Predictions of the Remaining UEFA Affiliated Nations to Qualify for the 2022 FIFA World Cup Supported by Artificial Intelligence Techniques

**IN3062 Introduction to Artificial Intelligence**

Peter Farkas, Xavier Murrow

GitHub:

# Introduction

This project sets out to predict the remaining teams affiliated with UEFA to qualify for the 2022 FIFA World Cup. To create these predictions, machine learning is applied using multiple different classifiers to achieve a spectrum of results and performance related data which are then compared and evaluated. Beyond simply predicting the remaining qualifying teams, this project also looks to answer questions such as: are there differences and/or overlaps among the predictions of varying classifiers? How accurate are the full-time score predictions? How realistic are the overall predictions? And how much do the predictions vary?

# Objective

The Objective of this project is to use data the SPI Ratings dataset by FiveThirtyEight together with match data of all international matches played since the 1st of January 2000, provided by Kaggle.com, to predict the remaining UEFA affiliated nations to qualify for the FIFA 2022 World Cup in Qatar.

# Model

The model used for this project is at the highest level a binary classification model, where the examples are the individual teams and the classes being whether they have qualified or not. However, to ultimately achieve these predictions a few lower-level predictions must be made, i.e., the scores of the individual matches. For example, should Wales be predicted to qualify, they must first win 2 matches, one with a defined opponent and one without. In the first match, Wales plays against Austria. Predicting the outcome of this match causes the model to become multi-label classification model due to two different labels (scores) being predicted with possible values being any non-negative integer.

The individual scores of the home and away team, however, are predicted separately in isolation from each other. This means that the model used for predicting the individual values is actually and multi-class classification model, where the classes are any non-negative integer.

Should Wales win against Austria they shall win and the winner of the parallel match in the same group, the same lower-level model will be applied to the new match. The result of this final match will determine the result for the highest-level model, meaning who qualifies.

# Dataset

To achieve the desired tasks, we were unable to find a singular ideal dataset and were hence required to merge multiple different ones together. The first step was to cut down Mart Jürisoo’s dataset containing all international match results ever played since 1872 to only include results recent enough to validly determine the strength of a nation’s team. The starting date chosen was the 1st of January 2000. This dataset strictly includes men’s full international teams and hence does not include e.g. Olympic teams and U-23 teams.

The next step was to merge this dataset with FiveThirtyEight’s SPI Global International Rankings dataset. This second dataset provided offensive and defensive values representing the number of expected goals scored or respectively conceded against an average of all teams and a “SPI-value” representing the strength of each team. The merging of these two datasets allows us to include the previously mentioned values of each team added to each of their matches and hence predict scores according to them.

The dataset containing the yet to be played UEFA qualifying playoff matches was created by us using FIFA.com as a source. The structure of this dataset is identical to that of the merged dataset used to train and test the algorithm.

# Processing Data

The first thing we did with our chosen datasets was merge them, this is mentioned above. To do this we used the merge function within the Python Pandas library and excel editing tools. This allowed us to create the dataset for our training and testing.

We split our data into training and test data. The test data had a size of 100 data points, and the training data a size of 18,650. The reason behind our test data being so small is due to computational limitations. As training the classifiers was already taking very long, we where not willing to compromise on the accuracy.

The classifiers we used where Support Vector Machine (SVM), Logistic Regression (LR), Decision Trees, and K-Means. The first 3 are supervised learning techniques and K-Means is unsupervised. Whilst testing the different classifiers we found that classifiers using supervised learning produce the more accurate scores. That said, we did want to see how a unsupervised classifier would fair in our model and chose K-Means as it has a predict function within the plugin for it in sklearn. These classifiers are given the training data to learn how to predict the results, then are given the required data of the teams facing each other and creates the prediction; this is done with the test data.

There results of the predictions for all the data is displaced in many ways. The training accuracy and test accuracy is recorded, confusion matrices for the training and the test results, then ending results for preforming each prediction 1000 times. This is done to create much more reliable data for us to discuss in this report.

# Results

## Chart, waterfall chart Description automatically generatedChart, waterfall chart Description automatically generatedChart, scatter chart Description automatically generatedGoal Scoring

Figure 1 Confusion Matrices of Home Score Predictions

Chart, scatter chart, waterfall chart

Description automatically generated

The plots above show the confusion matrices of the initial training of the classifiers (in a usual execution of the code) used to predict the home scores, whilst the ones below are of the classifiers used for away scores.

Chart, waterfall chart

Description automatically generatedChart, scatter chart, waterfall chart

Description automatically generatedChart, scatter chart, waterfall chart

Description automatically generatedChart, waterfall chart

Description automatically generated

Figure 2 Confusion Matrices of Away Score Predictions

There are a few notable differences in the predictions of the home and away classifiers. The values predicted for the number of goals score by the home team have higher extremes (maximum and minimum values), e.g. the decision tree classifier predicting a maximum value of 8 for home goals whilst only predicting 6 for away goals. This discrepancy can be explained by the argument of home field advantage, meaning teams that play in front of a home crowd will have a louder cheering audience and this can affect their performance.

Whilst the discrepancy of goals scored home and away can be explained by external factors affecting the statistical performance of teams, the varying f1-scores of the classifiers cannot.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | F1-Score | | | |
| **Home** | | **Away** | |
| **Train** | **Test** | **Train** | **Test** |
| Logistic Regression | 0.3480 | 0.3700 | 0.4262 | 0.4100 |
| Support Vector Machine | 0.3455 | 0.3300 | 0.4306 | 0.4400 |
| Decision Tree | 0.6323 | 0.2500 | 0.6783 | 0.3700 |
| K-Means | 0.2947 | 0.2900 | 0.3676 | 0.2900 |

Figure Instance of F1-Scores for all Classifiers

(The data included in the table above are taken from a random execution of the code.)

The data in the table above shows a clear increase in the accuracy of the predictions of the individual classifiers when predicting the number of away goals over the number of home goals. Even though the table above only captures data from a singular instance of training and testing the classifiers, the pattern is consistent through all instances. The only value that does not increase is the testing f1-score when using K-Means, this is due to our K-Means being set to only use two clusters.

Whilst we are uncertain about the cause occurrence of this discrepancy, we speculate it may be due to the smaller difference of the extreme values of away goals scored.

## Text Description automatically generatedText Description automatically generatedText Description automatically generatedText Description automatically generatedMatch Results

Figure 4 Instance of Match Results and Qualification

The above figure showes a randomly selected qualification prediction and its results. The decision tree classifier (DTC in Figure 4) will produce a variety of results that do not show a clear pattern. It will predict some believable results like a 2:0 win for the Czech Republic over Poland, whilst also predicting Turkey, the nation ranked 44th in the world, to win 7:1 against 6th ranked Portugal.

Using Logistic Regression (“LG” in Figure 4) will cause the predictions to have very high scoring games, e.g. a 22:22 draw between Wales and Austria, whilst using K-Means and Support Vector Machines (SVC in Figure 4) will predict very low scoring games, with scores rarely exceeding a total of 2 goals scored in the case of SVCs. Football is in general a low-scoring sport and hence such predictions are resonable as long as teams are of similar in skill level.

Our K-Means classifiers are set to only use two clusters. We decided to do this because we were less interested in how accurately it could predict the resulting score of a match, but rather how an unsupervised learning classifier would compare to supervised learning classifiers.

One problem with our approach that will lead to concern and misleading data is the handling of tiebreakers. In football, a draw between two teams after the maximum playable time is handled by a penalty shootout. In our approach the victory is handed to the away side. There is no specific reason as to why we do this, but there are reasons as two why we do not have an elaborate way of handling such situations.

First, we do not possess any data on the performances of international teams in penalty shootouts and are hence not able to reasonably predict the outcomes. Secondly, awarding the victory randomly was considered, however, it was also decided against as penalty shootouts are not random and do heavily depend on many different factors.

## Qualifying Teams

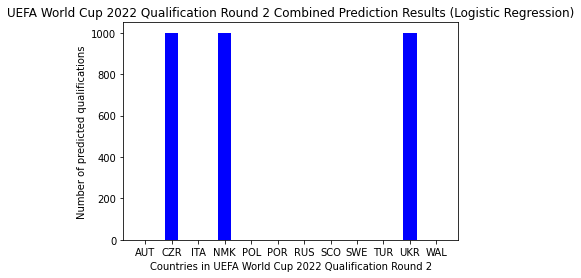
The objective of this project was to predict the remaining teams to qualify. We did this by using multiple classifiers and seeing how their predictions behave. Some classifiers gave predictions that were consistent, whilst others would give a different result with each prediction. To capture these inconsistencies, we ran our code 1000 times, each time retraining all classifiers to avoid any repeated results.

Figure 5 Results of 1000 Predictions Using Logistic Regression

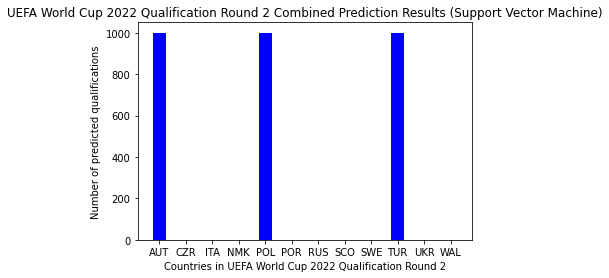
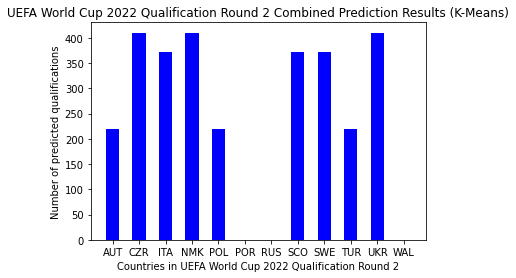
The two most consistent classifiers when predicting the teams that will qualify were Logistic Regression and Support Vector Machine. In the case of Logistic Regression, it predicted the Czech Republic, North Macedonia, and Ukraine to qualify every single time. In comparison, the Support Vector Machine predicted Austria, Poland, and Turkey to qualify all 1000 times. This consistency in prediction is one could argue is not ideal. Even though it gives you a clear result, it does not account for any divergences. Predicting a sport where the result is ultimately only affected by the players on the field and weaker teams may perform better on the day even though statistically this makes no sense, e.g. Sheriff Tiraspol’s 2:1 victory of Spanish giants Real Madrid on the 28th of September 2021

Figure 6 Results of 1000 Predictions Using Support Vector Machine

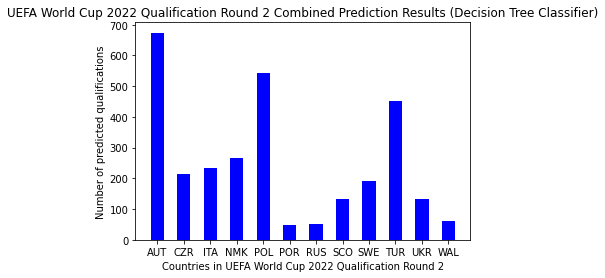
Using both K-Means and decision trees as classifiers lead to better results. K-means has its preferred teams and gives a rather diverse result, however, according to it Portugal, Russia and Wales have absolutely no chance of qualifying.

Figure 7 Results of 1000 Predictions Using Decision Tree Classifier

Using decision trees leads to the most reasonable results, as all teams have qualified at least once. Its clear favorites are Austria (~67% chance of qualification), Poland (~55% chance of qualification), and Turkey (~45% chance of qualification).

Figure 8 Results of 1000 Predictions Using Decision Tree Classifier

It is also interesting to look at a summary of these results as this shows the overlaps between the different predictions. Overall, the most predicted teams are Austria, Czech Republic, North Macedonia, Poland, Turkey, and Ukraine.

These predictions are by far not the most accurate they could be. This is instantly recognisable by the highest ranked teams Portugal and Italy not being among the most predicted teams overall. Two major factors play into this. Firstly, they are both home sides in their first matches giving them a disadvantage in our model, due to them automatically loosing should they not outscore their opponents. Secondly, they have very low defensive and rank values, meaning they are very good teams, however, we speculate that this may not be recognised when training the classifiers.

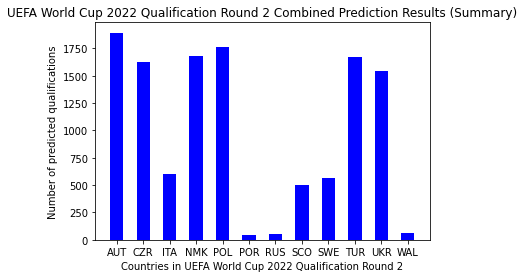


Figure 9 Summary of 1000 Predictions

# Conclusion

Whilst the prediction result of this project may not be the most realistic and require much improving, they do set a strong foundation for further development and a lot of insight in the behaviours of the classifiers used. The first aspect that requires solving and would greatly increase the plausibility of the predictions is the handling of tiebreakers. Data on penalty shootouts of international teams would need to be found and used to train an additional set of classifiers which predict a winner in the case of a tie.

A second aspect that could help improve the predictions is the way ranking and defensive values are handled. A potential solution for the ranking of the teams is to represent the ranking values in ascending order, where the lowest ranked team has a value of 1 and the highest ranked a value of 219.

Another solution to this could be by represent these values as fractional values. One could divide the ranking value by the number of nations and then subtract it from 1. This would then align with the increase in values of the “SPI-ratings” yet remain small values. The same equation could be applied to the defensive values of the teams, but rather than dividing by the number of teams, one would divide by the highest available defensive value. This would allow these values to be represented as a value in a range between 0 and 1, where teams closer to a value of 1 would be defensively better.

# References

llSourcell (2017), *Predicting the Winning Football Team.* Available at: <https://github.com/llSourcell/Predicting_Winning_Teams/blob/master/Prediction.ipynb> (Accessed: 20 December 2021)

jayb (2020), *SPI Ratings*. Available at: <https://github.com/fivethirtyeight/data/tree/master/soccer-spi> (Accessed: 28 November 2021)

Jürisoo, M. (2021), *International football results from 1872 to 2021.* Available at: <https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017> (Accessed: 18 December 2021)

Lamb, R. (2019), *Using machine learning to predict sport scores – a Rugby World Cup example.* Available at: <https://towardsdatascience.com/using-machine-learning-to-predict-sport-scores-a-rugby-world-cup-example-f699fd552673> (Accessed: 20 December 2021)

FIFA (2021), *FIFA World Cup Qatar 2022™.* Available at: <https://www.fifa.com/tournaments/mens/worldcup/qatar2022/qualifiers> (Accessed: 21 December 2021)