Homework 3:

Advanced Data Analysis in Python

Due before class Tuesday, January 4, 2022

The purpose of this homework is to review and practice fundamental machine learning concepts. In The folder ML on GitHub, you will find a csv entitled "cses4_cut.csv" containing a subset of the CSES Wave Four data set. In the data, you will find demographic variables, the codebook for which is in "csesCodebook.txt." I have also appended the columns "age" and "voted." The idea is to build a predictive model of whether or not a respondent likely voted in their last presidential election (this is a cross-country survey). The values the variables can take (notice especially the missing values) are listed at `https://cses.org/wp-content/uploads/2019/03/cses4_Questionnaire.txt`. Some are ordered, and can be treated as an ordered scale. Others are categorical, and you will need to one-hot encode (see ML1.py).

You will need to perform all of the steps typical of a supervised machine learning analysis. You can use any subset of the variables included, or all of them. You should try multiple approaches, optimizing the model and its hyperparameters. You can use either a classifier (voted or not) or a regressor (such as a logit, treating the outcome as binary and the predicted fits as probabilities). You will need to assess the accuracy of your model in a way that is appropriate for the chosen model. This task includes feature engineering, variable selection, train-test split, optimization, and accuracy assessment. You may choose to utilize a dimensionality-reduction technique. How you approach the problem is up to you, but you need to explain the choices and steps made, any resulting tables and/or figures, etc., in a

brief informal write-up as a pdf (does not need to be general audience). You should include all of the code that you run in a py file. Comment the code for readability.

You may choose to use theoretically-guided variables or simply maximize predictive accuracy empirically. You do not need to, but may choose to, substantively explore and explain what variables are most influential in the model.