TEST 2

EDA, ML Classifiers and Data Cleaning

For this exam you will be responsible for performing exploratory data analysis, creating, and testing a variety of ML algorithms, and properly preparing the data for training. You will need to provide evidence you have done each part as well as explain your design decisions in a separate document. You should provide the best performing model as well as sufficient evidence that the other classifiers are sub-par.

You are free to do this assignment with any resources you can obtain. That means you can use the internet, your book, class code, and class notes. The only thing you may not use is other groups as a resource for code or algorithmic solutions. You must do your own work and own research. You may share sources (I found this site to be helpful, I found a solution using chapter 4 in our book, etc.). You may not copy and paste code; you should understand the code you write.

You can use IDLE, Jupiter Notebook or Anaconda to do this assignment. Just upload your file(s) to canvas when you are completed. You may provide images of EDA, visualizations of models and examples for how you cleaned your dataset. Your group should turn in evidence of each portion of this exam in the form of source code and documentation. To give you some hints, hyperparameterization of the models, exploration of API calls, different methods of data cleaning, different methods of feature scaling and a variety of metrics should all be analyzed.

For each model you should try to provide as many images of graphs as possible. Your code should produce accuracy for each model. There are other metrics you can explore as well, determine what metrics are available to you and how these impacts your decision about which model is best for this assignment. Your hint here is that accuracy may not be the most appropriate metric to make your sole decision by. Please tell me why. If there are any features that you believe are problematic or include some bias, please explain this as well.

Your group should turn in one exam with all three names, please include a separate document that explains how each person contributed to the exam. This is a chance for you to show me what you know!

As I explained in class, the group with the best overall model (not necessarily best accuracy) will be awarded an extra 2% on your final grade.