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

# Introduction

# 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.

# 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 1972 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.

# 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.