2-1. The two attributes that I chose that are likely predictors on whether a player makes it to the hall of fame is points and field goals made. The pattern I found is that the data is skewed left for both the attributes. This shows that there are a few players who can score a lot of points in the league.

2.2. Using classification accuracy rate as the evaluation measure, the random forest model best predicts whether a player will be inducted. Regardless of the hyperparameters, numIterations and numFeatures, the classification accuracy rate remains 97.58%.

2.3. Using the Area under the ROC curve (ROC area), The random forest model using 100 iterations and 3 features yields the highest ROC area under the curve. It had an area of around .970.

2.4. When seeking to identify which players are most likely to be inducted into the Hall of Fame, it is recommended to assess our predictive model's accuracy using the area under the Receiver Operating Characteristic curve (AUC-ROC). The AUC-ROC is advantageous because it evaluates the balance between the true positive rate and the false positive rate. This method is more informative than simply counting correct predictions, as it sheds light on the frequency and impact of both false positives and negatives. Given the rarity of Hall of Fame inductions, the AUC-ROC is particularly appropriate for our nuanced analytical needs.

3.1 .The predictors I would use is points, total seasons, and assists. The measure I use is the overall standings and records that the NBA uses to keep all the historical data. The majority of the people who are on the list are either in the Hall-of-Fame or will be in the Hall-of-Fame because they are still a current player.

3.2. The caveat with using cumulative values is that it needs to be supplemented with a per game basis rather than a cumulative summary. This is because there are young players who have historic runs in the first 5 years of their career and that needs to be accounted for in the model. By having a cumulative summary, the data can be skewed if there was an injury or suspension in the career of the player. Furthermore, the game of basketball as evolved, the introduction of the three point line, pace-of-play, and playing styles, it is more beneficial to use stats that tracks on a per game basis. By focusing on averaging statistics, we can accurately assess players who might have shorter careers but had an exceptional stint, providing a clearer picture on who is worthy for the Hall-of-Fame

3-3. We created new variables by calculating the points per game and assists per game. After incorporating these variables in our WEKA dataset and by removing the position, league, and retired2004, it made the model worse with a classification accuracy of 97.17% and an AUC of .969.

4-1. I would use a profit chart to determine whether or not I should select the players to invest in.

4-2. You would use the AUC as the evaluation measure. This is because the AUC will give a global view if the model classifies the data correctly

4-3. a. the model I am using is the Random Forest with 3 number of features and 100 iterations.

b. the probability threshold will be 0.36

4.5 The profit before the competition is -$ if we were to test on the training data that references the test set.

inst#	actual	predicted	prediction	probability
4	0	1	0.64	0.64
25	0	0	0.52	0.48
102	0	1	0.51	0.51
108	0	1	0.61	0.61
150	0	1	0.61	0.61
151	1	1	0.94	0.94
152	1	1	0.97	0.97
153	1	1	0.78	0.78
154	1	1	0.97	0.97
156	1	0	0.53	0.47
157	1	1	0.97	0.97
158	1	1	0.69	0.69
159	1	1	0.92	0.92
160	1	1	0.74	0.74
161	1	1	0.96	0.96
162	1	1	0.97	0.97
163	1	1	0.93	0.93
164	1	1	0.93	0.93
165	1	0	0.57	0.43

Here are the players who we should invest in. It needs to be noted that this is tested on the training set that references the test set. The profit I would get is 1,640,000