PROJECT DELIVERABLE 2

Enrique Aponte
Seif Abdelkefi

1. **Problem Statement**

   Input: two international teams A and B, each one having a set of corresponding score features (fifa ranking, goal_keeper_score, mean defense score, mean_midfield_score, mean_offense_score)
   Output: classify team A as a winner or loser.

2. **Data Preprocessing**

   https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022
   Same data set specified in the first deliverable.
   The original data set contains 23921 data row. Each row corresponds to a game played between two international teams and contains some of the informations about that game such as teh date in which the game was played, the number of goals scored by each team, the city and the country where the game was played, the fifa ranking of each team, the fifa scores of each of the lines of each team.
   The total number of these columns is 25.
   However, not all the information is needed for our model to work.
   First, we decided to keep only 10 columns to train with: fifa_ranking_A, fifa_ranking_B, goal_keeper_score_A, goal_keeper_score_B, mean_defense_score_A, mean_defense_score_B, mean_midfield_score_A, mean_midfield_score_B, mean_offense_score_A, mean _offense_score_B.
   Then we removed all the rows where there is at least one information missing, ending up with 4301 rows.

3. **Machine learning Model: training and testing performance**

   We changed the model from convolutional neural network to Support Vector Machines, because the training time of the SVM model is faster and is almost as precise as neural network models.
   To implement our model, we used the library sklearn. We parameterized the training phase by setting the kernel to linear, which is going to increase the dimensions of the features in order to find a better hyperplane.
   We divided that set into three different sets, the most recent 5% of rows (based on the date of the game) are the testing set (216 rows), the following 5% corresponds to the validation test (215 rows), and the rest of the 90% is for training (3871 rows).

   We trained the model and tested the accuracy of the prediction using the validation set. We got an accuracy of 0.52. We then started to play with the parameters and figured out that by removing the draw label and having only a binary classification, we increased the accuracy to 0.76!

Obviously, this decreased the number of the total data points: 2891 data points for the training set, 161 for the validation set and 161 for the test set.

The accuracy of the model applied on the test set is around 0.72. The decrease in accuracy is explained by the time gap between the games used to train the model and the games in the test models. There are 161 games that have been played in between, and logically the most recent games are a better indicator of future games. Therefore, by including those 161 games of the validation set in the training we would obtain a more accurate model.