CHAPTER 1
Introduction

The aim of the following report is to describe the study carried out on the application of Machine Learning methods for the prediction of football match results based on the past performance and statistics of the involved teams.

Large amounts of data have been collected during matches by TV channels, football clubs and betting agencies in the last few years. This enables the detailed analysis of the players and teams performance, and therefore the creation of predictive models of their future performance, and consequently the results of the matches. Nowadays there are still few research studies on the matter, but increasingly more researchers are deciding to apply their knowledge in the domain of sport analytic.

This document describes the complete process that has been followed, as well as the different techniques that have been used to create a new database, process the data, perform feature engineering, build a baseline model trained with bookmaker odds, and finally build and train some predictive models based on Artificial Neural Networks and Gradient Boosting techniques. All this, after a revision and critique of the previous relevant related literature.

In order to reach this objective, an iterative spiral methodology has been followed. Starting by proposing simple and basic solutions, and after a subsequent rigorous analysis on the results the experiments, more complex solutions are proposed gradually. Then, these new experiments are evaluated and new solutions are proposed again, bringing the study towards the optimal solution. In this case, we have started training Multilayer Perceptron networks of few layers with an initial dataset, and as experiments have been carried out, new models based on more complex ANNs have been gradually implemented, regarding both their architecture and their hyper-parameters, while the data have been enhanced using feature engineering techniques.

1.1 Motivations

The reasons that led me to conduct this research have been several. First, my passion for football motivated me to consider doing this work. Since I was a child I have been passionate about football and even more about the statistics and data of teams and players.

Secondly, from a professional point of view, it was a great opportunity for me to introduce myself into research and Data Science, the field that I want to dedicate to from

now on. Also the fact that the scope of Sport Analytics is still little studied motivated me. Nevertheless, the world of football is gradually incorporating Big Data and Artificial Intelligence. For this reason, I also think that in the future it could be a professional career path in which I could work because of my two passions: football and Data Science.

As I said, it was a great opportunity to introduce me to Data Science in an area that has not been studied much, so I was also encouraged to see it as a challenge. I was conscious of the complexity of the problem, since finding a correlation between the events of a football match and its outcome is no trivial problem given its high component of randomness and the data limitations that I had.

1.2 Objectives

It was difficult to set an initial objective due to the extremely wide-open nature of the study. From the beginning, however, the target was to design and train a model that could predict whether the home team would win, the away team would win or draw. Moreover, the objective was that the model could outperform the trained model with the odds made by the bookmakers. The odds are taken as a reference to be able to evaluate the performance of the model.

Another goal has been to approach or beat the results of past similar work. This objective is hard to achieve as each work uses a different data source and therefore different data, so it is not straightforward to compare. Nonetheless, we have trained some of the models used in previous works with our data in order to compare and improve our model.

Putting the quantitative results aside, it is also a challenge to find out what are the relationships that exist between the events, strategies and statistics that take place in a football match regarding the match outcome. And thus, attempting to model these interactions in order to understand the factors that decide the destiny of football.

As for personal goals, it also includes gaining experience and skills in data processing and Machine Learning technologies employed, as well as consolidating theoretical knowledge learned on statistics and Machine Learning over the last six months.

1.3 Methodology

The methodology used to carry out the work has been the following:

1. Review of previous related work.
    (a) Notation of the designed models and results obtained and/or conclusions reached in each work.
    (b) Write notes on the highlights of each paper, to be possibly used later in this document.
    (c) Brainstorm and design a possible predictive model architecture based on previous literature.
2. Collection and generation of datasets.
    (a) Search and gathering free and open soccer databases available on the Internet.

   (b) Inspection and analysis of the data content of each database so as to identify which ones are compatible among each other, provide a reasonable amount of data and can be used to train our model.
   (c) Creation of the bookmaker odds dataset from a database that contains the betting odds of the matches.
   (d) Creation of datasets from multiple compatible databases, merging datasets and generating new variables from existing ones. This step consists of a long-term work of programming, creativity and exploratory analysis of data.
3. Training of the baseline model trained with the bookmaker odds.
4. Data analysis
   (a) In-depth statistical analysis of the dataset.
   (b) Dimensional reduction and feature selection based on the data analysis.
5. Iterative process of experiments approaching progressively to the optimal model.
   (a) Study of the solution proposed for the problem, carrying out a research if necessary. In the first iteration, starting with a simple solution.
   (b) Implementation of the solution.
   (c) Running experiments with the model, data and configuration proposed.
   (d) Rigorous analysis of the experiment results, focusing on the results and the performance of the model, especially analysing its learning metrics.
   (e) Comparing the results with the previous experiments.
   (f) Summing conclusions and brainstorming possible enhancements of the model and the data. Back to a) if not possible improvement considered.
6. Final analysis of optimal model.
   (a) Detailed analysis of final results and comparison with baseline models and related works results.
7. Simulation of a whole season in order to compare the simulated league table with the actual one.

1.4 Thesis structure

????? ????????????? ????????????? ????????????? ????????????? ?????????????