PROJECT PROPOSAL

Learning Task:

The task is to carry out **letter recognition**. So given any capital letter, the learner should be able to identify what letter it is.

Dataset:

The data required for this task is gotten from the **UCI machine learning repository**. It consists of 20000 capital characters of the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted. It contains 16 different attributes asides from the actual label. This dataset can be found here: ***http://archive.ics.uci.edu/ml/datasets/Letter+Recognition***

Intended Algorithm

I intend to carry out this task using the **k-Nearest Neighbour** and the **Boosting** algorithm.

Implementation Language

This algorithm would be implemented in **Python**.

Planned experiments

About 16000 characters would be used as training set and the rest used as the test set.
During this learning task, I intend to show the effects of boosting on this algorithm. I would explore how boosting affects both the training error and test error.
I intend to explore how applying a boosting algorithm to the k-NN affects overfitting. I also intend to show how the value of k alters the learning process when k is either too large or too small.