REPORT

Domain/Data: We will be using the MADELON dataset, which was part of the NIPS 2003 feature selection challenge.

Problem: We want to create a model to best predict the data for the aritificial and unlabeled Madelon dataset by developing a benchmark and testing multiple different models.

Solution: We will look at teh results of all of our models and determine which one preforms the best on our dataset.

Metrics: The metrics for our analysis will include the ridge penalty, the lasso penalty, SelectKBest, and the gridsearch.

Benchmark: Our benchmark will be the scores received from step 1 under the ridge penalty.

RESULTS:

Step 1 - Our step one results gave a training score of 0.786 and a test score of 0.544. This suggests that our model here may be overfitting the data.
Step 2 - The best training score under the lasso penalty was 0.610 and the best test score was 0.617. This is slightly better than what we found in our benchmark. The best scores were received with a C value of 0.01.
Step 3 - Using the gridsearch, our training and test scores were 0.617 and 0.610 respectively. These were the best scores of all of our models.

RECOMMENDATIONS:

I would recommend the use of the KNN model because it was the model that gave the best scores for both the testing and the testing data.