Chapter 1

Introduction

1.1   Motivation

Data analysis and machine learning techniques have experienced huge growth, finding applications in very diverse fields, such as medicine, economy, industry, or sports. In particular, professional football has become a domain of greater interest for the application of predictive models, due to the huge quantity of data generated in each competition and due to the economic and media impact associated with this sport.

Competitions such as LaLiga generate detailed information on results, match statistics, team performance, and seasonal trends. This data is used by sports clubs, analysts, media outlets, and specialized platforms to gain insights that enable better decision-making, evaluate team performance, and anticipate possible future results.

Within the field of sports prediction, one of the most studied problems is estimating the outcome of a football match. Traditionally, many studies focus on predicting the result of the match (home win, draw, or away win), as this is a simpler formulation of the problem and allows for models with acceptable performance. However, predicting the exact result, i.e., the final score of the match, is a significantly more complex challenge due to the unpredictable nature of football and the influence of numerous factors that are difficult to model.

The difficulty of this problem lies in aspects such as the low frequency of certain markers, the inherent variability in team performance, the influence of chance, and the occurrence of unexpected events during matches. Despite this, predicting the exact result is of great interest from an analytical point of view, as it provides richer and more detailed information than simply predicting the outcome of the match.

Furthermore, studying this problem is of clear academic interest, as it allows us to analyze the extent to which historical data can capture relevant patterns in a sport characterized by a high degree of uncertainty. In this context, this Final Degree Project is an opportunity to explore the capabilities and limitations of predictive models applied to a real, complex, and widely studied problem.

1.2 Disciplinary context and problem statement

The use of analysis techniques in football has been addressed from multiple perspectives in scientific literature. There are studies focused on predicting match results, analyzing player performance, detecting tactical patterns, or evaluating the probability of certain events during a match[1]

When it comes to predicting results, the most common approaches include classical statistical methods, such as Poisson models for estimating the number of goals [2], as well as more advanced machine learning and deep learning techniques [3]. However, most of these studies simplify the problem by focusing on predicting the winner of the match or the total number of goals, leaving the predictions of the exact score in the background.

Predicting the exact result presents additional challenges, such as class imbalance, the scarcity of examples for certain scorelines, and the need to adequately model the relationship between the goals scored by both teams. These difficulties have been highlighted in various studies as one of the main challenges in predicting football scorelines[4]

In this context, this paper focuses on addressing the problem of predicting the exact outcome of LaLiga matches, using historical data available before each match is played. The objective is not only to obtain a model with good predictive performance, but also to analyze the difficulties inherent in the problem and assess the extent to which it is possible to extract relevant information from the available data.

This approach allows the work to be placed within the field of predictive analysis applied to sport, addressing a challenging and relevant problem in professional soccer.

Chapter 2

Objectives

2.1 General Objective

The main objective of this Final Degree Project is to develop and evaluate a predictive system capable of estimating the exact final score of football matches in the Spanish professional league. The system is based on historical match data available prior to each encounter and applies data analysis and machine learning techniques to address this prediction task.

Beyond obtaining predictive results, this project aims to analyze the feasibility of predicting exact match scores in football and to assess the limitations inherent to this type of problem, characterized by high uncertainty and variability.

2.2 Specific Objectives

In order to achieve the general objective described above, the following specific objectives are defined:

1. To collect and analyze historical data from football matches in LaLiga, identifying the most relevant variables for score prediction.

2. To derive new information variables from the collected data through a feature engineering process inspired by Knoledge Discovery in Databases (KDD) methodologies.

3. To design, implement and evaluate a predictive model for exact football match score prediction.

4. To analyze the results obtained, identifying strengths and weaknesses.

2.3   Scope and Limitations

The scope of this project is limited to the analysis of football matches corresponding to the Spanish professional league. Only historical data available before the start of each match are considered, ensuring that the predictive system does not rely on information that would not be known at prediction time. The historical data goes from the 2005-2006 season to the 2024-2025 season.

Several factors that may influence match outcomes are intentionally excluded from the analysis due to their unpredictable nature or lack of reliable data. These include last-minute player injuries, referee decisions, extreme weather conditions and unexpected in-game events.

Additionally, the project does not aim to develop a real-time prediction system or a commercial applications. The focus is placed on the academic analysis of the problem and on evaluating the capabilities and limitations of predictive models when applied to exact score predictions in football.

Bibliography

[1] R. Rein and D. Memmert. A survey of analytics in soccer. *Sports Medicine*, 46(5):1–17, 2016.

[2] M. J. Dixon and S. G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C*, 46(2): 265–280, 1997.

[3] D. Berrar, P. Lopes, and W. Dubitzky. Predicting the outcome of football matches using machine learning techniques. *Machine Learning*, 108(1):165–189, 2019.

[4] D. Karlis and I. Ntzoufras. Score prediction in soccer using probabilistic models. *Journal of the Royal Statistical Society: Series C*, 52(3):381–397, 2003.