**Artificial Intelligence Final Project (Spring 2021)**

Please choose **one of the following two projects** as your final project. Before starting working on the project, please read Chapter 2 of the textbook “End-to-End Machine Learning Project”, and become familiar with the general procedures of a data analysis project that uses machine learning methods.

**Choice 1: Regression on Ames Housing Dataset**

You can load the Ames housing dataset from [here](http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls).

You can find a description of the variables [here](http://jse.amstat.org/v19n3/decock/DataDocumentation.txt).

Take note that for categorical variables, NA here does not mean a missing value, but should be treated as a separate category.

1. Visualize the univariate distribution of each continuous input variable, and the distribution of the target variable. Write down anything you find worth noting. Is there something that might require special treatment?
2. Visualize the dependency of the target on each continuous feature via scatter plot.
3. Split data in training and test set. Do not use the test-set unless for a final evaluation in Step 6. Convert each categorical variable into numerical variables using [one-hot-encoding](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html).

***Example of one-hot encoding:***

*Gender: Male -> (1, 0), Female -> (0, 1)*

*Ethnicity: 1. Caucasian, 2. African American, 3. Hispanic, 4. Asian, 5 Native American, 6 Pacific Islander*

*One-hot encoded labels: 1 → (1, 0, 0, 0, 0, 0)*

*2 → (0, 1, 0, 0, 0, 0)*

*6 → (0, 0, 0, 0, 0, 1)*

1. Evaluate Linear Regression, Polynomial Regression, Decision Trees, and Neural Network. Note that you need to decide the choice of hyper-parameters for the models, such as the degree of polynomial regression, the maximum depth of random forests, and the number of layers for neural networks.
2. Based on your models, what are the most important factors for deciding the price of houses?

**Choice 2: Classification on the Telco-Churn Dataset**

The dataset and its description is available at [Kaggle](https://www.kaggle.com/blastchar/telco-customer-churn). The goal of this task is to analyze the behavior of telecom customers and understand what factors are important to retain customers.

1. Visualize the univariate distribution of each input variable and the target variable “churn”.

2. Split data into training and test sets. Convert each categorical variable into numerical variables using [one-hot-encoding](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html).

***Example of one-hot encoding:***

*Gender: Male -> (1, 0), Female -> (0, 1)*

*Ethnicity: 1. Caucasian, 2. African American, 3. Hispanic, 4. Asian, 5 Native American, 6 Pacific Islander*

*One-hot encoded labels: 1 → (1, 0, 0, 0, 0, 0)*

*2 → (0, 1, 0, 0, 0, 0)*

*6 → (0, 0, 0, 0, 0, 1)*

3. Evaluate the following classification models:

a. Logistic Regression

b. Support Vector Machine

c. K Nearest Neighbors

d. Decision Trees

e. Random Forests

Note that you need to decide the choice of hyper-parameters for the models, such as the value of k for k nearest neighbor method and the maximum depth for the random forest method.

4. Choose the best model by analyzing the accuracy, precision, recall, and F-1 score.

5. Which types of customers are less likely to end the service?

**Submission Instructions:**

- (80 pts) Submit your notebook as a pdf file including:

- code

- execution results

- description of what you do

- interpretation of the results

- (20 pts) Record a 5 minutes videos presenting your results.You should assume that the target audiences are not familiar with artificial intelligence. For Choice 1, the audiences are home buyers, and for Choice 2, the audiences are salespersons of the company. It is recommended to upload the video to YouTube and then submit a link to that video to Blackboard.

- The notebook and the video link should be submitted to Blackboard by **Wednesday, May 26th at 11:59 PM**