# CENG 414

## Introduction to Data Mining

Spring 2021-2021

## Homework 2

Due date: 01 05 2022, Sunday, 23:55

# Introduction

In this assignment we will work on Classifiers and Classifier Accuracy Calculations. We will elaborate on two related sub tasks.

# Tasks

**Task 1.** Implementing the k-NN algorithm.

In this part of the assignment, you will implement the well-known k-NN algorithm yourself in Python. Remember that it is an instance based learning algorithm including the following main steps:

- Given the instance whose label to be predicted ($x_q$), find k nearest neighbors of $x_q$ within the training data set. At this step, you are not expected to implement an index structure, just calculate the similarities and find k nearest neighbors.
- Decide for the label of $x_q$ according to the majority vote for predicting categorical labels and weighted mean for numerical labels.

**Task 2.** Applying the k-NN algorithm.

In this part of the assignment, you will apply your algorithm to a real-world dataset.

- You will use the 'Auto MPG' dataset. Dataset is provided for you as an attachment in ODTU-CLASS. You can find more information about the dataset in the following link. However, please download the attached dataset as it is slightly different from the one in the link. https://archive-beta.ics.uci.edu/ml/datasets/auto+mpg
- You will predict the 'mpg' (fuel consumption in miles per gallon) of cars.
  - Report the result for k values from 2 to 10, under 3-fold cross validation. Since your prediction results are numeric values, report the prediction performance in terms of MSE, RMSE and MAPE.
  - Report the prediction time for all cases.

**Task 3.** Comparing with the classifiers in the scikit-learn library.

In this part of the assignment, compare the performance of your best implementation with the following supervised learning methods in scikit-learn library: KNeighborsRegressor and DecisionTreeRegressor.Note that you may need to adapt the domains of the attributes according to the classifier. You do not need to optimize the parameters for these classifiers. Just use the default settings.

- Report the results under 3-fold cross validation in terms of MSE, RMSE and MAPE.
- Report the prediction time for all cases.

# Submission

Submissions will be done via ODTUCLASS. You are expected to submit a zip file containing your code and report presenting the analysis result.