**LaLiga Data Analysis**

*Data Science Project Report*

**Submitted To**: Department of Sipalaya InfoTech



**Prepared By:** Susan Dhungana

**Date:** 2081/09/27

# **Abstract**

This research offers an elaborate study about the team performances in La Liga for the season 2024-2025, scraped from ESPN. This study is concerned with many aspects of data science, from data preprocessing, exploratory data analysis, visualization, and predictive modeling to evaluating and predicting team standings. By using Python's powerful libraries, such as pandas, matplotlib, seaborn, and sklearn, this project brings forward critical trends and insights into the competitive terrain of the league.

The main objectives are to forecast the overall number of points each of the 20 teams will achieve after 38 matches, find out who might win the league and identify teams that could be facing relegation this season. Another point of analysis will look at how an indicator such as goals scored, conceded, or a win or a lose match is related to the overall performance index of the team. Thus, this project offers better understanding for the current season, and most importantly demonstrates data-driven techniques whereby sports will never be the same again, with predictive analytics benefits to analysts, fans, and stakeholders.

# **Overview**

As one of the most competitively fierce football leagues in the world, La Liga has the great teams attacking each other for supremacy; at the same time, with Python for data analysis this project analysis reflects the dynamics of the league in terms of team performance and predictions concerning the standings of a team by the end of the season through analysis that focused on team rankings, goals, points, and the relationship of these with performance in the overall game.

# **Introduction**

Analyzing the football competition performances of La Liga, Spain's elite football competition, has implications for the strengths and weaknesses of teams in the league. It broadens the understanding of the league's configuration. Using Python for data extraction, cleaning, visualization, and then predictive modeling to analyze and predict La Liga standings forms the core of the project.

# **Tools and Libraries**

The following Python libraries are used in this project:

* Pandas: For data manipulation and cleaning.
* Matplotlib and Seaborn: For data visualization.
* Scikit-Learn: For machine learning model training and evaluation.

# **Data Source**

The data is extracted from ESPN’s La Liga standings table using the pandas.read\_html function. The extracted dataset consists of following columns:

* **Rank**: The current position of the team in the league.
* **Teams**: Name of the football teams.
* **GP**: Games Played.
* **W**: Number of Wins.
* **D**: Number of Draws.
* **L**: Number of Losses.
* **F**: Goals Scored (For).
* **A**: Goals Conceded (Against).
* **GD**: Goal Difference.
* **P**: Points Accumulated.

# **Methodology**

## **Data Collection**

The dataset derives information from the La Liga table from ESPN that processed itself into a clean data set, and "Teams" column was further processed to delete unnecessary prefixes.

## **Data Cleaning**