Assignment 2 Guidelines

- Deadline for submission is Monday 12th of December 03:00 pm.

- Submission will be through the following link:
  https://forms.gle/AWWUQJwvyWQxonkF6

- The name of the notebook should be similar to this: T1_46_1234_T1_46_1235
  - This means that the two students worked on this notebook are from T1 with IDs (46-1234 and 46-1235).

- Please write the members' **name, ID** and **Machine Learning Tutorial number** inside the notebook.

- The assignment will be in group of 2. So 1 or 3 are not accepted.

- You can work with a colleague from different tutorial number.

- Please submit the notebook that has all cells run and the outputs produced and shown in the notebook. **PLEASE DONOT REMOVE** the outputs.

Assignment Description

- You are provided 3 datasets below:
  - https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
  - https://www.kaggle.com/datasets/mastmustu/income (Note: you need to convert the problem into a binary classification to find whether the income is above 45K or not.)
  - Gene expression csv in the zipped file.

- For each of the datasets provided apply proper scaling and feature engineering.

- You need to train and test 3 different classification models (logistic regression, KNN, SVM) on each dataset separately so in total you will train and test 9 different models.

- For the KNN, try 20 different K values, select the best k using the elbow method **with plotting**.

- For the SVM:
  - Compare linear kernel versus 2 non-linear kernels. Select the best.

- Compute accuracy, precision, recall and F1-score for each of the 9 models.

- What is the model that outperformed other models on heart disease dataset? Justify your reasons?

- What is the model that outperformed other models on income dataset? Justify your reasons?

- What is the model that outperformed other models on gene expression dataset? Justify your reasons?