - 6 points = 3 points each for a descriptive and visual explanation.
    - 4 points for the correct approach.

5. You will use the wine for this task (dataset: **sklearn.datasets.load_wine()**), perform NN and NearestCentroid on the dataset to perform classification. Select different parameter values (i.e., parameter tuning) and discuss the influence. Finally, report result of comparisons (consult lecture notes) **(80 Points)**

- **Answer Policy**: Following describes the policy:
    - 20 points = 10 points for NN, and 10 points for parameter selection and tuning
    - 20 points = 10 points for NearestCentroid, and 10 points for parameter selection and tuning
    - 10 points for reporting the winner
    - 10 points for reporting the computational comparison
    - 10 points for visualising decision boundary