**Sultan Qaboos University**

**Department of Computer Science**

**COMP4603 - FL 2022**

**HW1 - Introduction to Machine learning (Classification problem)**

**Dataset:** the breast cancer dataset we will use for this HW was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. It consists of 699 samples (instances). The aim is to classify each instance as benign or malignant.

The following is the description of the dataset:

| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 699 | **Area:** | Life |
| --- | --- | --- | --- | --- | --- |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 10 | **Date Donated** | 1992-07-15 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 827435 |

More details are given in the accompanying text file “breast-cancer-wisconsin\_features.txt” and in the following website:

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>

**Questions**:

1. Load the dataset as a dataframe with column names included.
2. Show dataset description.
3. Check if there are missing values or not.
   * If so, repair the dataset either by deleting the corresponding rows or by replacing the missing values with other relevant values that you believe are more representative.
4. Show some scatter plots using pairs of features. Provide a brief discussion.
5. Extract the features into a matrix X and labels (last column in the dataset) into a vector y.
   * Convert the labels to 0, for benign, and 1 for malignant.
6. Split the dataset into training and test sets.
7. Train a K-NN classifier with the number of neighbors k=1.
8. Evaluate the trained classifier on the test set.
9. Report classifier performance in terms of classifier accuracy on the test set.
10. Discuss the results.
11. Apply feature scaling (using minmax and z-score scalers) and redo 7), 8) and 9).
12. Compare and discuss the results obtained before and after feature scaling.
13. Apply dimensionality reduction using feature selection. Keep only 2 features.
14. Show the scatter plot of the reduced dataset.
15. Redo 7), 8) and 9).
16. Compare and discuss the results obtained before and after feature selection.

**Note**: You have to provide:

1. a well commented and organized Python code (.py).
2. A report document.