Domain

```
$ psql -h joshuacook.me -p 5432 -d dsi -U dsi_student
```

Data

This is the Madelon data set. It is a data set with 5 real features and 15 combinations of those features. The rest of the features are all features that have been randomly composed.

Problem

To find the real features using machine learning pipelines

Solution

Using Logistic Regression and KNN Classifier as our model.

Metric

Using the metric in the classifiers of Logistic Regression, Losgitic Regression l1, and KNN

Benchmark

Will be found using Logistic regression l2

Step 1

Using a high C value of 10000, the benchmark that was found was the train score is 0.787, but the test score is 0.56

Step 2

Using l1 in the Logistic Regression showed that the train 0.77 and the test was 0.57. That is a slightly better result than using the high C in the first Logistic Regression.

Step 3

Using KNN with the Grid Search was able to produce the the best result. It also looks like the best parameter for the KNN was 5 neighbors. The next thing to do is to be able to look at the coefficients that were best used in the grid search.