Mandatory Assignment #1

Assignment specification

1. Group Size= **2 Only**
2. A Jupyter notebook saved as **.ipynb file**. Please use comments and/or markdown cells wherever necessary explanation is required.
3. You must score above **80 marks** to pass the assignment.
4. Time: 3 weeks

Please address the following questions in your submission.

Problem 1: Regression Problem (20 marks)
The data in the file regression_*housedata.csv* are collected from 1,000 homes being sold in Oslo. The response variable of interest is the Price (the price of the house). The input variables are bedrooms, sqft_living (the living space area), sqft_lot (the area of the land the house sits on), floors (the number of levels of the house), sqft_above (area of the house excluding the basement), sqft_basement (basement area).

Use Multi-Linear Regressor, Decision Tree Regressor, and Support Vector Regressor to build a regression model for the prediction of house prices. Perform Model Evaluation using two metrics: Root Mean Squared Error and Coefficient of Determination. Which of the regression models is the best fit for the data?

Problem 2: Classification Problem (20 marks)
Using the Brain Cancer data set to fit classification models in order to classify whether a given brain cancer is diagnosed as "Meningioma", "LG glioma", "HG glioma", or "Other".
Explore Polynomial regression, naive Bayes, and KNN models. Describe your findings. And using appropriate metrics to tell which is the best model.
The instructions to download dataset are given in the following link: https://islp.readthedocs.io/en/latest/datasets/BrainCancer.html

Problem 3: Clustering Problem (25 marks)
The data in the file *clustering_diabetesdata.csv is* collected from 768 patients tested for diabetes. The dataset consists of following features:
1. Number of times pregnant
2. Plasma glucose concentration)
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

The following tasks need to be performed:
1. Pre-process the data as it contains some values which are equal to zero. **(5 marks)**
2. Choose an optimum value of k and use k-Means to identify any clusters in the dataset. **(10 marks)**
3. Use Hierarchical clustering to identify any clusters. Draw the dendrogram. (**10 marks)**

Problem 4: Multi-Layer Perceptron Problem Using Keras (35 marks)

The data scientists at one of the retail stores have collected 2019 sales data for different products across various stores in different cities. The data in the file *deep_learning_task_dataset.csv* consists of 5000 datapoints and consists of both input and output variables, the description of which is given in the table below. Divide the dataset into training (80%) and test (20%). You need to predict the sales for test data set.

Table 1: Input and Output Variables for Training Dataset

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

The following tasks need to be performed:

1. Pre-processing of dataset. (**10 marks**)
2. Define the architecture of your Deep Learning Model. Use the markdown cell to explain the architecture of your model. **(5 marks)**
3. Training and testing your model. **(10 marks)**
4. Calculate the $\mathbf{R^2}$ of your predictions. The closer the value of $\mathbf{R^2}$ to 1, the higher points you will score. **(10 marks)**