**Homework 1:**

***Supervised Learning***

**Datasets and why are they important?**

**Nursery**

The first dataset I have chosen comes from 1980’s Ljubljana, Slovenia. During that period, there used to be an excessive number of applications to nursery schools but not enough school capacity to admit all of the applicants. Therefore, the government discretized some certain aspects of the applicants by assigning points and discrete values to them. In the end, there were a total of three main categories that the applications were evaluated: parents’ occupations, family structure and financial standing, and social and health picture of the family. In this dataset, these three subcategories are described with 8 different features (attributes) and these attributes have around 3-5 options each. The target value (label) is also categorized in five different categories, ranging from ‘*not recommended*’ to ‘*special priority*’. In this dataset, there are 12960 instances, meaning 12960 children who are classified. I separated my training and test data as 80% and 20% respectively of the total number of instances.

The reason I have chosen this dataset is because it comes from real life and the end result (labels) was something that had a great impact on thousands of families and their children. In reality, government formulized a way to classify the children and in this project, I will be trying to capture their classification methods and models. Throughout this paper, I will be referring this dataset as the ‘nursery dataset.’

**Letter Recognition (Letters)**

This dataset contains 20,000 instances of digitally typed and distorted in different ways letters from 20 different fonts of the English alphabet, meaning the result (label) has 26 different available options. Each letter appears between 734-813 times in the dataset, suggesting a somewhat balanced distribution. Each instance is defined by 16 different features including height, width of the char box, the number of pixels available and other features that describe edges of the letters. Each attribute is discredited from a continuous scale to an integer range between 1-15.

The reason I have chosen this dataset was to be able to compare and contrast with the first dataset. Since this dataset contains attributes that are derived from a continuous range, intuitively, the classification process seems to be different compared to the first one, a dataset with very discreet and polarized features. Another reason I have chosen this dataset was to be able to work with one of the most popular concepts in the industry: OCR (optical character recognition). I wanted to see how different machine learning algorithms would be able to capture the basics of this method and be successful when it comes to recognizing shapes and classifying them as the correct letters. Throughout this paper, I will be referring this dataset as the ‘letter dataset.’

**Project Overview**

This homework will have analyses of 5 different supervised learning algorithms. I will be exploring how different variables and kernels affect the results of these algorithms and how do they perform when ran on the two datasets I have. I separated each dataset as 80% training data, and 20% testing data to be able to test the trained classifiers on the test data.

**Decision Trees**

**Nursery**

**Letter Recognition**

**Neural Nets**

**Nursery**

**Letter Recognition**

**Boosting**

**Nursery**

**Letter Recognition**

**Support Vector Machine**

**Nursery**

**Letter Recognition**

**K-Nearest Neighbors**

**Nursery**

**Letter Recognition**