CS105 - Data Analysis Methods

UC Riverside

Project Phase 3 - Data Analysis Phase

In the final project for this course, you will apply the techniques learned in this class to analyze datasets of interest to you.

In this last phase, we first ask you to import all your code into the new repository for Phase 3. We had separated Phase 1 and Phase 2 into separate repositories for grading purposes, but it will be nice if all your code and analysis was in one repository so you can highlight this on your CV.

For this phase, you are asked to perform data analysis. This can include building a model to perform prediction (like applying linear regression or kNN) or clustering. You can also use the models for data analysis, not just 'predictions'. For example, in linear regression we saw that the resulting coefficients tell us how the features are correlated to the target variable. So, this analysis might help you identify features of importance with respect to a target feature in the dataset.

For analysis, identify two hypotheses (or questions) you are curious about regarding your data and then proceed to use ML techniques to try to answer these questions. If you decide to build models for predictions or classification, be sure to setup your problem so that you have a training, validation and testing dataset and you evaluate at least two ML / Data Mining techniques and compare their results.

To get started, accept the assignment link here `https://classroom.github.com/g/vGRQ7IMj`. Please import your code from Phase 1 and Phase 2. Assume this is the repository that you will showcase on your resume, so be sure the README has a description of the dataset and the problem.

This submission MUST include a README that :

- identifies and describes the dataset and problem at a high-level ;
- Phase 1 and Phase 2 contributions (can be part of the main readme or link to a different readme file);
- Summary of the data analysis process performed for Phase 3 and the results / observations obtained; Contain information about how to run your code (include any dependencies, etc.).

0.1 Submission via Github

Submit your jupyter notebook (and other code) to Github. Note, your notebook file should be named **teamname_proj_CS105**. You should include comments at the beginning of the notebook that includes your name and student id. Submissions will not be accepted via email; you must turn your assignment via Github.

0.2 Lab Presentation

In addition to submitting your code to Github, you must give a mini presentation about your project in lab or during lecture during Week 10.