Assignment 2

Instructions:

1) Use **Python** programming. **You may use libraries**.
2) Handle missing data as and when required using any approach.
3) There are two questions, each of 50 marks. You will be submitting two python code files named as **"q1.py"** and **"q2.py".**
4) You will prepare a **README** file to explain how to execute your code.
5) You will print the outputs in a **".txt"** file and also provide the plots.
6) All source code files, results files and documents should be kept in a folder named **"roll1_and_roll2_a2". Zip the folder and upload it on Moodle**.
7) Kindly download the dataset from: " https://archive.ics.uci.edu/ml/datasets/Lung+Cancer "

Question 1: Unsupervised Learning (30)

1) Apply **PCA** (select number of components by preserving **95%** of total variance). **(in-built function allowed for PCA)**.
2) Plot the graph for PCA**.**
3) Using the features extracted from PCA, apply **K-Means Clustering**. Vary the value of **K from 2 to 8**. Plot the graph of **K vs normalised mutual information (NMI)**. Report the value of K for which the **NMI is maximum**. **(in-built function not allowed for K-Means).**
4) Prepare a **report** including all your results.

[10+5+10+5]

Question 2: Supervised Learning (70)

1) Normalise the data using **Standard Scalar Normalisation**. Randomly divide the Dataset into 80% for training and 20% for testing. Encode categorical variables using appropriate encoding method **(in-built function not allowed for normalization, sampling and encoding)**.
2) Implement the **binary SVM classifier** using the following kernels**: Linear, Quadratic, Radial Basis function**. Report the **accuracy** for each. **(in-built function allowed)**.
3) Build an **MLP classifier (in-built function allowed)**. for the given dataset. Use **stochastic gradient descent** optimiser. **Keep learning rate as 0.001 and batch size of 32**. Vary the number of hidden layers and number of nodes in each hidden layer as follows and report the **accuracy** of each:
    a. 1 hidden layer with 16 nodes
    b. 2 hidden layers with 256 and 16 nodes respectively.
4) Using the best accuracy model from part 3, vary the learning rate as 0.1, 0.01, 0.001, 0.0001 and 0.00001. Plot the **learning rate vs accuracy** graph.
5) Use **forward selection method** on the best model found in part 3 to select the best set of features. **Print the features**.
6) Apply **ensemble** learning (**max voting** technique) using SVM with quadratic, SVM with radial basis function and the best accuracy model from part 3. Report the **accuracy**.
7) Prepare a **report** including all your results.

[10+10+20+5+10+10+5]