# Prediction Model for Baseball Hall of Fame Players

Diego Maldonado Castro

## 1 INTRODUCTION

In professional baseball, the Hall of Fame represents the pinnacle of individual achievement and recognition. Induction into this prestigious institution is reserved for the most exceptional players, managers, umpires, and executives who have significantly impacted the sport. The process of selecting Hall of Fame inductees traditionally relies on the subjective judgment of voters, which, while thorough, can sometimes be influenced by biases and varying criteria for evaluation.

With the advent of machine learning and advanced statistical analysis, there is an opportunity to bring a more objective and data-driven approach to predicting Hall of Fame inductees. This project aims to leverage machine learning models and ensemble techniques to predict the likelihood of a baseball player being inducted into the Hall of Fame based on their career statistics.

## 2 DATA DESCRIPTION

The dataframe used for model training takes into account all those who were candidates for the Hall of Fame (players) and their statistics which represent their overall performance as a pitcher and/or batter. Each field represents the average of a game statistic per player, which will give the average performance of each player represented in a season.

### 2.1 Fields

- AB (At Bat): Number of official at-bats a batter takes against a pitcher in the season.
- R (Runs): Runs scored by the batter.
- H (Hits): Hits by a batter.
- 2B (Double Play): Act of a batter hitting the pitched ball and safely reaching second base without being declared out by the umpire.
- 3B (Triple Play): Act of a batter hitting the pitched ball and safely reaching third base without being declared out by the umpire.
- HR (Home Runs): Play in which the batter hits the ball in such a way that allows him to make a complete circuit of the bases and score a run.
- RBI (Runs Batted In): Number of runs batted in by the batter.
- SB (Stolen Bases): Stolen bases.
- CS (Caught Stealing): When a runner attempts to steal a base but is thrown out before reaching the second base, third base, or home.
- BB (Base on Balls): When a batter receives four pitches during a plate appearance that the umpire calls balls, and is awarded first base without the possibility of being declared out.
- SO (Strike Out): When a batter accumulates three strikes during their at-bat.
- IBB (Intentional walks): Base awarded to a batter by a pitcher with the intent to remove the batter's opportunity to swing at the pitched ball.
- SH (Sacrifice Hits): When a batter successfully advances one or more runners by bunting the ball for an out, or would have been out if not for an error or fielder's choice.
- SF (Sacrifice Flies): When a batter hits a fly ball to the outfield or foul territory that allows a runner to score.
- W (Wins): Games won.
- L (Losses): Games lost.
- SHO (Shutouts): Act by which a single pitcher pitches a complete game and does not allow the opposing team to score a run.
- SV (Saves): When a relief pitcher finishes a game for the winning team under certain circumstances.
- IPouts (Outs Pitched): Innings pitched measured by the number of outs the pitcher records (each inning consists of 3 outs).
- H (Hits): Hits against the pitcher.
- ER (Earned Runs): Any run that scores against a pitcher without the benefit of an error or passed ball.
- HR (Home Runs): Home runs against the pitcher.
- BB (Walks): When a pitcher throws four pitches outside the strike zone, and the batter is awarded first base.
- SO (Strike Out): The number of times the pitcher causes a batter to accumulate three strikes during their at-bat.
- BAOpp (Opponent's Batting Average): The batting average of all batters who have faced the pitcher.
- ERA (Earned Run Average): The average number of earned runs allowed by a pitcher per nine innings pitched.
- WP (Wild Pitches): An errant pitch by the pitcher allowing a base runner to advance.
- HBP (Hit by Pitch): The number of times a batter was hit by a pitch.
- BK (Balks): An illegal move by the pitcher during their delivery.
- BFP (Batters Faced by Pitcher): The number of batters who faced the pitcher.
- R (Runs Allowed): Runs allowed by the pitcher.

*2.1.1 Handling null values.* Since the null values are the result of not playing in a position and not a problem of the recollection of data, the most accurate way of handling the null values for purposes of training the model would be filling them with zeros. Once the null values where filled we can proceed.

## 3 METHODOLOGY

### 3.1 Preprocessing

In order to make easier the training of models, the preprocessing tool of MinMaxScale was applied to each column of the dataframe. This gives a standardized dataframe 'X' with each column maintaining its scale.

For the inductedx column that will represent or vector y of binary classification, the label encoder tool will be applied, allowing the

model to properly process the y vector of binary classification. The result being a vector of ones and zeros.

### 3.2 Model selection and training

There will be 5 models from the sklearn library in charge of the predictions. These models will be:

(1) SVC(Support Vector Classifier)
(2) KNeighborsClassifier
(3) GaussianNB
(4) Logistic Regression
(5) DecisionTreeClassifier

Before training the models we use the 'train-test-split' method to divide the data that will be used for training and testing. With all the models now defined, we train them with the Xtrain dataframe and ytrain vector.

### 3.3 Ensemble

After the training of the models they will be integrated into an ensemble. The ensemble being a Voting Classifier.

Apart from that the ensemble is going to be compared to other ensembles to see if the data set can be better predicted with other methods that aren't individual models. The ensembles being:

(1) Voting Classifier
(2) Random Forest Classifier
(3) Gradient Boost Classifier

## 4 MODEL EVALUATION

For the evaluation of the overall performance of our models, the metrics that will evaluate the them will be:

- Accuracy
- Precision
- Recall
- F1

Taking into account that only around 20% of players nominated to the hall of fame are actually admitted into the hall of fame, we have to take into to consideration that there are significant more candidates than hall of fame players. This is why the metrics used for evaluation will be weighted or balanced to take into consideration the majority of one class against the other.

Support Vector Classifier

- Balanced Accuracy: 0.716106014271152
- Weighted F1: 0.8424867339803767
- Weighted Precision: 0.8689451523958567
- Weighted Recall: 0.8615023474178404

K-Nearest Neighbors

- Balanced Accuracy: 0.7058659994439811
- Weighted F1: 0.8247700921172899
- Weighted Precision: 0.8319108759609799
- Weighted Recall: 0.8403755868544601

Logistic Regression

- Balanced Accuracy: 0.6302474284125661
- Weighted F1: 0.7833380851913099
- Weighted Precision: 0.8300956412217693
- Weighted Recall: 0.8215962441314554

Gaussian Naive Bayes

- Balanced Accuracy: 0.6912241682883884
- Weighted F1: 0.7761857052444141
- Weighted Precision: 0.7778687514790871
- Weighted Recall: 0.7746478873239436

Decision Tree

- Balanced Accuracy: 0.7059123343527014
- Weighted F1: 0.7874869179771293
- Weighted Precision: 0.7886735810651855
- Weighted Recall: 0.7863849765258216

### 4.1 Randomized tests

In order to give a better idea of how the model would perform, randomized samples of size 30 from the test sets were given to the models. The results of the performance of the models are presented in the next graphic:

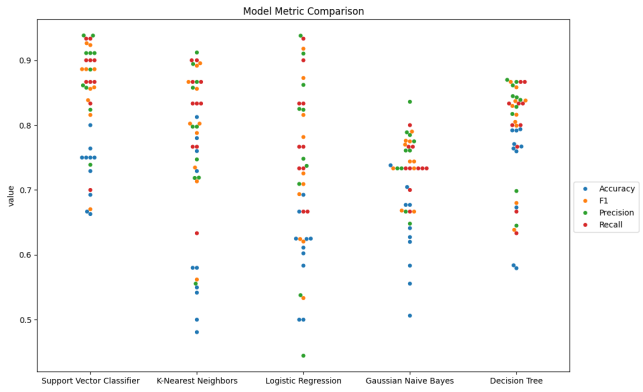


**Figure 1: Swarmplot of model metrics**

### 4.2 Model results

Taking into account the randomized tests and overall performance of the models with the test sets, the SVC(Support Vector Classifier) is the clear winner.

## 5 ENSEMBLE

In order to see if better test results can be achieved, we're going to test three distinct ensembles and see if the results improve compared to the individual models.

The three ensembles are:

(1) Voting Classifier
(2) Random Forest Classifier
(3) Gradient Boost Classifier

The results being:

Voting Classifier

- Balanced Accuracy: 0.7710592160133445
- Weighted F1: 0.8668547145454132
- Weighted Precision: 0.8723525014695469
- Weighted Recall: 0.8755868544600939

Random Forest Classifier

- Balanced Accuracy: 0.7897321842275971
- Weighted F1: 0.876001907477698
- Weighted Precision: 0.8791664184337661
- Weighted Recall: 0.8826291079812206

Gradient Boost Classifier

- Balanced Accuracy: 0.7594291539245668
- Weighted F1: 0.8589402161679197
- Weighted Precision: 0.8642973293132595
- Weighted Recall: 0.8685446009389671

## 5.1 Randomized tests

As previously done with the individual models, in order to give a better idea of how the ensembles performs, randomized samples of size 30 from the test sets were given to the models, adding the best single model for comparison reason. The results of the performance of the models are presented in the next graphic:

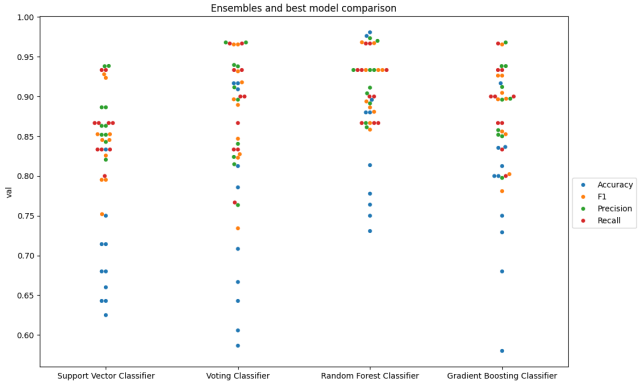


**Figure 2: Swarmplot of different ensembles and best individual models metrics**

As seen by the graphic and the general metrics there's a small improvement compared to the best individual model. Although, the improvement isn't that significant. Even in the case of most improvement which was the 'balanced accuracy', the metric only went up around 7%.

The ensemble that provided the best results was Random Forest Classifier.

# 6 CONCLUSIONS

The performance for the individual models proved to not be that reliable or satisfactory. As for the ensembles, each one presented better results in comparison to the individual models. Even then the best ensemble can't be seen as "reliable".

This could be attributed to various factors. Mainly that there isn't a defined bar or standard that makes a hall of fame player. A Hall of Fame player is determined by a jury, which is in constant change meaning the determining factor isn't always the same.

Also the data, that is the average career of each player doesn't give us the whole picture of how this player performed. The average while a good indicator of a career, it doesn't give full insight that could be critical at the hour of determining a hall of fame player.

Overall, the combination of a non constant criteria and the type of data that doesn't give full information about a player, could be the reason why the models and ensemble performances, while somewhat high weren't that reliable. However, even with some inconveniences the some models and ensembles can give a good general idea of which player could be hall of fame material.

# 7 REFERENCES

Lahman, S. (2023). Lahman's Baseball Database. Retrieved from http://seanlahman.com/