

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.

Benchmark

Will be found using Logistic regression l2

Step 1

Using a high C value of 0.0001, the benchmark that was found was the train score is 0.833, but the test score is 0.0299

Step 2

Using l1 in the Logistic Regression showed that the train was perfect, but the test showed almost no correlation.

Step 3

Using KNN and Logistic Regression was able to produce the same results as each other.

With these results, I would look into doing a different model to try and find a better solution to the problem.