How I got into the Hall of Fame

Unai Perez Mendizabal* , Damian Rubio Cuervo*
* Computer Science
Universitat Politcnica de Catalunya, C/ Jordi Girona 1, Barcelona

***Abstract*—The Hall of Fame is a recognition to the best players in the history of American baseball. But, what are the particular set of skills that make a player worthy of entering into the Hall? The Hall of Fame dataset collects the statistics of the careers of more than a thousand baseball players. Only a few made it into the Hall of Fame. By means of exploratory data analysis, we have aimed to find the motives and argumentations behind the Hall of Fame. Finally, we have found that, even though personal taste of the committee members plays a significant role, mostly offensively skilled and outstanding players tend to make it into the Hall.**

INTRODUCTION

This paper aims to study the *Hall of Fame* dataset. It comprises a list of different american baseball players, including a collection of statistics about each one of them, corresponding to their whole careers. The Hall of Fame refers to an official recognition that a few organizations may (or may not) give to certain players with outstanding careers. The *Veterans Committee* (VC for short) and the *BaseBall Writers Association of America* (BBWAA) are two of these organizations. The statistics included in the dataset are the following:

- **`Player`**: Character. Variable representing the name of the player. It will not be used during the analysis, but only to label the individuals.
- **`Number_seasons`**: Count data representing how namy seasons the player has played in the league.
- **`Games_played`**: Count data representing how many games the player has played.
- **`At_bats`**: Count data representing the number of times a player has been in the hitter position.
- **`Runs`**: Count data representing the number of runs a player has achieved.
- **`Hits`**: Count data representing the number of hits a player has achieved.
- **`Doubles`**: Count data representing the number of doubles a player has achieved. This happens when the player runs through two bases after one hit.
- **`Triples`**: Count data representing the number of triples a player has achieved. This happens when the player runs through three bases after one hit.
- **`Home_runs`**: Count data representing the number of home runs a player has achieved.
- **`RBIs`**: Count data representing the number of runs-batted-in a player has achieved. That is the number of runs that were completed after one of the player's hit.
- **`Walks`**: Count data representing the number of walks a player has made. This happens when the pitcher throws the ball poorly four times and the batter is allowed to just walk to the first base.
- **`Strikeouts`**: Count data representing the number of times the player has been eliminated due to three strikes.
- **`Batting_average`**: Countinuous numerical data representing the average number of hits a player made when in the hitter position.
- **`On_base_pct`**: Continuous data representing how frequently a batter reaches the base taking into account the times he has been in the batter position.
- **`Slugging_pct`**: Continuous data representing a measure of the batting productivity of a player. It is calculated with the formula below, where `1B` to `4B` refer to all possible outcomes of a hit (from single to home run) and `AB` refers to `At-Bats`.

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

- **`Fielding_ave`**: Continuous data representing the average times a defensive player properly handles a batted or thrown ball.
- **`Position`**: Factor. Variable that represents the position of the player in the field. The possible values are catcher, designated hitter, first base, outfield (That encloses the positions of Right field, Left Field and Center Field), second base, short stop and third base. These are all defensive positions, except for the designated hitter.
- **`Hall_of_Fame`**: Categorical variable. This is the target variable. Originally, it has three categories that represent players not included in the hall of fame with 0, players included in the hall of fame by the Baseball Writers' Association of America (BBWAA) with 1, and players included in the hall of fame by the Veterans Committee (VC) with 2.

The dataset has 1340 observations, from which, only 125 are players that made it into the Hall of Fame.

DATA PREPROCESSING

Values `1` and `2` of the class target `Hall_Of_Fame`, which is a factor, count as being in the Hall of Fame. The different values make a reference to the evaluation committee that accepted the players into the Hall. Depending on what procedure or analysis is run over the dataset, it might be useful to keep the different committees well labelled and differentiated. On

the other hand, there are very few Hall of Fame players, so dividing them into two smaller categories seems unprofitable; the samples would get very small. The two categorizations have been kept for potential usage.

As for the individuals, the name `ELMER_SMITH` is duplicated. A quick search in Wikipedia tells that there are two former players by the name of Elmer Smith. The data matches: One had played 10 seasons and the other 14, so it is not mislabelled data, and removing it would mean data loss. According to Wikipedia, the one that played 14 seasons is however more generally known as Mike Smith. So its name was replaced to `MIKE_SMITH`, which does not match any other player in the set. Also, the players' names are originally considered as one variable. They are going to be used as the row names and the variable is going to be removed.

Missing data and imputation

Missing data in the dataset is represented as question marks (`?`). This happens to 20 observations in the `Strikeouts` variable. Only 1 out of the 20 is a Hall of Fame player. Actually, the hetereogeneity of the dataset, regarding the proportion of Hall of Fame and non Hall of Fame players, is pretty similar in the whole population and in the sample with missing values (0.093 vs. 0.05). So, this sample has been considered relevant and has been imputed.

For this task, chained equations, PCA and kNN imputations have been tried. The chained equations method is not deterministic, but in most, if not all, of the attempts the imputed values decreased the mean of the variable significantly. PCA, on the contrary, increased the mean too much. kNN gave the results that most closely followed realistic values.

Outliers

Looking at the summary of the dataset, it seems like some variables have suspicious maximum and minimum values. For example, variable `RBIs` has a minimum value of 21. This value is far away from the mean of the variable and far below its first quartile. It can be seen in the boxplots of figure 2 that the range of the `RBIs` variable changes drastically

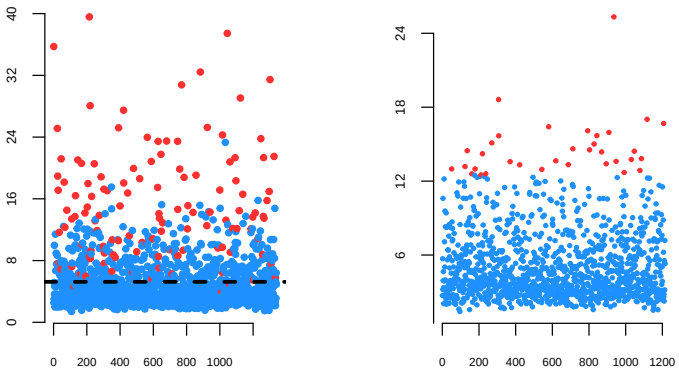

Fig. 1. Mahalanobis distances with and without Hall of Fame

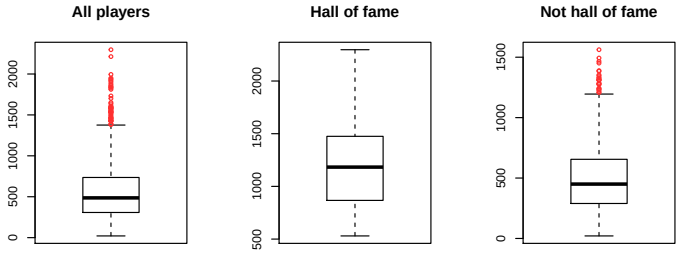

Fig. 2. Univariate outlier detection over RBIs variable

when including or omitting the hall of fame players. The minimum value of `RBIs` for the hall of fame players is much higher than the overall minimum. This means that the dataset's discordant maximum and minimum values may not really be representative of univariate outliers, becuase of the two really different populations that the dataset mixes. This will also be applied to the rest of continuous variables that show suspicious values, such as `Walks` or `Strikeouts`.

But, as for the averaging statistics, a different criteria should be applied. Specifically, the `Fielding_ave` variable has a maximum value of exactly `1`. Only one player, labelled as JOHN_ANDERSON, has the given value. This would mean that every time the player defended, the opposing team scored no runs. Still, the player did not make it into the Hall of Fame. However, the mean of the variable is pretty high (0.9663701). Without any more knowledge, this will not be considered an outlier.

As for multivariate outliers, the robustified Mahalanobis distance has been used for their detection. The plots of figure 1 represent the distances of the individuals. The plot on the left shows in red those players that made it into the Hall of Fame, and in blue the ones that did not. It is obvious that the plot considers most of the players in the hall of fame as outliers, and there is a reason for that: They were considered for the hall of fame for their outstanding stats and performances indeed. These players kind of behave like outliers in real life, but that is the point. So, they will not be treated as outliers.

The plot on the right of the figure 1 shows only players that are not in the hall of fame. It can be seen that some players still exceed other players. These could actually be considered outliers. If we consider to discard the 2.5% of the individuals, this is, the 34 with the greatest robustified Mahalanobis distance for being outliers, then those individuals plotted in red will be discarded. The player that most outlies from those excluded from the hall of fame actually has stats that are outstanding, so one can not help but wonder why he was not included. This player is Pete Rose and, as noted by `Wikipedia`, he has been involved in tax evasion and gambling controversy, which probably affected the commitees' decision making.

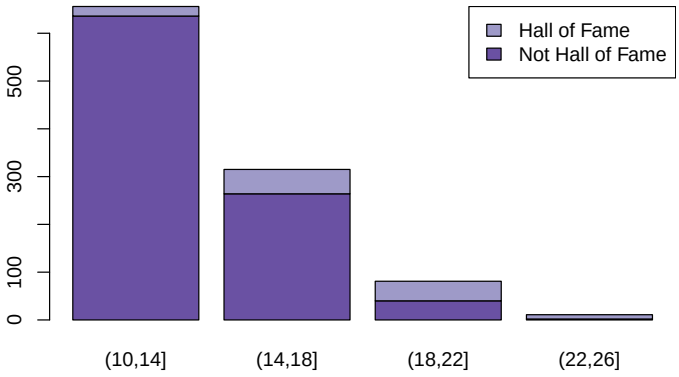

Fig. 3. Stacked barplot for played seasons

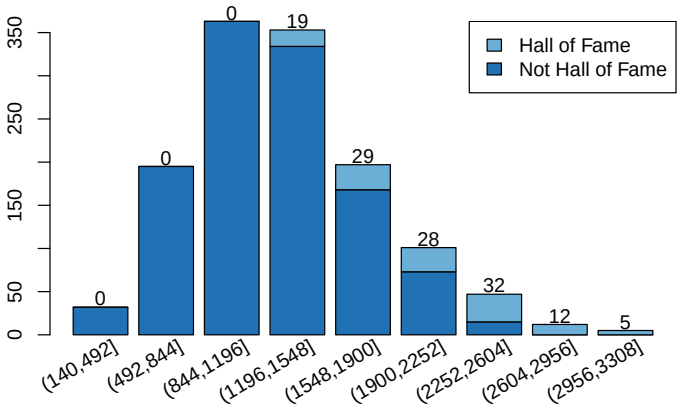

Fig. 4. Stacked barplot of played games

DATA EXPLORATION

Via visual data exploration, we might be able to obtain some useful information about the criteria of admission into the Hall of Fame. For example, one can reasonably think that the players that have the longest careers should be good enough or, at least, some kind of recognisable figure that is worth the acceptance into the Hall of Fame. Nevertheless, as barplot in figure 3 shows, there is no strong correlation between the season amount and the Hall of Fame. As careers grow longer, players that keep playing do end up getting accepted, but the amount of players accepted per each bar does not vary too much.

Variable `Games_played` should be strongly correlated with `Number_seasons`, meaning the variables are statistically dependent. Running a χ^2 test with the two variables gives a p-value of 0.018, which allows us to reject the null hypothesis H_0 that states that the variables are independent. And thus, the barplot on figure 4 does not show a but a slight increase in the number of Hall of Fame players. All variables are mostly related to `Number_seasons`, because the more seasons a player plays, the highest chance he has of scoring more in this statistics. Actually, the χ^2 p-value for the whole dataset is 0, so no independece can be taken for granted.

Oddly enough, baseball is a mostly defensive sport. The defensive part of it is when players need a more strategic and collaborative response. And, still, most of the variables are related to the offensive part of baseball, and, therefore, they are closely tied to the amount of games and seasons. That is the case for all but the `Fielding_ave` variable, which refers to the amount of plays in which the player has taken part into the putout of another player. Barplot in figure 5 shows the relation between the best recorded defenders and the players that got in the Hall of Fame. The better defenders they are, the highest is the chance to make it in. But also, the amount of discarded players increases as the number increases too. This leads to the conclusion that being a good defender is important to be a good player, and also that players that are just good defenders are not taken all that much into account.

A player needs to stand out in as much statistics as possible to make into the Hall of Fame. In the end, as the plot in figure 1 makes the most clear statement regarding what it really takes to be considered a great player: You need to be an outlier and, indeed, stand out in all the possible ways. Let's take a deeper look into all of this with some analysis.

CLUSTERING

In order to see what variables of the dataset characterize more the individuals we are going to perform some component analysis. This will also allow us to perform some clustering techniques with just the part of the information that is more relevant for our analysis. The analysis we are going to do is called PCA. When doing such an analysis we can see how the different variables are respresented and how much of the information is given by each of the analysis. In figure 6 we can see how the variables of the dataset correlate with the relevant dimensions. As we can see, the variables best represented in the first dimension are `RBIs` and `Games_played`, while only `Fielding_ave` seems to correlate with the second dimension.

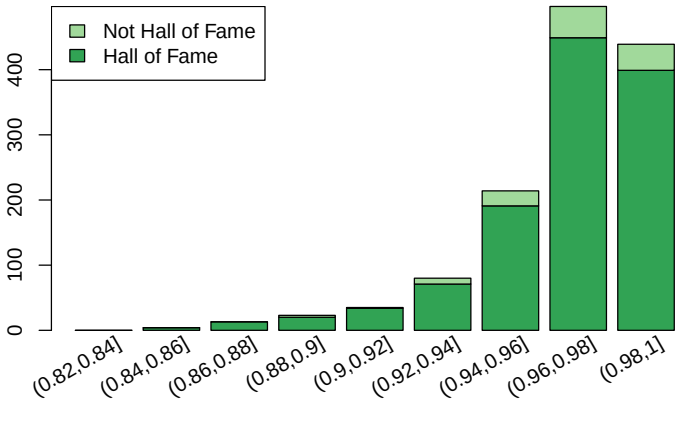

Fig. 5. Stacked barplot of fielding average

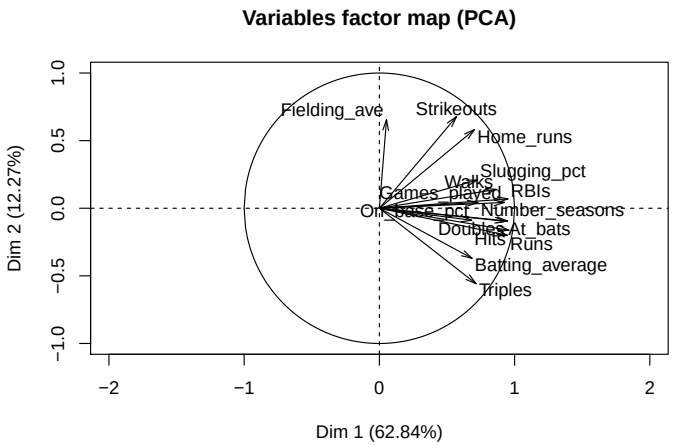

Fig. 6. PCA Analysis of the Hall Of Fame.

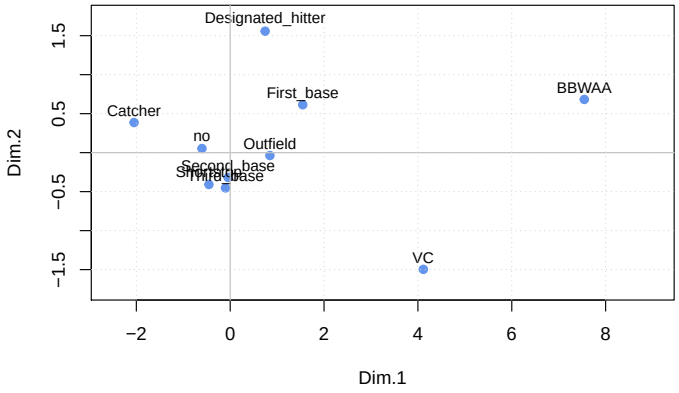

Fig. 7. Projection of the supplementary categorical variables

As we want to reduce the dimensionality of the data to be able to visualize and analyze it in a better way, we will have to decide how many dimensions are relvant to us. For that purpose we will use all that dimensions that will allow us to guarantee that we are preserving at least 80% of the information on the sample. In this case, that threshold is reached with just the three first dimensions. From now on, an in order to perform the clustering we will use the coordinates of the individuals in just those three dimensions.

The approach to clustering will be based on performing first a Hierarchical Clustering cutting it with the a priori knowledge we have, that is that there are only to clusters, one formed by players that entered the whole of fame and other with the ones that did not. In order to perform the clustering we will use the Ward distance metric. The results of this clustering are not trully representative of the classes that we know that exist in the data. If we cut the tree so that it gives us two clusters we will have 241 and 1065 individuals in each cluster. The hall of fame players are divided as 91 in the first cluster and 34 in the second cluster. This tells us that 72.8% of them have been classified as been in the first cluster, but there they only represent 37.76% of the total of individuals.

In order to improve these results we are going to perform a consoloditaion step. This step will consist of a k-means clustering procedure that will start from the centroids of the clusters found in the hierarchical clustering and then iterationg until the optimal local solution is found. With the purpose of assesing to this step a metric that tells us if it improves the results we are going to use the Calinski-Harabassz index. This index has been run over both clusters, the hierarchical clustering without consolidation and the consolidated one. The results can be seen in figure 8. The greater the value of the index, the better the clustering has been. Thus, we should consider the consolidated clustering with just two clusters, the best result for this problem.

Now, the clustering provides us with the following results. The clusters split the baseball players so that cluster one contains 410 and cluster two contains 896 individuals. The hall of fame players are divided as 121 in the first cluster and 4 in the second cluster. This tells us that 96.8% of the hall of fame players have been allocated into the first cluster, and that now they represent up to 29.51% of the total of individuals in that cluster. These results prove to be much better than before. The distribution of the indivudals in the clusters over the first two dimensions found in the principal component analysis can be seen in figure 9.

CLASSIFICATION

Once the data has been explored, we propose two methods for the classification. To reduce generalization error, the hold-out technique, along with cross validation and oversampling has been applied.

Decision Trees

In order to be able to create a classifier that will be able to predict whether a baseball player with certain records enter the hall of fame or not we are going to build a decission tree.

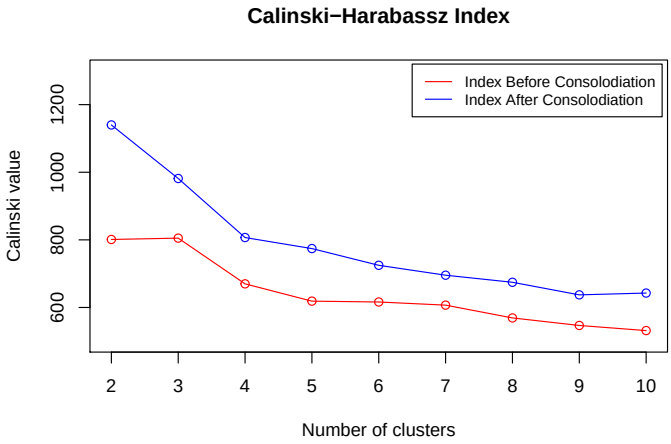

Fig. 8. Clustering Evaluation by CH Index.

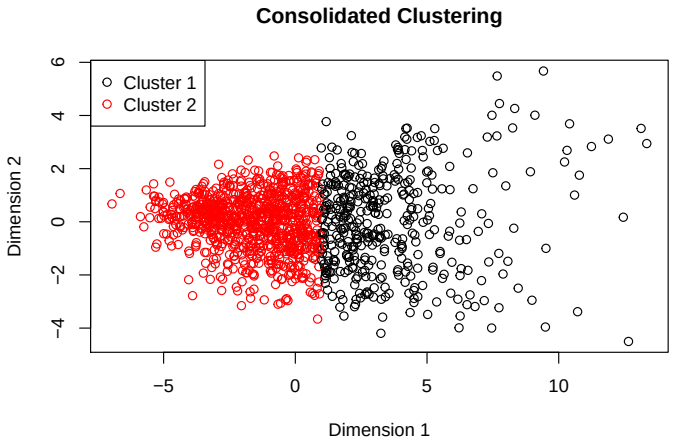

Fig. 9. Clustering of Hall Of Fame.

TABLE I
DECISION'S TREE CONFUSION MATRIX FOR HALL OF FAME.

	Predicted No	Predicted Yes
True No	362	31
True Yes	10	32

For that purpose we start by creating two subsets of the data that will be used to do holdout validation. As we have a really unbalanced dataset we've been oblied to perform oversampling regarding the players that enter the hall of fame, since the amount of them present in the sample is much lower than those that did not.

In first place we start by creating a pure decission tree. The shape of the tree can be seen in figure 11 and its performance over the test sample are shown in the table I.

With these values we can calculate some interesting statistics about pur classifier. The accuracy of the classifier, that is the proportion of instances that are properly classified is 90.57%. Even though this value is pretty high, it is not really representative for our example since we want to ensure that the maximum possible of players that enter the hall of fame are properly classified. Thus, we can measure the miss rate for that case, and we get it is 23.81%. It can be seen that a high volume of the interesting indivduals are been wrongly classified over the test sample. This can be due to overfittiong of the model.

To solve this issue we are going to try to prune the tree. The shape of the tree can be seen in figure 12 and its performance over the test sample are shown in table II.

With these new values we observe that now the accuracy of the classifier is 90.57% and the new value for the miss rate for the case of a player being in the hall of fame is 19.05%. Now, when comparing these values with the previous ones, we can see that the proportion of players that entered the hall of fame and were predicted to do so has increased.

TABLE II
PRUNED DECISION'S TREE CONFUSION MATRIX FOR HALL OF FAME.

	Predicted No	Predicted Yes
True No	360	33
True Yes	8	34

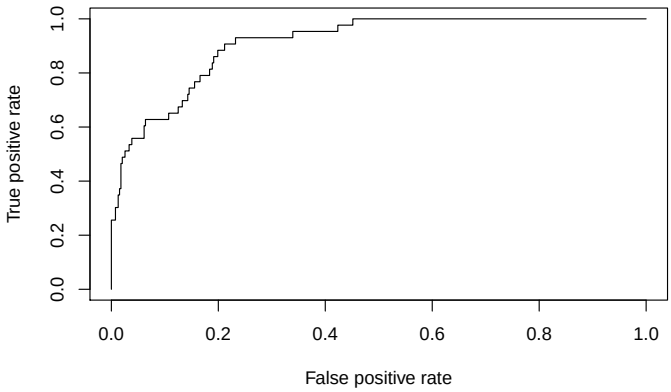

Fig. 10. ROC table of the best logistic regression model

Logistic regression

Being the target variable categorical, the most appropriate way to perform a regression is through logistic regression. The output of this regression follows a Bernoulli distribution and permits the transformation into the categories of the variable. To do so, the version of the dataset with binary target variable has been used.

The variables of the data are heterogeneous. Variable `Number_seasons` is an integer that ranges from 10 to 26, and `Hits`, while also being an integer, ranges from 48 to 4189. Also, four of the variables are percentages (with decimals) going from 0 to 1. The variables need to be standardized so that they can be used together without causing distortion in the results.

For the selection of the best possible model, the dataset has been split into a training and a test sets, each being two thirds and one third of the whole dataset, and a 10-fold cross validation process has been applied for the training. Due to the big disproportion between the Hall of Fame players and the no Hall of Fame players, it could be reasonable to think that the model could end up working with unbalanced samples and, therefore, underfit or biased. Actually, the random samples for the validation are pretty equally split. On the other hand, it could be easier to end up with unbalanced samples in the cross validation step. To avoid this, the Hall of Fame players have been oversampled. The best found model has an accuracy of 0.89, which is not great but is still really good. The ROC curve for the model is shown in figure 10. The Area Under the Curve (AUC) is 0.911.

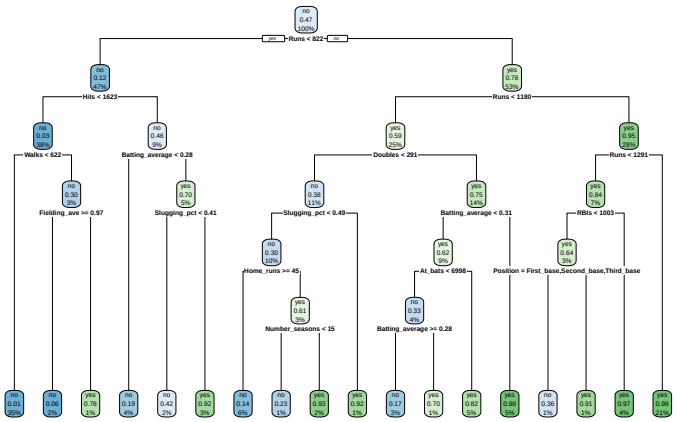

Fig. 11. Pruned decission tree for hall of fame.

CONCLUSION

To begin with, the Principal Component Analysis results in the variable projection shown in figure 6, which leads to two main conclusions. First, the most significant variables are related to offensive skills. This can be seen as the arrows that point the most to the right. Then, on the completely perpendicular direction, the only skill related to defense is found. As seen in the figure 7, the horizontal axis is tightly tied to being accepted in the Hall of Fame.

The same figure 7 also shows that no position is strictly correlated with getting accepted into the Hall of Fame. If anything, the designated hitter is related with the best fielding average, which does not make any sense, as they just do not defend.

To further see if we were able to find a correlation between the statistical data we possessed, and the players entering the Hall of Fame we have decided to perform clustering. This clustering has been performed in two steps were consolidation played a key role to obtain better results. As an output we have obtained a reasonable split of the baseball players in two groups as it was expected from a priori knowledge on the classes of the problem. This clusters allowed us to identify, with an acceptable degree of certainity the players that have entered the Hall of Fame.

In order to build a model that properly classifies the players as entering the hall of fame or not we have followed two different approaches. First of all, we have started building a decission tree. In the first attept of building such a model the results were kind of disappointing, but after the proper prunning we have succeeded on avoiding the overfitting of the tree to training dataset. Thus, the results of the prunned decission tree outperformed those of the original one.

The second classification model built was a logistic regression one, a variation of linear regression aimed at predicting binary target variables. As stated in the corresponding section, after performing 10-fold cross validation the obtained model, although improvable, is pretty good and is able to achieve an accuracy of close to 90%.

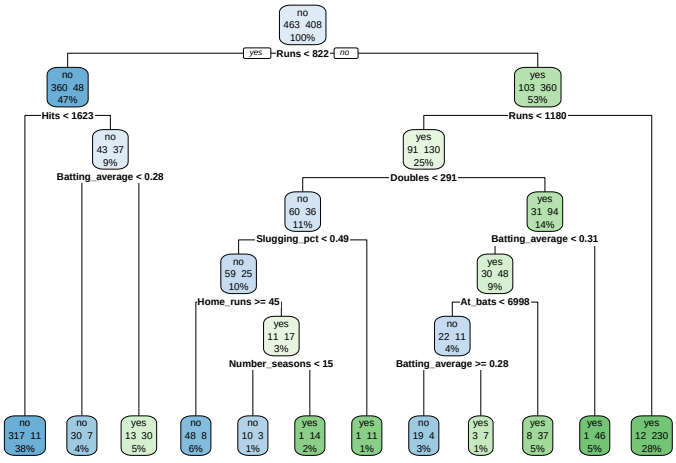

Fig. 12. Decission tree for hall of fame.