METCS677

Final Project

Tommy Lin

tomlin1

**NBA Hall of Fame Classifier**

**Introduction**

For my final project, I decided to train a model that can accurately predict whether an NBA player will be inducted into the NBA Hall of Fame.

The dataset used in this project can be downloaded from this [Kaggle link](https://www.kaggle.com/datasets/ryanschubertds/all-nba-aba-players-bio-stats-accolades). The dataset contains 4980 players and 39 features in total, but I will be narrowing it down to just 10 features.

The models I have decided to use are the K-Nearest Neighbors Classifier, the Logistic Regression Classifier, and the Random Forest Classifier. At the end of the project, I will use the best classifier to predict which current NBA players have a shot at making the Hall of Fame.

**Preprocessing**

The first thing I did was create a class column for the dataset. While there is no column indicating whether a player is in the Hall of Fame or not, I leveraged the fact that players who had an asterisk ‘\*’ at the end of their names were in the Hall of Fame to create the true labels.

Next, I separated players who are retired for 4+ years from the players who are not. This is because only the former group can be inducted into the Hall of Fame. My training and testing will be done on the former group of players.

Then, I used the RFECV selector from sklearn to select the 10 best features from the 39 total. I ended up including the following features: All Star, G, FG3%, Championships, TRB, PTS, FG%, FT%, WS, and PER. Most of these attributes made sense to me like the number of All-Star appearances, a player’s win shares, and the number of championships. Other features like total rebounds and free throw percentage were a bit surprising to me. Regardless, I used these 10 features going forward.

After splitting the preprocessed data into a subsets of 80% training data and 20% testing data, I was ready to implement the models.

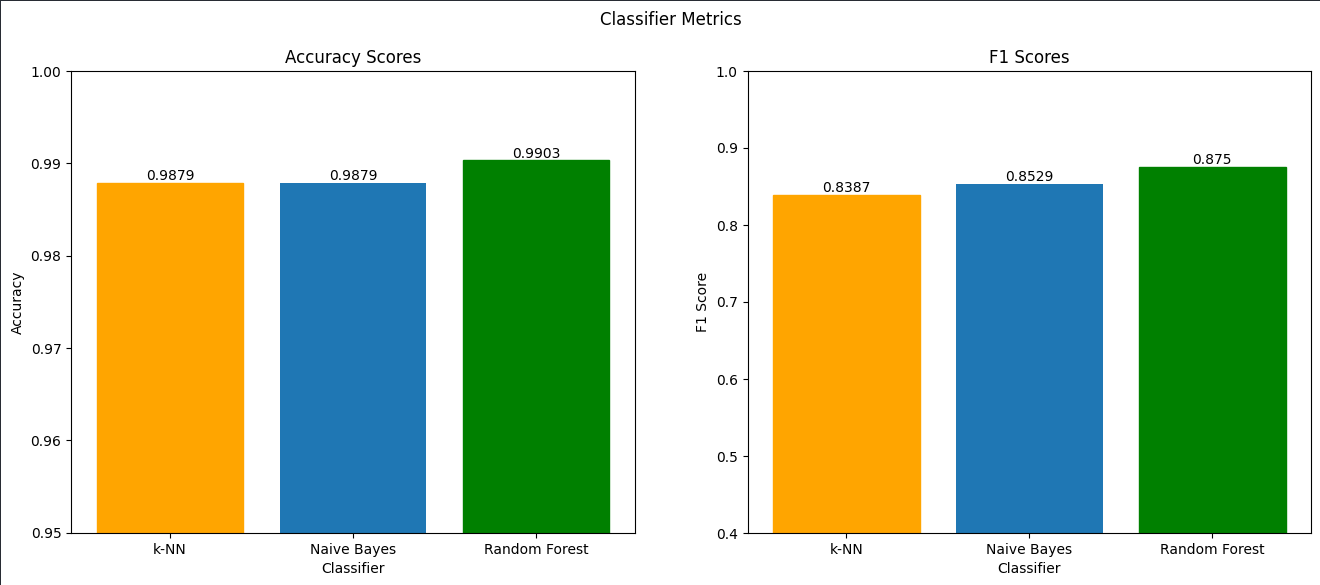
**Models**

I first implemented the k-NN classifiers on different odd k values ranging from [1, 29]. I found that the model produced the best results when k is equal to 13. For the Random Forest Classifier, I found that the best parameters were max\_depth=7, n\_estimators=25, and class\_weight=None. I used f1 score instead of accuracy to determine how good a model is because it is better than accuracy for imbalanced datasets, which my dataset is. There was no need to tune hyperparameters for the Logistic Regression Classifier.

| **Model** | **Accuracy** | **F1 Score** |
| --- | --- | --- |
| k-NN | 0.988 | 0.839 |
| Logistic Regression | 0.988 | 0.853 |
| Random Forest | 0.990 | 0.875 |

**Results**

After implementing all three models, I found that the Random Forest Classifier performed the best. It had both the highest accuracy at 0.990, and the highest f1 score at 0.875. However, the other models displayed comparable results.



Using the fitted Random Forest Classifier, I made predictions on the subset of players who have not been retired for 4 years yet. This dataset includes current players and recently retired players.

Looking at the list of players that are predicted to enter the Hall of Fame in the future, I saw a lot of familiar names. Players like Stephen Curry, Lebron James, Kevin Durant, and Giannis Antetokuonmpo are without a doubt future Hall of Famers. Other players like Kyle Lowry, Paul Millsap, and LaMarcus Aldridge were classified as Hall of Famers but aren’t as obvious to me as the previously mentioned players. All in all, the results were interesting.