**Assignment 2.1 Short Answer Template**

These questions are based on the Assignment 2.1 Exercises document.

**Section 1 (Ranking Problem):**

If our goal is to predict the outcome of the next match between a pair of players:

* Is predicting the outcome of a match between two players, given their scores, a regression or classification problem? Explain your answer in one sentence.
* Are the scores,ra, rb, and so on examples of hyperparameters, parameters, or labels? Explain your answer in one sentence.

**Section 2.2 (Text Data):**

* What accuracy do you get with the standard decision tree (no pruning, control of depth)?
* *To get the plot of accuracy, you can re-use your code from 2.1. Make sure to set the test size to 0.10. Plot the accuracy (train and test) as you vary the depth of the tree from 1 to 15.*

* What does this say about the accuracy of decision trees using text data?
* What accuracy do you get by varying the decision tree depth? Do you get higher accuracy with the entire vocabulary, or does choosing a smaller set of words increase accuracy? What does this suggest about text classification in terms of how much “information” a particular word might be providing the algorithm to make its decision with.

**Section 2.3 (Audio Data):**

You made plots of the “before” and “after” decision trees. Refer back to these plots for the following questions:

* Based on how much the tree changes in these two datasets, what can you say about the stability of decision trees?

* If two students develop decision tree classifiers using the audio dataset, are their trees likely to be similar or not?

**Section 2.4 (Decision Trees Summary):**

* Which dataset has the best accuracy, using decision trees? These aren’t exactly apples-to-apples comparisons, but using both your quantitative measures of accuracy and your subjective impression from manipulating the data, which of the three datasets gave the best performance?
* What types of data seem to be best suited for decision tree classification? In this question, be more general than just the three data types - explain what might make any new dataset good or bad for decision tree classification. Answer by explaining in terms of inductive bias.

**Section 3 (Cross-Validation):**

* You should have a plot with three lines, showing train accuracy, test accuracy, and random predictions. Does it seem that decision trees outperform random guessing when using the audio dataset?
* Plot accuracy vs. depth again, using the smaller amount of training data. How has the plot changed: how different are the training and test set error for the decision tree. How different are the accuracy predictions for random guessing using this smaller test set? Keep in mind that you are making fewer predictions, so random guessing might do well at some times.
* Why would you use one out cross-validation as opposed to a different method?

* When would you use a third validation set in addition to using a training and test set?