Problem description

The purpose of the semester-long project is to give you hands-on experience working with the machine learning pipeline. For the first stage of the project, we would like you to identify your guiding question, find the dataset you plan to use to answer the question, and perform preprocessing using pandas and sklearn. Your proposal should be 1-2 pages, excluding any figures.

Deadline

Please send your project proposal draft to your assigned mentor TA by September 24th, so that they can provide you feedback before you turn in your proposal to Andras. The draft will be graded by your TA for completion and effort. The finalized project proposal is due at the end of the day on September 30th via Gradescope.

Requirements

Clearly describe the problem you want to solve. (4 points)

- What is the target variable?
- Is the problem regression or classification?
- Why is this interesting/important?

Describe the dataset. (8 points)

- Number of data points and number of features.
- If dataset is not well-documented, write a description for each feature (if feature is categorical, describe each category; if feature is numerical, include the unit of the quantity and what it measures)
- If dataset is from Kaggle/UCI/already described, write a short description about 2-3 public projects where the data has been used, and how the features were used.

Preprocess the dataset. (8 points)

- Apply MinMaxEncoder or StandardScaler on the continuous features
- Apply OneHotEncoder or OrdinalEncoder on categorical features
- Apply the LabelEncoder on the target variable if necessary.
- Describe why you chose the preprocessor you used for each feature.
- How many features do you have in the preprocessed data?