Madison Chamberlain

Neural Networks

To generate predictions for the African soil dataset, I used 1,157 rows, which can be denoted as n = 1,157 and 3,594 predictors which can be denoted as p = 3,594. So in this situation, p>n, which tends to lead to a longer runtime and overfitting. Since Neural Networks tend to overfit anyway, this dataset proves extra difficult in the game of not overfitting. Without specifying any parameters, the default parameters for the MLPRegressor() function generated a prediction which was accurate 72.61% of the time. Using GridSearch() and testing many parameters for optimal accuracy gave me 72.85% accuracy when the optimal parameters were used. This suggests that the neural networks model is ok, but not great for predicting the pH of the soil based on all 3,594 predictor features. This is likely due to the fact that using 3,594 features to predict a single number is unnecessary, and there are likely to be many features which simply create noise and decrease accuracy. With that being said, if we first used PCA to determine which features were really contributing to the pH, and then used an SVM or a polynomial fit with just those primary components, we may be able to generate a far more accurate prediction of pH for the African soil dataset. This is likely the case for most to all datasets which have p>n.