ECE356 Lab 4 Report
Group 10

Victor Yan
vlyan
20612514

Lun Jing
l7jing
20558988

Feature Selection

The process of selecting comes from two main aspects: the selection of the tables, and the selection of the attributes within these joined tables. We decided to join the Master table of players with the Batting, Pitching, and one attribute from the AllstarFull table as they seemed like they would contain more of an individual player's data.

The attributes with numerical data regarding a player were then all selected as features initially from the multi-joined table. We then had a Python script train a classification decision tree and checked to see the "importance" values of each feature. The script loops to train a new decision tree through each iteration and counts the number of instances where a feature has an importance of zero. After looping 5 times, any features that had zero importance 4 or 5 times would be removed. This process would be repeated again until no features had zero importance appearing 4 or 5 times. We also tried to remove features that appeared 3 times afterwards, but found that it was not effective.

The table below shows the data for each run-through when removing the features, starting with the first run containing all of the features. The accuracy and F1 scores were taken from running the lab 4 Python script for generating the decision tree and taking the first of the five iterations. G indicates the range values found for the Gini measure, and E for Entropy. The "AVG" suffix indicates the average values found for the respective impurity measures.

Table 1. Scores and Omitted Features of Progressive Runs

	First Run	**Second Run**	**Third Run**
Accuracy (First Iteration)	G: 91.7%-95.0% G_AVG: 93.86% E: 90.1%-94.6%	G: 91.7%-94.6% G_AVG: 93.12% E: 92.1%-95.8%	G: 91.3%-96.3% G_AVG: 93.34% E: 90.9%-95.0%

	E_AVG: 92.7%	E_AVG: 94.0%	E_AVG: 92.83%
F1 Score (First Iteration)	G: 38.1%-63.1% E: 42.8%-66.7%	G: 54.5%-68.3% E: 48.6%-64.3%	G: 45.7%-70.9% E: 47.6-70.0%
Omitted Features	None	Sum(B.RBI) Sum(P.G) Sum(P.IPouts) Sum(P.H) Sum(P.HR) Sum(P.IBB) Sum(P.WP) Sum(P.BK) Sum(P.R) Sum(P.GIDP)	SUM(B.R) Sum(A.gameNum) Sum(B.2B) Sum(B.GIDP) Sum(B.HR) Sum(P.HBP) Sum(P.SF) Sum(P.SH) Sum(P.SO)

From the data seen here, we decided to select the second run's set of features. This is because the first run appears to be overfitting, with too many features, and the third run, while having similar values for accuracy and F1, has a larger variance in the values, making it more unpredictable, thus potentially underfitting. The second run is a better balance between slightly higher accuracy and F1 scores with slightly less varying in values between runs.

Decision Trees

Figure 1 and figure 2 show the decision trees for Gini and Entropy respectively.

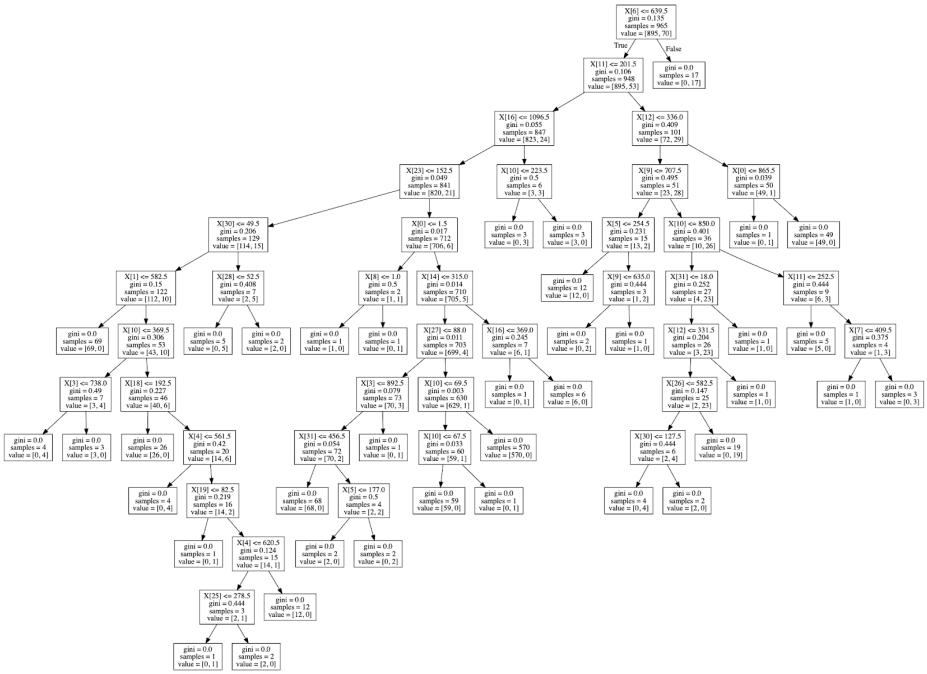

Figure 1. Gini Decision Tree

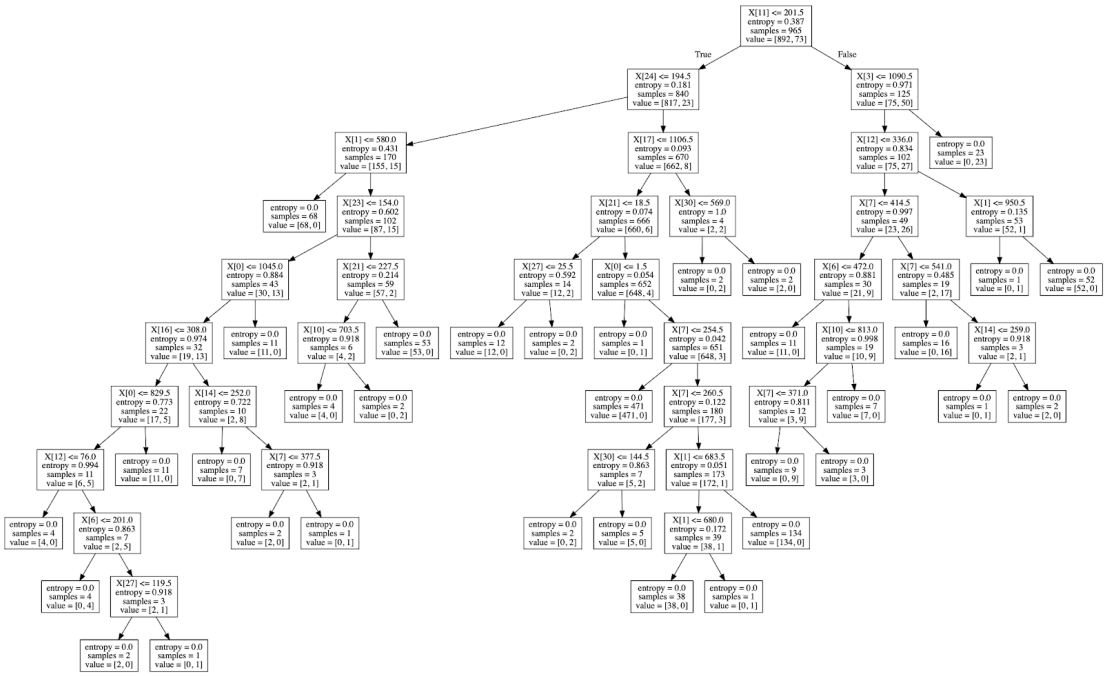

Figure 2. Entropy Decision Tree

Output Analysis and Comparison.

From the figure below, please note that the first iteration is the test result from the the 80/20 split. Comparatively, the accuracies

in the randomly pick 20% data test iterations are higher than those in the first trial. This is because, in the first trial, 80% data is used for training and creating the decision tree, while the left 20% test data are not included. In this case, the test data will have little chance to be identical to some entry in the train data set. Whereas in the other 5 random data select iterations, the data is picked from the entire data set, which has very high chances to include the data that appeared in the training data set. In this case, the prediction results will be guaranteed to be correct, which leads to a boost in the final accuracy.

Another thing to compare is the accuracy results between Gini and Entropy. From the test data in Table 1, despite the test data which appeared in train data set, the Entropy seems to have a slightly better accuracy than Gini (in the 80/20 split test, although this is not reflected in the data in Figure 3), but not too much difference. And in the rest five iterations, the accuracies of Gini and Entropy are both increasing and almost same.

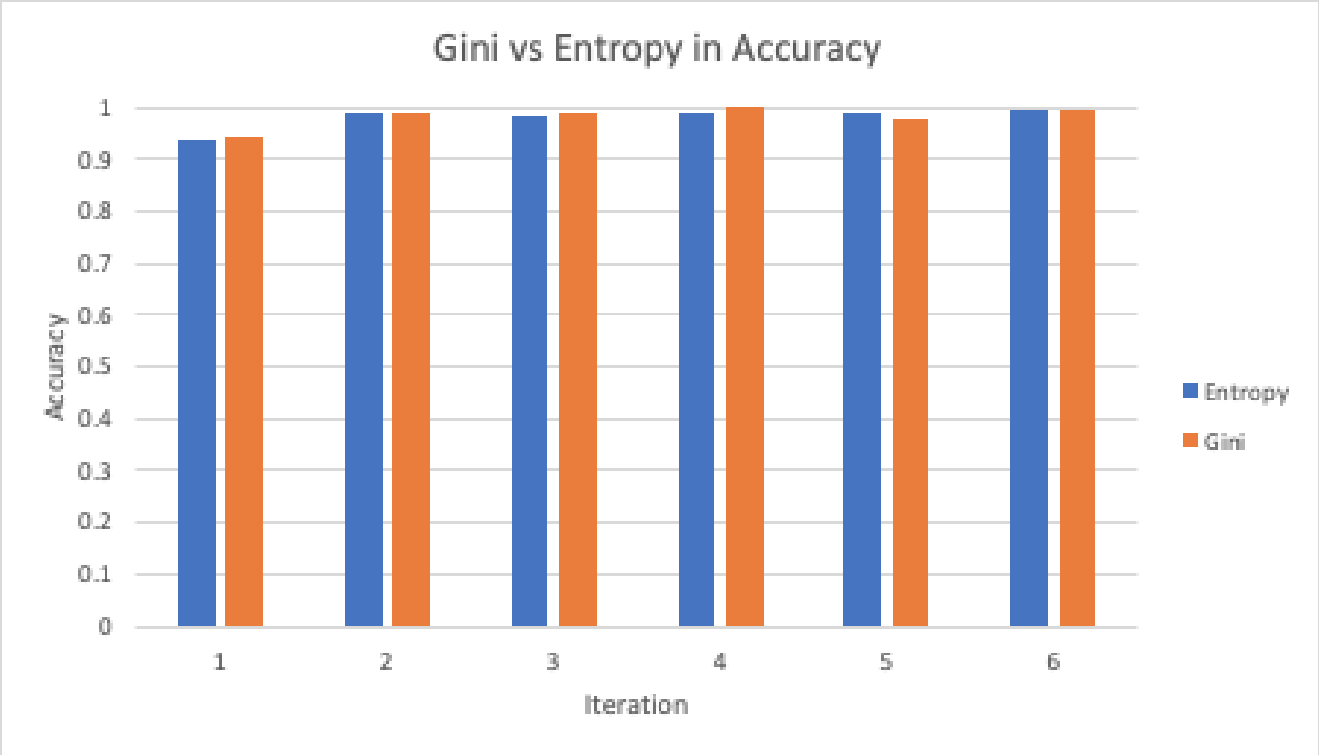

Figure 3. Gini vs Entropy in Accuracy