ECE 356 Lab 4 Report
Shashank Kotturi, Patrick Liu, Samiur Rahman

Analysis & results

(a) Compare the percentage of each class: nominated or elected.

In the HallOfFame table, there are individuals that can be categorized according to the following three cases:

1. An individual that has only ever been nominated, and not elected
2. An individual that has only ever been elected (elected to HOF during first nomination)
3. An individual that has been considered numerous times before being elected to the HOF

Individuals satisfying Case I are thus considered strictly as nominees. Any other case indicates that the individual was successfully elected to the HOF.

- Case I: **943 nominees**
- Case II & Case III: **317 elected**

Therefore, the percentages of each class are as follows:

$$Nominated = \frac{943}{943+317} = 0.7484126984 \approx 75\%$$
$$Elected = \frac{317}{943+317} = 0.2515873016 \approx 25\%$$

(b) Indicate your initial set of features and how you reduced them.

Initial Feature Selection

There is an attribute in the HallOfFame table, *category*, that is used to divide the set of all individuals in the table across four distinct domains: *Player*, *Manager*, *Umpire*, and *Pioneer/Executive*. This is the first feature that is supplied in our extracted data.

There is also an attribute in the HallOfFame table, votedBy, that is used to denote the committee responsible for screening nominees for election into the Hall of Fame. Prior to our feature selection, we investigated the prominence of these committees and how they influence the candidates' election into the Hall of Fame. We observed a series of trends in Table 1:

Table 1 - Voting Committee Influence on Electoral Results for Varying Candidate Categories

	votedBy	BBWAA	Veterans	Centennial	Old Timers	Special Election	Final Ballot	Nominating Vote	Run Off	Negro League
category	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛
Player	⬛	🟨	🟨	🟨	🟩	🟩	🟥	🟥	🟨	🟩
Manager	⬛	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥
Umpire	⬛	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Pioneer / Executive	⬛	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥

Where:

- Green denotes elected
- Yellow denotes that the respective committee has both elected as well as denied nominees to the Hall of Fame, and therefore additional screening (statistical analysis) is required
- Red denotes that these individuals have never been elected by the respective committee

We note that Managers, Umpires, and Pioneers/Executives are definitively elected, or definitely rejected from being inducted into the Hall of Fame across each committee. Based on the above findings, we considered including this attribute as a feature in the dataset, as we were unable to find additional statistics in the provided database that can be used to support electing Umpires and Pioneers/Executives to the Hall of Fame.

Players can be divided further according to their position. Although there is no need to go into detail with their positions, we can generalize them across two roles: pitchers and fielders. Pitchers will always have pitching statistics. Fielders will always have batting and fielding statistics. In consideration of their statistics, each of these players are eligible to earn awards over the course of their career with respect to their exceptional play in batting, pitching, and fielding.

Table 1.1 - Initial Feature Set (Part 1)

Table	HallOfFame	Batting	Pitching	Fielding	Managers
Attributes	category votedBy	AVG (H, AB) HR RBI H SLG (H, 2B, 3B, HR, AB)	WLPCT (W, L) ERA (ER, IPOuts) S IP (IPOuts) SO	P A	W L R

In addition to game-related data, we will also consider the quantity of All-star appearances made (from the AllstarFull table), as well as the quantity of awards received (from the AwardsPlayers and AwardsManagers tables), as they are both indicative of exceptional performance that can be tied into a player's and/or manager's Hall of Fame consideration.

Table 1.2 - Initial Feature Set (Part 2)

Table	AllstarFull	AwardsPlayers	AwardsManagers
Attributes	count	count	count

We initially considered including postseason statistics as features in the dataset, but decided against it. We agreed that they would be brought into stronger consideration in the event our classification accuracy was not close enough to the target.

Consequently, our initial feature set consists of 20 features, which altogether, are comprised of 22 raw features from the database. Although using this feature set was sufficient in achieving a mean classification accuracy greater than 85%, further consideration was made in order to optimize our selected feature set.

(c) Justify the choice for your selected features in terms of their relation with the player being either nominated or elected. These features are from Batting and Patching in addition to other tables.

Eliminating Raw Stat Comparison

If we treat each nominee as their own dataset, then we are essentially comparing datasets of varying sizes. In order to make the data comparable, we must scale their statistics accordingly. This is where we incorporate the raw data into advanced statistics.

Batting

$$Slugging\ Percentage\ (SLG)\ =\ \frac{1B+2(2B)+3(3B)+4(HR)}{AB}$$

A player's SLG represents the average number of bases a player records at each at bat. This takes into consideration the number of singles (1B), doubles (2B), triples (3B), home runs (HR), and at bats (AB) a player has recorded over their career. It can be interpreted as the overall productivity of a player as a batter. Consequently, we remove Home Runs, RBI, and Hits as standalone features to be compared. The SLG can be considered a weighted batting average, as it applies a weight factor to the type of hit that was recorded. Therefore, we chose to omit the batting average in favour of the SLG as the sole feature analyzed from the Batting table. The higher the SLG, the better, and vice versa.

Pitching

$$Earned\ Run\ Average\ (ERA) = \frac{9(ER)}{IP}$$

A pitcher's ERA represents the number of runs they have conceded (ER) on average per nine innings pitched (IP) - the length of a professional baseball game (not including extra innings). The lower the ERA, the better, and vice versa.

$$Walks\ plus\ Hits\ per\ Inning\ Pitched\ (WHIP) = \frac{BB+H}{IP}$$

A pitcher's WHIP represents the average number of baserunners allowed over the course of an inning (IP). Baserunners are considered batters that the pitcher has conceded a base on balls (BB), commonly referred to as a walk, and/or a hit (H) to. The lower the WHIP, the better, and vice versa.

We choose to omit the remaining statistics from the Pitching table in favour of using ERA and WHIP, as they are the more characteristic features of pitchers. Win-Loss Percentage is not exclusively influenced by the pitcher, saves are a feature typically exclusive to relief pitchers and its inclusion would discriminate against non-relief pitchers, and again, innings pitched and strikeouts are raw stats that are not necessarily indicative of a pitcher's effectiveness in their role. A strikeout cannot be weighted differently from a standard putout, as both contribute towards a single out when recorded.

Fielding

$$Fielding\ Percentage\ (FPCT) = \frac{PO+A}{PO+A+E}$$

A fielder's FPCT represents how reliably they properly handle a batted or thrown ball. It takes into account the player's number of putouts (PO) and assists (A), considered over their total number of chances - the quantity in the denominator - which also includes the number of errors (E) committed by the player. By incorporating the raw data into this advanced statistic, we choose FPCT as the sole feature from the Fielding table.

Managers

Additionally, we incorporate a Managers' Wins and Losses into a single statistic: Win-Loss Percentage (WLPCT). We choose to disregard rank, as it's relative to the number of teams in a division. Due to the variations in the number of teams in a division, it's better that this statistic is disregarded due to its inconsistent behaviour. Therefore, we choose WLPCT as the sole feature from the Managers table.

AllstarFull

There are implications with including a player's all-star game selections. The first professional baseball all-star game took place in 1933. With this database containing player data starting in 1871, using all-star selections as a metric discriminates against players with careers that were not active before 1933. As a result, it is omitted from the feature set.

AwardsPlayers and AwardsManagers

Likewise, the quantity of awards collected by a player and/or manager is also omitted for the same reason. The awards that are handed out today had been introduced at different points in time, meaning that players and/or managers whose careers were not active before their introduction were not eligible to compete for such awards. Since the quantity of the total number of awards handed out each season has been subject to change over the course of history, these statistics are omitted in favour of using consistent data.

Reduced Feature Set

We are now left with the following feature set, consisting of seven features overall, comprised of a total of sixteen raw stats:

Table 2 - Reduced Feature Set

Table	HallOfFame	Batting	Pitching	Fielding	Managers
Selected Attributes	category votedBy	SLG (H, 2B, 3B, HR, AB)	ERA (ER, IPOuts) WHIP (BB, H, IPOuts)	FPCT (PO, A, E)	WLPCT (W, L)

(d) Report the {classification accuracy, confusion matrix} of your decision tree classifier in addition to either {recall, precision} or F1 score.

Using 5 randomly selected 80/20 samples for each impurity measure, we had the following results:
Gini

- Classification accuracy: 91.11%

- Recall: 64.95%
- Precision: 85.23%

Confusion matrix:

Classification \ Prediction	Y	N
Y	202	109
N	35	1274

Entropy

- Classification accuracy: 88.95%
- Recall: 59.20%
- Precision: 80.75%

Confusion matrix:

Classification \ Prediction	Y	N
Y	193	133
N	46	1248

One observation is that the recall for both Gini and entropy are fairly low in comparison with the other metrics reported. This is because we are training our model on a relatively small dataset of < 2,000 data points, very few of which have a classification of ‘Y’. As a result, our model will tend to predict ‘N’ where the classification is ‘Y’, resulting in a higher number of false negatives and therefore a lower recall.

(e) If there is any overfitting phenomenon, mention how you addressed it.

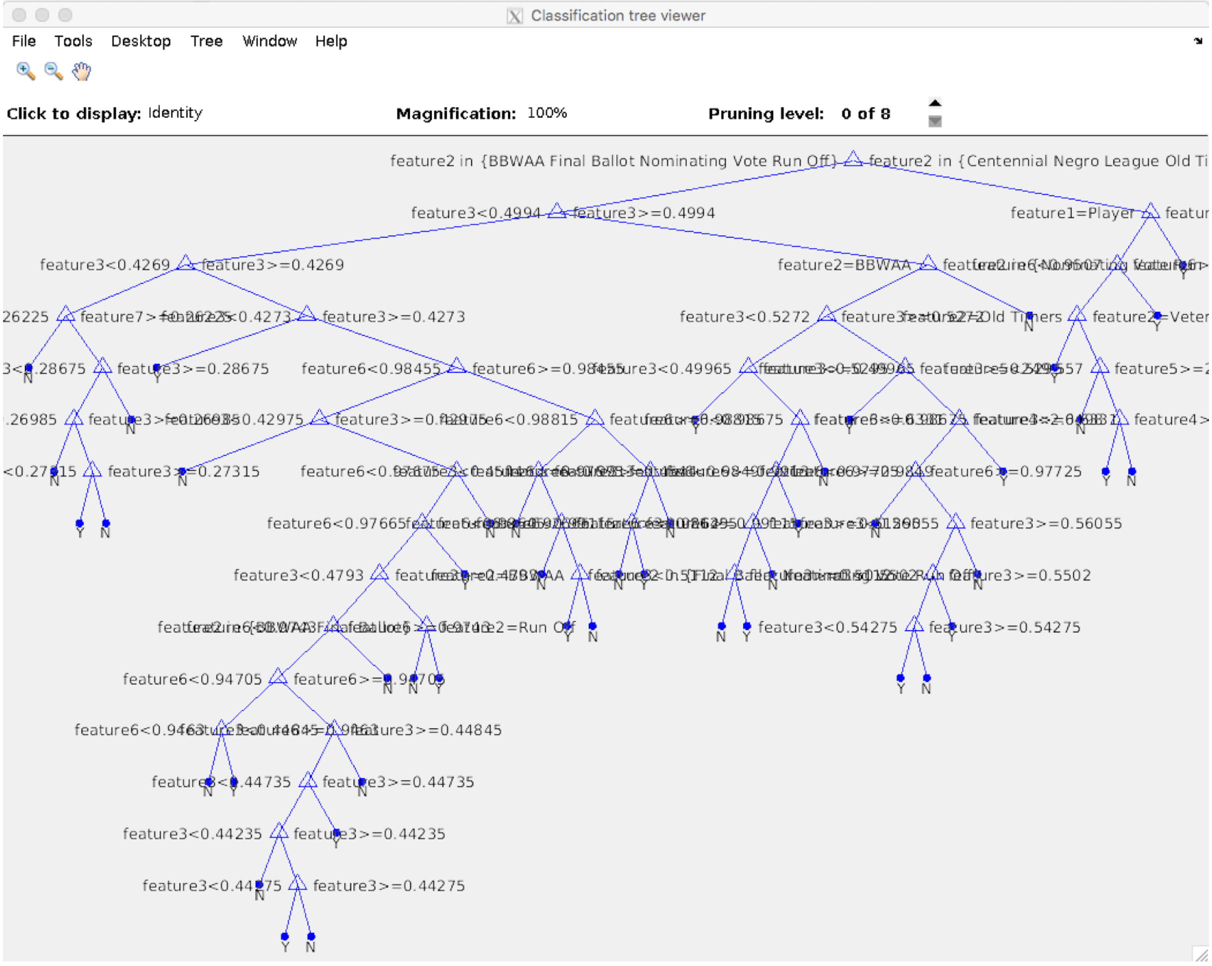
With our initial feature set, while the accuracy was around 85%, this could have been improved. The 22 features is excessive and resulted in an overly complex model.

We addressed this overfitting phenomenon by reducing the feature set through feature selection, which resulted in a higher accuracy. The features removed were extraneous and resulted in the model being overfitted, using relatively irrelevant features to determine branches within a decision tree.

Comparison

(f) A snapshot for the produced decision tree for each impurity measure should be provided. Compare the decision trees in terms of number of nodes and the breadth/depth of the tree versus the accuracy you achieved.

Gini



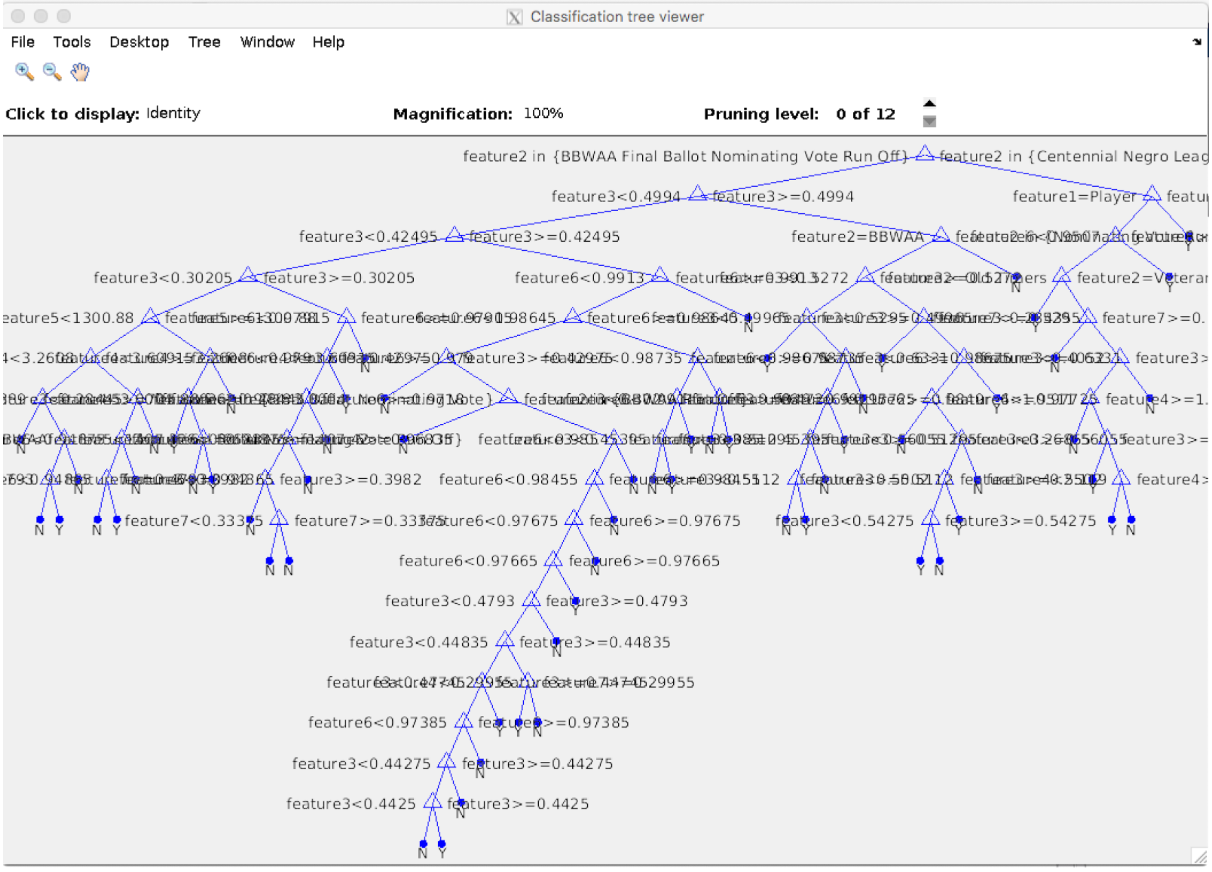
Click to display: Identity

Magnification: 100%

Pruning level: 0 of 8

Number of nodes = 91

Entropy



Number of nodes = 123

When comparing the decision tree generated using the impurity measure of entropy to that of Gini index, the tree generated using entropy had a greater breadth and depth. In addition, the decision tree generate using entropy also had a greater number of nodes. We observed that in these test runs, the tree with a greater number of nodes, breadth, and depth, has a slightly lower average accuracy. The entropy tree was very deep, which may be a symptom of overfitting, and therefore the lower accuracy measurement.

(g) Report & compare the accuracy plot/table for gini & entropy impurity measures for the 5 runs with commenting on the plot/table. Observations on the output relative to the selected features and the impurity measures are encouraged.

Gini

Dataset number	Accuracy
1	91.67%
2	90.12%
3	91.67%
4	90.43%
5	91.67%

Entropy

Dataset number	Accuracy
1	87.35%
2	90.12%
3	89.51%
4	89.20%
5	88.58%

When comparing the accuracy tables for Gini & entropy impurity measures, we observe an average accuracy of 91.112% and 88.952% respectively. Therefore, using this set of data points, the features we selected, and randomized 80/20 training/testing sets, we conclude that using the Gini impurity measure results in a higher-accuracy decision tree than with entropy.

Gini (initial feature set)

Dataset number	Accuracy
1	87.04%
2	87.04%
3	87.96%
4	86.11%
5	87.96%

Entropy (initial feature set)

Dataset number	Accuracy
1	85.80%
2	87.65%
3	87.04%
4	87.65%
5	83.02%

This pattern is consistent with the accuracy measures for both impurity measures using the initial larger set of features. However, the accuracies in those cases were both lower than those of the model after feature selection. Therefore, we conclude that the effect on accuracy of overfitting is greater than that of the impurity measure used to generate the decision tree.