

**Evaluating Predictors for the 2022 World Cup Using Decision Trees and
Random Forests**

Nicole Morgen

Department of Mathematics, Carroll College

MA 499: Senior Thesis

Advisor: Dr. Jodi Fasteen

April 25, 2024

Abstract

It is the goal of this project to accurately predict the results of the 2022 Qatar World Cup and develop a sophisticated model that can be applied to future World Cups. The FIFA 2022 World Cup was home to exciting upsets, devastating losses and unexpected results. FIFA rankings and past performance statistics were insufficient predictors for results, advancements and performances in the World Cup. In this project, machine learning algorithms and predictors beyond rankings will be used to predict the results of the Qatar 2022 World Cup. The FIFA ranking prediction method will serve as a baseline for accuracy. The two machine learning algorithms that will be considered for this project are decision trees and random forests, the latter of which can determine the validity of various parameters. The random forest algorithm created a more effective model for predicting the results of World Cup matches, exceeding the FIFA baseline accuracy by nearly 20% on individual games and by 6% in terms of team advancements. Some of the better predictors for World Cup matches were win proportions, FIFA points, the standard deviation in the age of players on the team and the average number of goals scored per game. While the random forest algorithm was a better predictor of games than both the FIFA baseline and decision tree models, it did not entirely accurately forecast the knockout bracket of the 2022 Qatar World Cup.

Introduction

Predicting the outcomes of sporting events remains difficult because of the unpredictability of the day, match and players, which includes the world of soccer. In this paper, we will consider some of the most well-respected methods, including Elo Ratings, for evaluating teams' strengths and their probability of winning in matches against their prospective opponents in the world of soccer in order to evaluate their relative effectiveness. The Elo Ratings utilize a logistic function that transforms a difference in rankings to a win probability (Palánek). Specifically, FIFA's official rating system, based on the Elo Ratings methodology, will be used as a baseline for predicting World Cup results because they are the current standard for predicting results. Some ranking systems commonly used in sports beyond Elo include a neighbor evaluation, a least squares rating and a network based system (Lasek et al., 35-37). However, these models are built around who a team plays and their past performance. While these rankings take into account every match that a team plays and are adjusted accordingly, they do not use information about players, goal histories or future vs. past performance. Because of this deficit, we will develop a method for rating teams that serves as a more accurate predictor for matches using decision trees and random forests, two machine learning algorithms that can take into account a variety of factors at the same time. Throughout the paper, we will use the actual Qatar 2022 World Cup results to evaluate the accuracy of both the baseline and the machine learning

algorithms to see which performs better in this context. Furthermore, by using Random Forests, we will be able to analyze what factors best serve as predictors for this particular World Cup, with the potential of application to future World Cups.

Elo Ratings

The Elo rating was originally developed to rate chess players, but is now used in a variety of other contexts. The Elo rating system has been adapted not only to rate players from a variety of sports, including soccer, but also in educational and professional contexts, demonstrating the versatility of this equation (Pelánek). Under the Elo system of rating, each team receives or loses points after each match that they compete in. If they win, points are added to their rating, while points are subtracted from teams when they lose. In the Elo ratings, the points earned or lost by a team after each match is proportional to the surprisingness of the result (Pelánek). These ratings also take into account the importance of the match; matches that are higher pressure and have more at stake will either reward or punish the teams participating to a greater extent than less influential games. In this section, we will delve into how Elo ratings are used to calculate soccer team ratings, specifically, though this methodology can be applied to a variety of fields.

The official equation used to evaluate soccer ratings is:

$$R_n = R_o + K(W - W_E) \text{ ("World Football Elo Ratings.")}$$

R_n is the rating of a team following a new match.

R_o is the number of ranking points a team previously had before the match.

W refers to the result of the match, as either a win, a loss or a draw. The table below shows the points per result.

Result of Match	W value
Win	1
Loss	0
Draw	0.5

Table 1: W values for the results of the match in the Elo Ratings

K is defined as the importance of the match. This value ranges depending on the level of match, ranging from friendly matches to finals of the World Cup. Games won at a higher level will increase a team's rating more significantly. See the table below for the specific breakdown of values for the variable K .

Type of Match	Value of K
Friendly matches	20
Other tournaments	30
World Cup and continental qualifiers and major tournaments	40
Continental championship finals and major intercontinental tournaments	50
World Cup Finals	60

Table 2: K values for the types of matches in Elo Ratings

K also takes into account the goal differential of the match. It is increased by half if the game is won by two goals and by three quarters if the game is won by three goals. For any goal differential greater than three, K is increased by the amount: $\frac{3}{4} + \frac{N-3}{8}$, where N is the goal difference.

For example, if a team wins a match in the category ‘Other tournaments’ by four goals, the resulting K value can be calculated as follows:

$$K = 30 + 30\left(\frac{3}{4} + \frac{4-3}{8}\right)$$

$$K = 30 + 30\left(\frac{3}{4} + \frac{1}{8}\right)$$

$$K = 30 + 30\left(\frac{7}{8}\right)$$

$$K = 30 + \frac{105}{4}$$

$$K = 56.25$$

W_E is the predicted outcome of the match. It is calculated by evaluating the difference in teams’ point totals (dr): $W_E = \frac{1}{10^{\left(\frac{-dr}{400}\right)} + 1}$. A team’s current point total is a cumulative calculation that takes into account all previous matches. A team with a greater point total than its opponent will have a better expected result. In the Elo system, 100 points is added to the dr for the team that plays at home, giving a home field advantage.

This system of rating teams is very similar to the current system used by FIFA to rate its teams ahead of the World Cup.

FIFA Ratings

The formula for the men’s FIFA world rankings, known as SUM, recalculates a team’s ranking based on every game they play, just as the Elo ratings system does (“Revision of the FIFA/Coca-Cola World Rankings”). In this formula, the expected result of the match, the match’s result and the importance of the match all play a role in calculating a team’s new ranking. However, unlike the Elo rating, the FIFA ranking system does not take into account a home field

advantage or adjust the match importance level based on the goal differential.

However, it does include more values for the match importance level.

The FIFA world rankings equation is: $P = P_{before} + I(W - W_e)$

In this formula, the new team ranking (in points) is calculated by summing the points that the team had before the match (P_{before}) with the product of the importance of the match (I) and the difference between the result of the match and the expected result of the match ($W - W_e$).

For FIFA, the matches are more vigorously defined and account for a greater variety than the Elo Ratings' system, leading to a wider possibility of values for I . Importantly, any World Cup match from the quarterfinals and beyond are rated the same.

Type of Match	Value of I
Friendly matches played outside of International Match Calendar windows	5
Friendly matches played during International Match Calendar windows	10
Group phase matches of Nations League competitions	15
Play-off and final matches of Nations League competitions	25
Qualification matches for Confederations final competitions and for FIFA World Cup final competitions	25
Confederation final competition matches up until the QF stage	35
Confederation final competition matches from the QF stage onwards; all FIFA Confederations Cup matches	40
FIFA World Cup final competition matches up until QF stage	50
FIFA World Cup final competition matches from QF stage onwards	60

Table 3: I values for the FIFA system

While under the FIFA system, there are some instances in which W can have a value of 0.75 or 0.25 (due to a win or loss after regular time or because of penalty kicks), the majority of matches are scored the same as in the Elo system, with a 0, 1 or 0.5.

The expected result W_E is calculated by evaluating the difference in

teams' point totals (dr): $W_E = \frac{1}{10^{\left(\frac{-dr}{600}\right)} + 1}$. A team with a greater point total than

its opponent will have a better expected result. Note the similarity in formula to the Elo rating. There is no home field advantage for the predictions in the FIFA system.

To fully understand how the FIFA/Elo Ratings systems function, an example is provided below. In this situation, we will use the FIFA ratings equation because it is the most relevant for soccer ratings.

Example: Say a Team A with 1000 points plays Team B with 950 points in a friendly match during the International Match Calendar Window and Team A wins 2-1. In this scenario, for Team A, $P_{before} = 1000$, $I = 10$, $W = 1$ and $dr = 50$.

For Team B, $P_{before} = 950$, $I = 10$, $W = 0$ and $dr = -50$

$$P_{Team A} = 1000 + 10\left(1 - \frac{1}{10^{\left(\frac{-50}{600}\right) + 1}}\right)$$

$$P_{Team B} = 950 + 10\left(0 - \frac{1}{10^{\left(\frac{-50}{600}\right) + 1}}\right)$$

$$P_{Team A} = 1000 + 10\left(1 - \frac{1}{0.825 + 1}\right)$$

$$P_{Team B} = 950 + 10\left(0 - \frac{1}{1.21 + 1}\right)$$

$$P_{Team A} = 1000 + 10(1 - 0.548)$$

$$P_{Team B} = 950 + 10(0 - 0.452)$$

$$P_{Team A} = 1000 + 10(0.452)$$

$$P_{Team B} = 950 + 10(-0.452)$$

$$P_{Team A} = 1000 + 4.52$$

$$P_{Team B} = 950 - 4.52$$

$$P_{Team A} = 1004.52$$

$$P_{Team B} = 945.48$$

Note that in both cases, either a deficit or addition of 4.52 points was applied to each team. For FIFA, matches are predicted based on the number of

points that a team and its opponent have. In short, FIFA predictions are solely based on previous results of the team.

FIFA Ratings as a Baseline

In order to evaluate how well the decision tree and random forests will perform in predicting the results of the 2022 World Cup, we will create a baseline model based on the current FIFA system. This is the system currently used to predict the results of soccer matches and will function to demonstrate whether rankings and past performance is sufficient for predicting future performance. Using the FIFA rankings before the 2022 World Cup, we will predict the teams that move on to the knockout stages and build a bracket to determine the finals.

The World Cup is divided into two stages: the group stage and the knockout stage (Arrieta-Kenna). In the group stage, the 32 teams that qualified for the world cup are divided into eight groups, each group with four teams. Every team plays each team in their group for a total of six group play games per group, with three games for each team. During this stage of the tournament, draws are allowed and earn 1 point, while a win earns a team 3 points and a loss is 0 points. This is consistent only in the group stage, before the knockout rounds. The top two teams that accumulate the most points in their group during the group stage will advance to the round of 16. Only half of the teams will advance to the knockout rounds. The tournament is set up in such a way that the teams predicted to do the best (ranked highest) are usually in different groups so that all of the favorites have a good chance of moving on to the knockout stages and that the

later games will be as even and exciting as possible. In the event that two teams have the same number of points, goal differential and head-to-head matchups will determine who moves on. In the Round of 16, teams that accumulated the most points in their group play teams that advanced as second in their group. This method gives an advantage to teams that played the best in the group stage.

FIFA's methodology for predicting the expected result, as demonstrated above, relies only on a team's number of points before the match and their opponent's number of points. Whichever team has a greater point value is expected to win. The size of the difference in the number of points is reflected in each team's probability of winning. Teams that are very close together will have closer to a 50/50 split, while uneven matches will be reflected with more skewed probabilities.

To create a baseline for predictions, we will use FIFA's predicted results equation, $W_E = \frac{1}{10^{\left(\frac{-dr}{500}\right)} + 1}$. For the purposes of making its predictive power more reasonable, teams that have a winning probability between 0.45 and 0.55 in a game will be predicted to draw that game in the group stage in order to accurately reflect each team's relative strength. In the knockout rounds, draws will not be allowed in the model. Rather, there, the tie-break will be head-to-head results and then FIFA point differential. Because these algorithms predict only results, not goal differential, that tie-break system cannot be implemented here.

Example of Group A's Expected Results:

Team 1	Team 2	Team 1 FIFA Points	Team 2 FIFA Points	Expected Result Team 1	Expected Result Team 2	Winner
Qatar	Ecuador	1439.89	1464.39	0.48	0.52	Draw
Netherlands	Senegal	1694.51	1584.38	0.60	0.40	Netherlands
Qatar	Senegal	1439.89	1584.38	0.36	0.64	Senegal
Netherlands	Ecuador	1694.51	1464.39	0.71	0.29	Netherlands
Netherlands	Qatar	1694.51	1439.89	0.73	0.27	Netherlands
Ecuador	Senegal	1464.39	1584.38	0.39	0.61	Senegal

Table 4: The expected results of the group stage for Group A in the 2022 World Cup

In Group A, both Senegal and Netherlands are expected to advance to the knockout round. The Netherlands have a high enough point differential that they are expected to win all three of their games; their lowest probability of winning is still above 60%, which is a good indication that they are big favorites to advance from Group A. Neither Ecuador or Qatar are expected to advance.

In the group stage, the FIFA predictions always predicted that the favorite team would win all three of their games and that the team with the second highest rating in the group would win two of their games. With this method of predicting, the FIFA expected win formula only accurately predicted 25 of the 48 group games.

The table below demonstrates how FIFA's method of predicting matches up against the actual results of the 2022 World Cup. Those that were expected to

advance but did not are in red, while those that were expected to advance but did so in a different position are in orange.

Group	Predicted to Advance	Actually Advanced
A	1. Netherlands 2. Senegal	1. Netherlands 2. Senegal
B	1. England 2. United States	1. England 2. United States
C	1. Argentina 2. Mexico	1. Argentina 2. Poland
D	1. France 2. Denmark	1. France 2. Australia
E	1. Spain 2. Germany	1. Japan 2. Spain
F	1. Belgium 2. Croatia	1. Morocco 2. Croatia
G	1. Brazil 2. Switzerland	1. Brazil 2. Switzerland
H	1. Portugal 2. Uruguay	1. Portugal 2. South Korea

Table 5: The FIFA predictions for who would advance past the group stage compared to the actual advancements.

Of the 16 teams that advanced, FIFA would have predicted 11 of them correctly, though it did not accurately portray the position that Spain would advance in. Although the FIFA method was only 52% accurate in predicting individual game results, FIFA predictions of who would advance past the group stage were relatively accurate.

We will update the points for each team based on their results in the group stage and use their updated scores to calculate who will progress through the

knockout stages to the final. In this section, draws will not be allowed and the team with the most points will be the team expected to win and move on. Below are the predicted and actual brackets from the Round of 16 to the Finals. In the predicted bracket, the teams in red are incorrectly predicted to advance. The winner of the World Cup is in bold.



Figure 1: Actual Bracket for 2022 Qatar World Cup Knockout Stages



Figure 2: FIFA baseline Predicted Bracket for 2022 Qatar World Cup Knockout Stages

In the quarterfinals, the FIFA predictions were nearly 88% accurate, while in the semifinals, they were only 50% accurate. In the finals, the FIFA predictions were wholly inaccurate, predicting the final occurring between two teams that didn't make it past the quarterfinals. Overall, this makes the FIFA prediction method 65% accurate in choosing teams that would qualify for the knockout stages and teams that would win games during the playoffs, while only 54% accurate in predicting the results of each game. Both of these scores will be used to evaluate the accuracy of both the baseline and of the later models.

$$\begin{aligned}\text{Team Accuracy Score} &= \frac{10.5 + 7 + 2 + 0}{16 + 8 + 4 + 2} = 0.65 \\ \text{Game Prediction Accuracy} &= \frac{25 + 7 + 2 + 0 + 0}{48 + 8 + 4 + 2 + 1} = 0.54\end{aligned}$$

Prediction Factors

Because FIFA scores are an evaluation of a team's previous performance, they serve as useful predictors for future performance. However, as demonstrated above, FIFA scores are not sufficient for accurately predicting the results of World Cup games. In order to form a better model for predicting games, we will consider a variety of other factors, including average age of the team and the average number of goals scored by the team. The table below offers the names and descriptions of the factors that will be used in our machine learning algorithms for the remainder of the project. Factors like average goals scored per game and winning proportion offer a look at a team's past performance in a more nuanced way than FIFA ratings, while factors like the standard deviation of the age of the

players might take into account some form of team compatibility or balance. Total appearances on the roster is a measure of a team's experience. It is the hope that including factors that take into account other aspects of the team will improve the predictions for the 2022 Qatar World Cup.

Factor	Description
Result	The actual results of the FIFA games, listed as either 0, 0.5 or 1 to indicate a loss, draw or win. This is in reference to Team 1, so a result value of 0 indicates that Team 1 lost and Team 2 won.
Overall Average Goals Per Game (10 Years)	The difference in the average number of goals per team in the game, home or away, the averages calculated over the last ten years.
Overall Average Goals Per Game (5 Years)	The difference in the average number of goals per team in the game, home or away, the averages calculated over the last five years.
Overall Proportion	The difference between the two teams of their average win proportions for ten years
Average Proportion (5 Years)	The difference between the two teams of their average win proportions for five years.
Total Goals on Roster	The difference between Team 1's total goals on roster and Team 2's total goals on roster. These goals are a sum of the players' total goals scored professionally.
Total Appearances on Roster	The difference between the total number of appearances at the international level in games on the roster between Team 1 and Team 2.
Average Goals	The difference between the average number of goals scored per player for Team 1 and Team 2.
Average Appearances	The difference between the average number of international appearances per player for Team 1 and Team 2.
Average Age	The difference between the average age of a player on Team 1 and Team 2.
Standard Deviation Age	The difference between the standard deviation of the ages of the players on Team 1 and Team 2.
Number of Players > 12 goals	The difference between the number of players with more than 12 recorded goals on Team 1 and Team 2.
Players with > 30	The difference between the number of players with more

goals	than 30 recorded goals on Team 1 and Team 2.
FIFA Points	The difference between the number of FIFA points between Team 1 and Team 2 at the start of the tournament.

Table 6: Data Dictionary for the factors used in our machine learning algorithms

Each game in the dataset has a value for factors like the average number of goals scored per player on the team over the last five years and the win proportion of the team in the last ten years. These are each expressed as a difference between the teams that are playing in the game. Each game is an entry in the database and for every entry in the database, we will subtract the second team's average number of goals from the first team's. For example, if the average number of goals by Team 1 in the last five years is 2.1 and for Team 2 is 1.83, the value in the database is 0.27. The values of 2.1 and 1.83 do not appear in the database because it is formed around games, not teams. This is true for every predictor in the database. Each value is an expression of subtraction between the two teams that are playing the game. The results are situated so that they are also in reference to Team 1. A recorded loss implies that Team 1 lost and Team 2 won. Losses, wins and draws are the possible results. In the following section, we will use decision trees, a machine learning algorithm, to evaluate the effectiveness of our predictors above.

Decision Trees

Decision trees are machine learning algorithms that can be used in classification and regression problems effectively. In this context, our decision trees will be trained through the Classification and Regression Tree (CART) algorithm and used to classify teams as either winning or losing. This algorithm functions by splitting the training data into two subsets using a feature k and a threshold t_k (Géron 182). For instance, in the World Cup, a feature k could be the difference of points between the two teams and the threshold t_k could be 0. This decision tree would probably make the same predictions as the FIFA score, where the team with the greater number of points would be predicted to win. However, decision trees can take into account many variables at the same time, which makes them a powerful classification tool. To continue this process with more than one factor, the algorithm would use another feature and another threshold to split the subsets and repeat the process until all the features are used appropriately (Géron 177). The maximum number of times that a decision tree splits the data is known as the max depth of the tree.

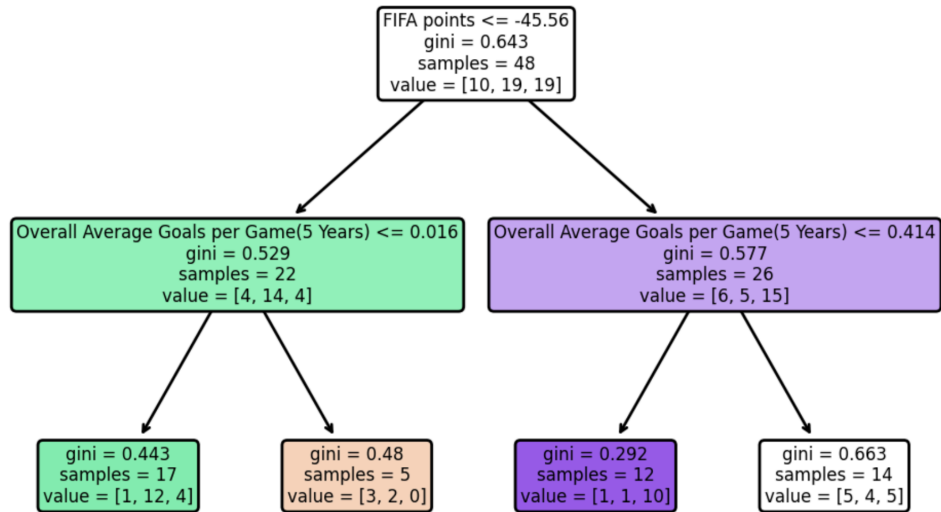


Figure 3: Decision tree using factors FIFA points and average number of goals per team in the last five years. Results are formatted as [Draw, Loss, Win].

The top of the decision tree is known as the root node, and the two nodes immediately below it are termed children nodes. Nodes that do not have any children, or that remain unsplit, are known as leaf nodes. In the above figure, there is one root node, two children nodes directly resulting from it and their four children nodes are all leaf nodes (Géron 179). The colors of each node indicate what result the majority of the sample is. The leaf node that is beige is the only node whose sample contains a majority of draws as a result, while the green nodes are mostly losses and the purple majority victories. The darker colored nodes have a stronger majority than the paler ones. The nodes that are the closest to white have no clear majority.

The decision tree in Figure 3 was created using two of the factors in our prediction database: FIFA points and average number of goals scored by the team over the last five years leading up to the World Cup. For a closer look at the components of the decision tree, reference Figure 4 below. Note that this tree uses

two factors and has a max depth of two, which only allows the algorithm to split the data twice, in this case, once per factor.

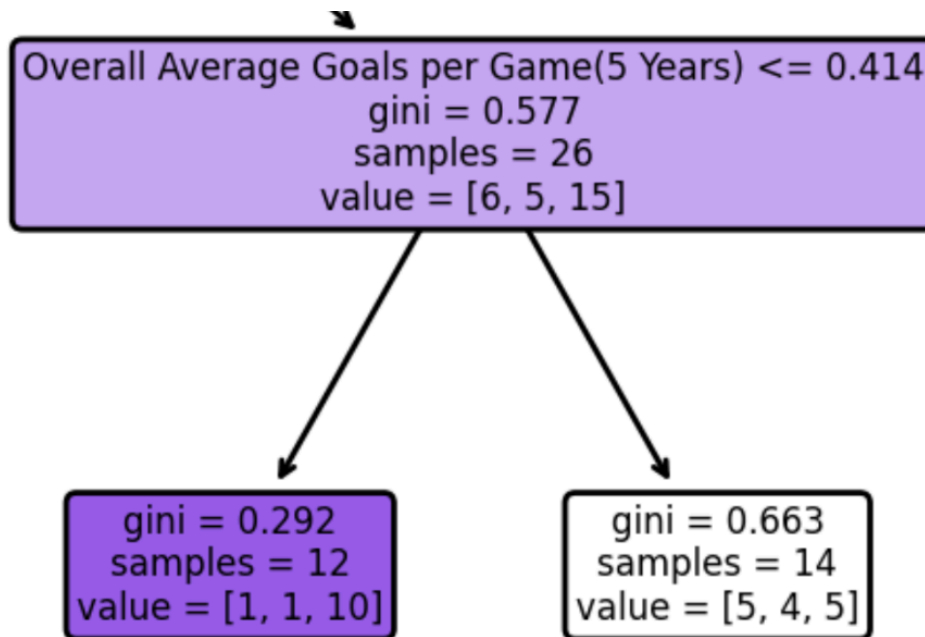


Figure 4: Partial Decision Tree based on FIFA points and average number of goals per team in the last five years.

The above figure looks at the right child node and its two children nodes.

Note that the parent node has Overall Average Goals per Game (5 Years) ≤ 0.414 as its first row. The split that this node is making is between games whose teams have a difference of fewer than 0.414 goals per game over the last five years and games whose teams have a difference that is greater than 0.414 goals. Games that are between teams with a difference in goals over the last five years fewer than 0.414 are sorted to the left node, while the rest of the games are sorted to the right. Samples indicate how many games are at each node.

The gini score refers to the purity of the sample. If all of the samples in the node are in the same category, the gini score would be 0.0 because that sample would be perfectly pure, while a gini score of 0.5 indicates impurity as a random

selection of classes. The closer the gini score to 0.5, the less pure the sample and the worse our predictions will be. The gini score for a three category node is calculated by the formula:

$$\text{gini} = 1 - p_1^2 - p_2^2 - p_3^2$$

where p_n refers to the proportion expressed by $\frac{\text{number of samples in group } n}{\text{total number of samples in node}}$ (Géron 180).

In the case of the leftmost leaf node, the calculation is as follows.

$$\text{gini} = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{4}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.663$$

This leftmost node is not very pure, as indicated by the relatively high gini score.

The decision tree algorithm functions by finding the best possible set of splits across all possible splits through calculating a weighted gini score for each split it makes. This weighted gini score acts as a cost function, which the decision tree seeks to minimize by evaluating all possible combinations of splits and calculating the gini score for each group it forms through its splits. The dividers that minimize the weighted gini score are selected as the final model's decisions. The greater the number of splits the tree is allowed to make, the more accurate the decision tree can be, but the more computationally expensive the algorithm is.

For the purposes of this project, we will build a decision tree based on all the factors in our database that will be allowed to make two splits (`max_depth = 2`). This model will be compared to the baseline FIFA model.

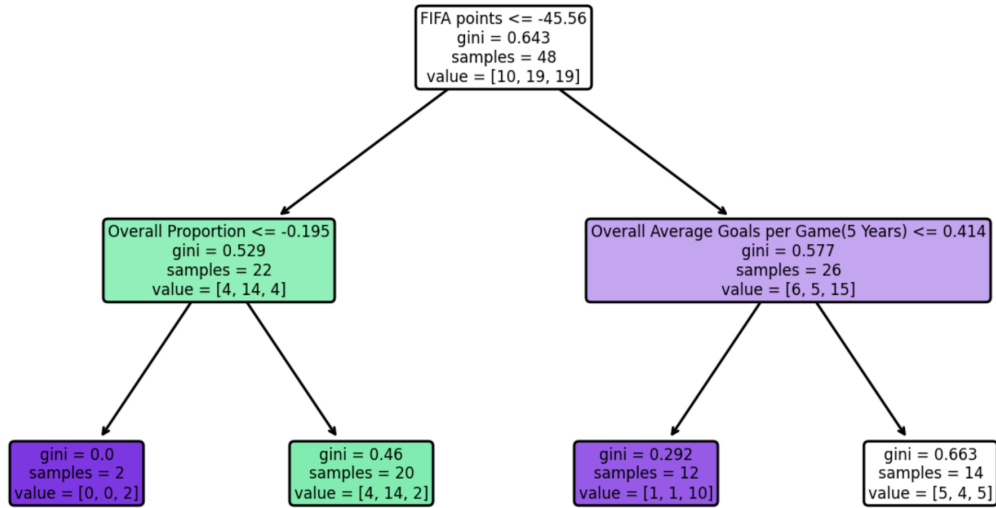
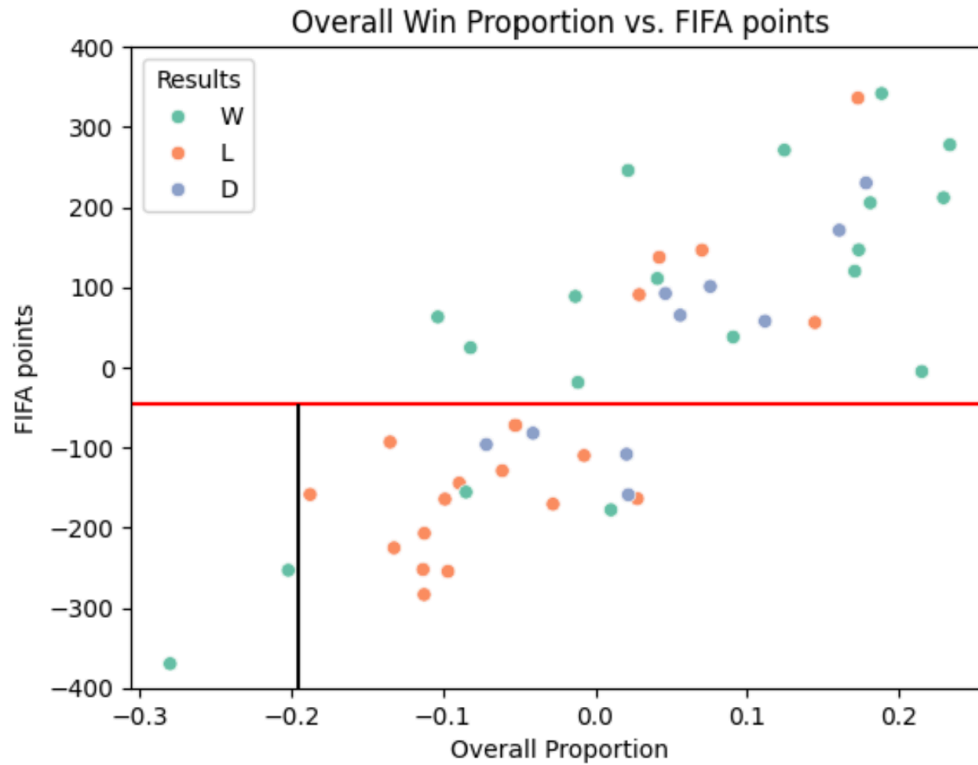


Figure 5: Decision tree to predict group stage of 2022 World Cup.

In this decision tree, the results are formatted as [D, L, W]. As seen above in Figure t, the first split was along FIFA points where games between teams with a difference in FIFA points less than -45.6 were sorted into one group and those with a greater difference were sorted into the other. Note that the majority of games sorted left are losses (14/22), while a majority of the games sorted right are wins (15/26). That implies that having more FIFA points than one's opponent increases the probability of winning. From there, the decision tree made splits along the teams' difference in overall win proportion over ten years and in their difference in overall average goals per game over five years.



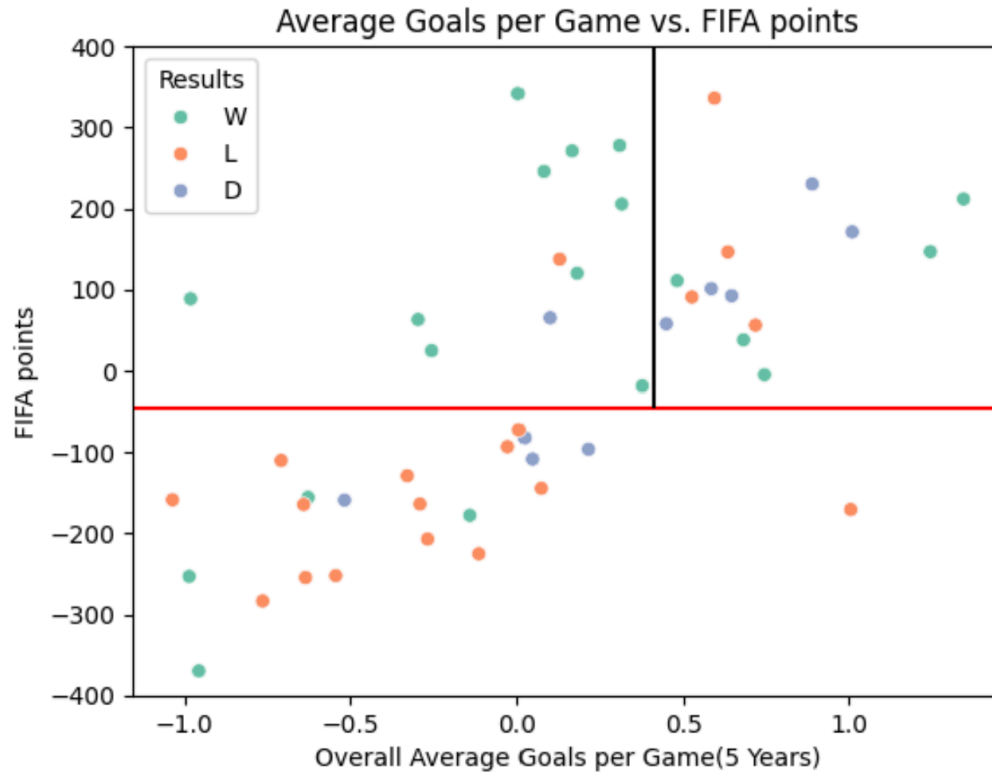


Figure 7: Left side of the tree data splits along FIFA points and Overall Average Goals per Game (5 Years).

On the right side of the tree, the split is rather similar. Games that result in wins have an average goal difference of less than 0.414 between their opponents, while teams that have a greater goal differential than their opponent over five years are more likely to lose or tie. While the splits might seem not intuitive when looking at the decision tree, the scatterplots demonstrate why the splits are being made there. These are the splits that best separate the wins, losses and draws from each other.

This decision tree had an overall accuracy rate of 0.65, with a training accuracy rate of 0.66 and a testing accuracy of 0.6. Note that the FIFA baseline only had a group stage accuracy rate of 0.52, so the decision tree is better at predicting the group play game results than the FIFA baseline. The table below

demonstrates how each model would select the teams that move on to the knockout stages.

Group	Decision Tree Predictions	FIFA Predictions	Actually Advanced
A	1. Netherlands 2. Senegal	1. Netherlands 2. Senegal	1. Netherlands 2. Senegal
B	1. England 2. United States	1. England 2. United States	1. England 2. United States
C	1. Argentina 2. Mexico	1. Argentina 2. Mexico	1. Argentina 2. Poland
D	1. France 2. Denmark	1. France 2. Denmark	1. France 2. Australia
E	1. Spain 2. Costa Rica	1. Spain 2. Germany	1. Japan 2. Spain
F	1. Croatia 2. Morocco	1. Belgium 2. Croatia	1. Morocco 2. Croatia
G	1. Brazil 2. Switzerland	1. Brazil 2. Switzerland	1. Brazil 2. Switzerland
H	1. Uruguay 2. Portugal	1. Portugal 2. Uruguay	1. Portugal 2. South Korea

Table 7: Random forest and FIFA baseline advancement predictions compared to the actual advancements

Note that the decision tree performed slightly worse than the FIFA baseline predictions in terms of what teams would advance out of the group stage. It accurately picked 10/16 teams to qualify for the knockout rounds, while the FIFA baseline accurately picked 10.5/16 teams. However, the game accuracy rate in the group stage was better in the decision tree classification, but that did not, in this case, lead to better results.

Random Forest Algorithm Predictors

Decision trees are excellent machine learning algorithms that are readily applicable to classification problems. However, random forests are algorithms that are even more effective because they build on decision trees. Random forests function by training many decision trees on random subsets of the features provided and then by averaging their results (Géron 78). An algorithm that combines many models (like multiple decision trees) is called *Ensemble Learning* and is more effective than the original models. Random forests are also useful because they can measure the relative importance of various features (Géron 200). Because random forests are testing different features in individual decision trees, they can average each feature's relative utility in the many Decision Trees to quantify their impact. This will be useful in deciding which factors are best for predicting World Cup results.

For the purposes of this project, we will split the group play games (48) into a testing and training split to determine the best parameters for predicting World Cup games. Only once those have been established will we apply the metrics to the knockout stages. This is critical because the knockout stages cannot be evaluated until the random forest algorithm selects the 16 teams it finds most likely to advance. Once it has selected the 16 teams for advancement, we will evaluate how it performs in the knockout stages of the tournament by building its hypothetical playoff bracket.

Because there are only 48 games in the group stage, the training/testing split will have a large influence on the random forest algorithm's decisions based

on which section of the group stage games it can see. In order to accurately predict the World Cup games, we will split the data five times and average the results of those five models. All five models will use a testing size of 20% of the data, randomly selected by a random seed and each will have a max depth that maximizes its relative testing accuracy. The results of these five training/testing splits are summarized below:

Random Forest Algorithm #	Random Seed	Max Depth	Testing Accuracy	Training Accuracy
1	912	6	0.6	0.95
2	88	6	0.6	1.0
3	635	6	0.5	0.90
4	38	5	0.7	0.92
5	303	3	0.5	0.71

Table 8: Random forest algorithm various test/train split results

The testing accuracy of the models ranged from 0.5 - 0.7, while the training accuracy was higher. By averaging the results of these models, we can create a random forest algorithm that takes into account the entire dataset while avoiding overfitting the data.

The random forest algorithm can generate both testing/training accuracy scores and an importance rating for each factor used. The higher the importance rating, the better the predictor the variable is. Table 9 below gives the results for which variables were most impactful in making good predictions.

	Feature Importance
Overall Proportion	0.127953
FIFA points	0.121166
Standard Deviation Age	0.097899
Overall Average Goals per Game (5 Years)	0.091785
Overall Average Goals per Game (10 years)	0.091280
Average Proportion (5 years)	0.082609
Average Age	0.074753
Average Appearances	0.073605
Total Appearances on Roster	0.073511
Average Goals	0.056467
Total Goals on Roster	0.055896
Number of Players > 12 goals	0.032044
Players with > 30 goals	0.021032

Table 9: Importance ratings for factors in the Random Forest Algorithm

The relative importance of a variety of factors imply that FIFA rankings are not sufficient to determine the probability of winning the World Cup. Rather, the team's overall winning proportion over the last ten years is critical to predicting World Cup games followed by FIFA points. Factors like the standard

deviation of age of the players and overall average goals per game for the last 5 and 10 years are also valuable, demonstrated by their relatively high importances.

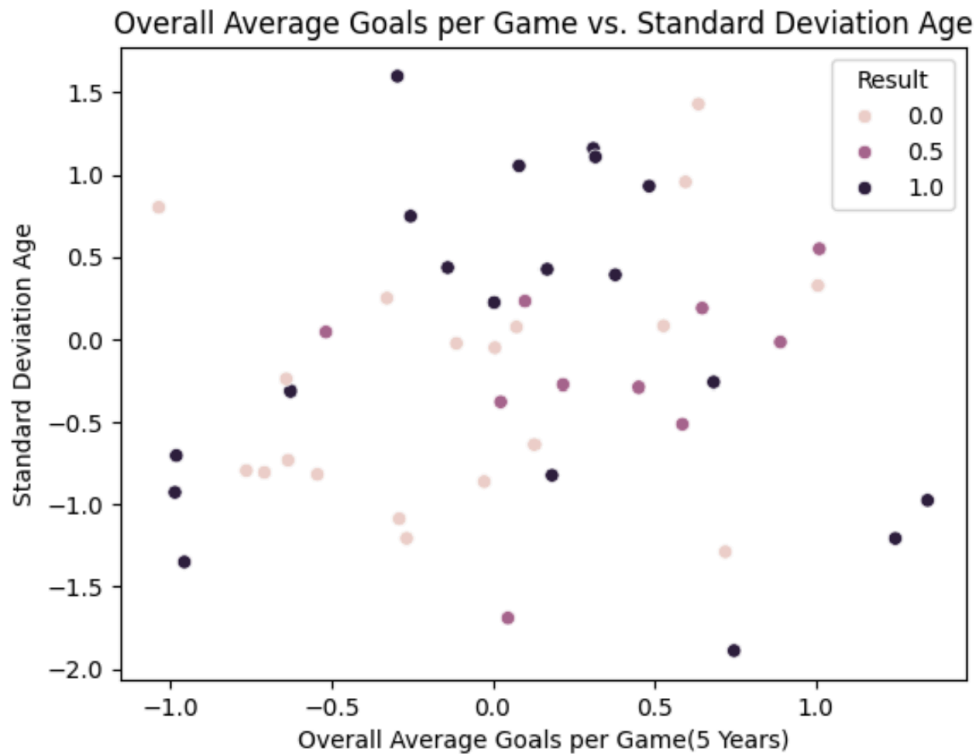


Figure 8: Overall Average Goals per Game vs. Standard Deviation Age as important factors for predicting the 2022 World Cup results

As seen in Figure 8, standard deviation functions relatively well to group games with similar results together. A majority of the winning games are clustered towards the top of the graph, implying that on average, having a greater standard deviation on the team makes a win more likely. However, there are quite a few teams with a greater standard deviation that did not win their games, demonstrating that increasing a team's standard deviation of age does not guarantee better results. On the other hand, the factor of average number of goals

on a team is a bit more elusive, as it does not immediately easily split the data into distinct sections. However, there is still some grouping of wins and losses when both features are analyzed in the graph below and both factors were useful for predicting the group stage of the 2022 World Cup.

Recall from the decision tree section that Overall Proportion and FIFA points were two predictors particularly well suited to separating the wins, losses and draws from each other. Comparatively, the factors that have lower importance, like the number of players on the team with a high goal count (either 12 or 30) are not as easily able to split the data.

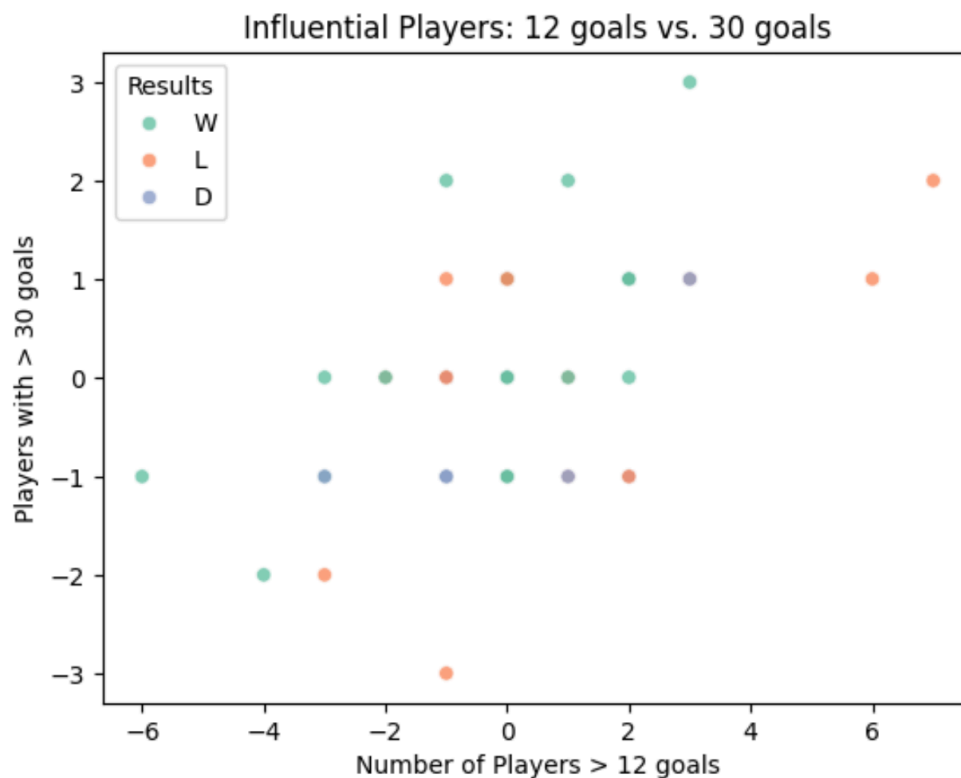


Figure 9: An evaluation of the factors Number of Players >12 goals vs. Players with > 30 goals

Note that here the losses are less easily separable from the wins. This indicates that the presence of a player with a lot of goals under their belt does not guarantee victory and that a team without a particularly famous player can still be dangerous. These factors, therefore, did not have a lot of weight in the random forest algorithm predictions.

Random Forest Algorithm Predictions

The table below evaluates the predictions of the random forest algorithm against the baseline predictions and the original.

Group	Random Forest Predictions	FIFA Baseline Predictions	Actually Advanced
A	1. Netherlands 2. Senegal	1. Netherlands 2. Senegal	1. Netherlands 2. Senegal
B	1. England 2. United States	1. England 2. United States	1. England 2. United States
C	1. Argentina 2. Poland	1. Argentina 2. Mexico	1. Argentina 2. Poland
D	1. France 2. Australia	1. France 2. Denmark	1. France 2. Australia
E	1. Spain 2. Germany	1. Spain 2. Germany	1. Japan 2. Spain
F	1. Morocco 2. Croatia	1. Belgium 2. Croatia	1. Morocco 2. Croatia
G	1. Brazil 2. Switzerland	1. Brazil 2. Switzerland	1. Brazil 2. Switzerland
H	1. Portugal 2. Uruguay	1. Portugal 2. Uruguay	1. Portugal 2. South Korea

Table 9: Advancement predictions for random forest algorithm and FIFA baseline.

One way to calculate the accuracy of each model is to count each team that is correctly ranked as either the winner or the runner up of their bracket as 100% correct, while a team that does advance but in the incorrect order is 50% correct. In this method, the FIFA baseline had an accuracy score of about 65% of who would advance past the group stage (and how), while the random forest algorithm had an accuracy score of 84%. However, of the 16 teams that advanced, the FIFA prediction system only correctly identified 11 of them, while the random forest algorithm accuracy predicted 13 of the 16 teams that would advance past the group stage.

$$\text{Team Accuracy Score} = \frac{13.5 + 6 + 2 + 1}{16 + 8 + 4 + 2} = 0.717$$

$$\text{Game Prediction Accuracy} = 0.825$$

Random forest Scores

Recall from above that the FIFA team accuracy score was 0.65, while the FIFA game prediction accuracy was 0.52. In both measures of accuracy, the random forest algorithm performs better than the FIFA prediction baseline, even if only slightly in the team accuracy score.

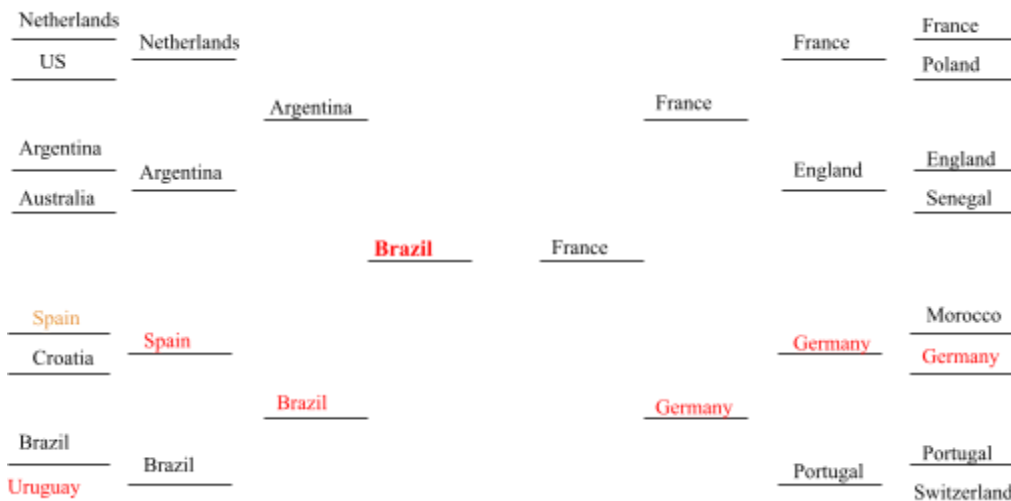


Figure 9: Random forest algorithm predicted bracket for 2022 Qatar World Cup knockout stages



Figure 10: FIFA baseline predicted bracket for 2022 Qatar World Cup knockout stages

The random forest algorithm had an overall game accuracy rate of 89.6% throughout the group stage. In the knockout stages, the random forest algorithm accurately predicted the contenders and results for four of the matchups in the round of 16 (compared to three in the FIFA predictions), and like the FIFA baseline predictions, accurately selected two of the four semifinalists (Argentina

and France). However, the random forest algorithm predicted a showdown between Brazil and France, while the FIFA predictions forecasted a final between Brazil and Belgium. While both the FIFA and the random forest algorithm incorrectly selected Brazil as the overall winner of the World Cup instead of Argentina, the random forest performed slightly better in correctly identifying France as a finalist and by initially predicting better teams to make it to the knockout stages. Overall, the random forest algorithm functioned as a better method for predicting the World Cup results, but was still subject to the unpredictability of sporting events.

Conclusion

The random forest algorithm was more effective at predicting the results of the 2022 FIFA World Cup than a method based on the FIFA rankings. It performed more than 20% better on predicting the results of individual matchups and 6% better on the results of what teams would advance past each stage of the tournament. It was also more successful at selecting the teams that would advance past the group stage and accurately predicted that France would reach the finals. However, like the FIFA baseline, it still failed to accurately predict Argentina's victory at the finals and instead selected Brazil.

The random forest algorithm used factors beyond FIFA rankings to determine the winners of each match and utilized information specifically about the team's previous winning records and evolution to improve its predictions. Specifically, factors like the team's overall winning proportion over a number of years and the

standard deviation of the age of the players on the team functioned as good predictors for World Cup results. Surprisingly, the presence of a player with a lot of professional goals to their name did not necessarily increase a team's probability of winning and did not serve as a good predictor for wins or losses in the World Cup. In short, predicting the results of the World Cup remains a challenge, but using information about a team's winning record, average goals, age metrics and their world ranking is relatively effective in at least identifying the teams that will advance past the group stage.

Works Cited

“2022 FIFA World Cup Squads.” *Wikipedia*, 11 Feb. 2024.

https://en.wikipedia.org/w/index.php?title=2022_FIFA_World_Cup_squads&oldid=1206103711.

Arrieta-Kenna, Ruairí. “Here’s How Each Team Can Make It to the Next Round of the World Cup.” *Time Magazine*, November 27, 2022,

<https://time.com/6237295/world-cup-group-stage-permutations/>.

“FIFA World Cup Knockout and Groups.” *FIFA*,

<https://www.fifa.com/fifaplust/en/tournaments/mens/worldcup/qatar2022/knockout-and-groups>.

Géron, Aurélien. *Hands-on Machine Learning with Scikit Learn, Keras and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems*. 2nd ed., O’Reilly, 2019.

International Football Results from 1872 to 2024.

<https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017>.

Lasek, Jan, et al. “The Predictive Power of Ranking Systems in Association Football.” *Int. J. of Applied Pattern Recognition*, vol. 1, Jan. 2013, *ResearchGate*,

<https://doi.org/10.1504/IJAPR.2013.052339>.

Men’s Ranking, InsideFifa, February 2024,

<https://inside.fifa.com/fifa-world-ranking/men?dateId=id13792>.

Pelánek, Radek. “Applications of the Elo rating system in adaptive educational systems,

Computers & Education.” Vol. 98, 2016, pp. 169-179,

<https://doi.org/10.1016/j.compedu.2016.03.017>.

“Revision of the FIFA/Coca-Cola World Rankings”. *FIFA*,

<https://digitalhub.fifa.com/m/f99da4f73212220/original/edbm045h0udbwkqew35>

[a-pdf.pdf](#)

“World Football Elo Ratings.” <https://www.eloratings.net/about>.

Appendix

Table of World Cup groups

Group	Teams
A	Netherlands (1) Senegal (2) Ecuador (3) Qatar (4)
B	England (5) United States (6) Iran (7) Wales (8)
C	Argentina (9) Poland (10) Mexico (11) Saudi Arabia (12)
D	France (13) Australia (14) Tunisia (15) Denmark (16)
E	Japan (17) Spain (18) Germany (19) Costa Rica (20)
F	Morocco (21) Croatia (22) Belgium (23) Canada (24)
G	Brazil (25) Switzerland (26) Cameroon (27) Serbia (28)
H	Portugal (29) South Korea (30) Uruguay (31) Ghana (32)