Lab 4: Clustering

Perform an Exploratory Data Analysis (EDA), Data cleaning, Building clustering models (at least three) for prediction, Presenting resultsusing on the datasets in Lab1

Exercise 1: Spam Classification

Spam or not spam
Using **spam.csv** dataset

Exercise 2: Predict the onset of diabetes based on diagnostic measures

Diabetes or not
Using **diabetes.csv** dataset
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.
For details: https://www.kaggle.com/uciml/pima-indians-diabetes-database#

Exercise 3: Mushroom Classification

Safe to eat or deadly poison.

Using **mushrooms.csv** dataset.

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended.

For details: https://www.kaggle.com/uciml/mushroom-classification

Exercise 4: Classification of Students' Academic Performance

The students are classified into three numerical intervals based on their total grade/mark:

•Low-Level (L): interval includes values from 0 to 69

•Middle-Level (M): interval includes values from 70 to 89

•High-Level (H): interval includes values from 90-100.

Using **xAPI-Edu-Data.csv** dataset

The dataset consists of 480 student records and 16 features. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, grade Level and section. (3) Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

For details: https://www.kaggle.com/aljarah/xAPI-Edu-Data

***