Project-2 Description

This is a continuation of the Project-1. In Project-1 the focus was on exploratory data analysis and visualizations. Project-2 focuses on machine learning.

Data (5 pts)

Use the data that you prepared in Project-1. Choose one of the variables as the output variable *y* and the rest (or subset of the rest) as the input variables **x** (features). If you think that none of the available variables are interesting as an output variable, you can add additional variable(s) to your dataset. Also, if you feel that the available data is not enough for learning, feel free to get more samples.

You may want to convert your categorical features using dummy variables, and to preprocess your data with normalization, etc.

The data should be split into training / validation / test splits. Use your best judgement to decide on the size of each split. Common choices are 80/10/10, 70/15/15, 60/20/20.

Modeling and Write up (20 pts)

- The goal is to train and evaluate at least three different models.
- The hyperparameters should be tuned using the validation split or using the cross-validation approach.
- You need to demonstrate that your models do not overfit your data.
- The results of evaluation on the test split should be presented in a table or as a bar chart.
- You will need to perform **analysis** of your results. The main idea is to explain *why* you got what you got. Possible ways to analyze the obtained results are:
    - Error analysis
    - Ablation studies
    - Further experiments

All data processing, experiments, and writeup must be done in a Jupyter notebook `modeling.ipynb`.

Your **introduction** should provide motivation for the choice of the output variable and a short description of the input variables. You may want to remind the context from Project-1 as I may forget it by the time of grading the Project-2. After that use the remainder of your write up to showcase your training and evaluation details. Also pay attention to your

writing. Neatness, coherency, and clarity will count. Your write up must also include a short **conclusion** and discussion. This will require a summary of what you have learned about modeling.

The project is very open ended. You can add sections as you see fit. However, a standard structure is:

- Introduction,
- Methods,
- Experiments,
- Analysis,
- Discussion,
- Conclusion.

Deliverables

Your submission should be a ZIP-file `Lastname.zip`, that contains:

- `ML.ipynb`
- dataset file(s)