**PCSE 595 – Spring 2022 - Assignment 1 - Report**

< Your Name Here >

**Description:**

In this experiment, I compared a K-Nearest Neighbors classifier to a Decision Tree classifier on their performance on the Wisconsin Breast Cancer Dataset. The results indicate that the <insert best classifier> classifier performs the best, with a test set accuracy of <insert best accuracy>.

**Methodology:**

I implemented a k-nearest neighbors (KNN) classifier and a decision tree classifier. The k-nearest neighbors classifier <describe how it works in a few sentences>. I used

The decision tree classifier <describe how it works>. I used information gain as a splitting criterion, featurized the data by finding all boundaries between classes for each feature, and used maximum tree depth to avoid overfitting.

**Dataset Description:**

The Wisconsin Breast Cancer Dataset consists of <number of samples> samples with <number of features> features each. The features were extracted from digitized images of a fine needle aspirate of a breast cell mass. These features include:

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

Each sample is labeled as benign (non-cancerous) or malignant (cancerous). There are <number of benign labels> benign samples, and <number of malignant samples> malignant samples. The original dataset contained a few missing values. Missing values were replaced with a value of 5 which is the midpoint of possible feature values (features range in value from 1 to 10).

**Experimental Details:**

I withheld 20% of the data as a test set and used the remaining 80% for training. I used 5-fold cross-validation to tune classifier hyperparameters, and used average fold validation accuracy as my primary evaluation metric. The hyperparameters included <list hyperparameters for KNN> for the KNN, and <list hyperparameters for the decision tree> for the decision tree. Based on the results of cross validation, I found the best hyperparameters were <best hyperparameters for KNN> for the KNN classifier and <best hyperparameters for decision tree> for the decision tree. Using these hyperparameter values, I trained each classifier on the full training set to generate the results shown below. I compare the classifiers against each other, and a majority class classifier as a baseline. The majority class classifier simply tags all samples as <the majority class> which is the majority class.

**Results and conclusion:**

The KNN achieved a training accuracy of <training results of KNN> and a test accuracy of <test results of KNN>. The decision tree achieved a training accuracy of <training results of KNN> and a test accuracy of <test results of KNN>. Both classifiers outperformed the majority class baseline which achieved a test accuracy of <test results of majority class baseline>.

Based on these results, I can conclude that both classifiers were effective when compared against the majority class baseline, and that the <algorithm that performed better> is the better of the two classifiers for this dataset.

**Additional Questions:**

1. Which feature (in the Wisconsin Breast Cancer Dataset) has the highest variance?

<insert answer here>

1. Which two features (in the Wisconsin Breast Cancer Dataset) has the highest covariance? What does that mean?

<insert answer here>

1. Why is training accuracy higher (assuming it is) than test accuracy?

<insert answer here>

1. What was the average validation accuracy for each algorithm using optimum hyperparameters? Is it closer to the test accuracy than the training accuracy? Why?

<insert answers here>