## Rubric: Midterm Project Report

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, perform exploratory data analysis, split the dataset, and perform preprocessing using pandas, matpotlib, and sklearn. Your proposal should be maximum 2 pages, excluding any figures, tables, and references.

### Deadline

Please submit your midterm project report on gradescope by October 13th 11:59pm. We highly encourage you to send your draft report to your assigned mentor TA a couple of days before the submission deadline, so that they can provide you feedback before you turn in your final version on gradescope. The report will be graded by your TA for completion and effort. If you have any questions, please post a question on piazza or come to Andras’s office hours on monday.

### Requirements

We describe what sections you should include in your report and what you should describe in those sections. Your goal is to write as if you were writing a scientific publication. Please use the section titles below. This will help the grading process.

**Introduction** (5 points)

Clearly describe the problem you want to solve.

* What is the target variable?
* Is the problem regression or classification?
* Why is this interesting/important?
* 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 the dataset is from Kaggle/UCI/already described, write a short description about 2-3 public projects or publications where the data has been used, what did the authors find, what ML question did they solve and how successful they were.

**Exploratory Data Analysis** (15 points)

* Perform a thorough EDA on each column in your dataset and pay special attention to your target variable
  + Use .describe or .value\_counts, create histograms or bar plots as appropriate
  + Create other type of figures using two columns
  + Create a scatter matrix
* There is no need to include everything in the report.
* Choose at least three but maximum five figures that you think are interesting or important and one of those figures should show the target variable. Do not include three figures of the same type (e.g., you’ll lose points if you submit three bar plots). Use at least two figure types.
* Write a figure caption for each figure.
* Make sure your figures are publication-quality figures. Please print out your report before you submit it and make sure the figures are readable, the font size is similar to the text font size, axes are labelled, the figures have sufficient resolution (usually 300 dpi).

**Data preprocessing** (10 points)

* Discuss how you split the dataset and why.
* Is your dataset IID?
* Does it have group structure?
* Is it a time series data?
* How should you split the dataset given your ML question to best mimic future use when you deploy the model?
* 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?

**References** (2 points)

* Cite previous work and add your citations to the ‘References’ section which should be the last section of your report.

**Github repository** (3 points)

* Please create a public github repository for your project and include the link in your report.
* For now, there should be three folders in your repository:
  + data: download your dataset into this folder.
  + figures: any figures you make should be saved in this folder in png, jpg, or pdf format. You can and should have more figures in here than what you have in your report.
  + src: your python notebooks should be in this folder. Make sure that your code reads in the data from the *data* folder and saves the figures directly into the *figures* folder.