Final Project Proposal: Statistical Learning (UNL STAT 983)

by Sarah Aurit, Alison Kleffner, Emily Robinson

March 19, 2021

Needed in Project Proposal (1-3 Pages):

- Project Description
- How and where we obtained our data (Maybe a quick description of it as well? Link to Kaggle)
- Scientific Research Questions you may want to address and corresponding machine learning and statistical learning methods

Classification Methods: In this paper we are going to be looking at two main research questions:

- Developing a model that would work well to classify new wines into three categories of poor quality, normal quality and high quality.It is desirable to classify wines using physicochemical properties since this does not involve human bias that would come into play with human tasters. Methods we are considering to be Compared:
  - Random Forests: In prior research, this seems to work well on just the white wine data set, so would look at how this would work using red and white wine.
  - XGBoost: gradient boosting framework that can be done in R. Previous papers on this data set have used different versions of gradient boosting like adaptive boost
  - KNN
- As described above, an issue with our data is the imbalanced nature of the data set, with most of the wines being classified into the normal quality category. So we are interested in looking at different resampling techniques in order to determine. The different methods we are going to compare are:
  - No resampling
  - SMOTE - an algorithm where oversample the minority classes (so in this case low quality and high quality). Running the risk of overfitting our model
  - Random under sampler: randomly remove cases of the majority class from the data set. Running the risk of losing valuable infomation in our model.

Comment: Need to still clean this up and write into paragraphs but want to wait until earlier stuff writeen in order to have it to reference.