Team Project 1
DS160-01
Introduction to Data Science
Spring 2021

Exploring Machine Learning Models (100 points)

Goal: *This project has two goals: 1) is for you to apply the exploratory analysis techniques you have learned this semester to prepare a dataset and 2) implement a machine learning (ML) model for which there is a package in Python or R. In other words, there is no need to try to write an algorithm from scratch.*

Instructions: This assignment will result in three deliverables:

1. Paper describing your analysis (template provided).
2. The well commented/documented notebook or R file containing your analysis code.
3. An 8 – 10 minute presentation of your analysis to the class (We will do these on final exam day)

Below are the specifications for this project.

1. With your partner, investigate several ML models and data sets to determine which model and which data set you want to work with. *I suggest you choose the data set first and then choose the model.* You may pick from the following ML models:
    a. Softmax Regression (Multiclass Logistic Regression)
    b. K-Nearest Neighbor
    c. Decision Tree
    d. Random Forest
    e. Naïve Bayes
    f. Support Vector Machine
    g. Agglomerative or Divisive Clustering
2. When you choose your data set, try to find one that has a lot of examples and a good mix of categorical versus continuous variables.
3. Perform a complete **Exploratory Data Analysis** (see template for further instructions).
4. Perform a **complete implementation of your model** including data preparation, experimental design (running the model with several different versions of the data), (see template for further instructions).
5. Document the output of the experiments with different versions of your model, (see template for further instructions).
6. Write a conclusion that summarizes your work, (see template for further instructions).
7. Prepare an 8 – 10 minute presentation of your work that you will deliver on final exam day. Both team members must present part of the work. Other than that, you have free rein to design the presentation as you want (e.g. PPT or not, live code, show PDF of paper, etc.)
8. Push all materials to a repository on GitHub. This includes your notebook/R files, your paper in Word format, your data set, presentation files (PPT, handouts, etc.) Ensure that you README.md provides a summary of your project. A good option would be to put your abstract in the README.md file, although you may do it a different way if you want.

This assignment must be completed in a teams of two. Please send me an email as soon as possible with your partner's name, so I know who is working together.

Project Submission: Upload a link to your GitHub repository for the project in the area provided in Moodle by the deadline specified.