**LaLiga Data Analysis**

*Data Science Project Report*

**Submitted To**: Department of Sipalaya InfoTech



**Prepared By:** Susan Dhungana

**Date:** 2081/09/22

Table of Contents

[**Abstract** 3](#_Toc187063758)

[**Overview** 4](#_Toc187063759)

[**Introduction** 4](#_Toc187063760)

[**Tools and Libraries** 4](#_Toc187063761)

[**Data Source** 4](#_Toc187063762)

[**Methodology** 5](#_Toc187063763)

[**1.** **Data Collection** 5](#_Toc187063764)

[**2.** **Data Cleaning** 5](#_Toc187063765)

[**3.** **Data Visualization** 5](#_Toc187063766)

[**4.** **Machine Learning Model** 6](#_Toc187063767)

[**5.** **Prediction of Final Points** 7](#_Toc187063768)

[**Results** 7](#_Toc187063769)

[**1.** **Data Insights** 7](#_Toc187063770)

[**2.** **Predictions** 7](#_Toc187063771)

[**Code Functionality** 8](#_Toc187063772)

[**1.** **Functions** 8](#_Toc187063773)

[**2.** **Outputs** 8](#_Toc187063774)

[**Conclusion** 9](#_Toc187063775)

[**Future Enhancements** 9](#_Toc187063776)

[**References** 9](#_Toc187063777)

# **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.

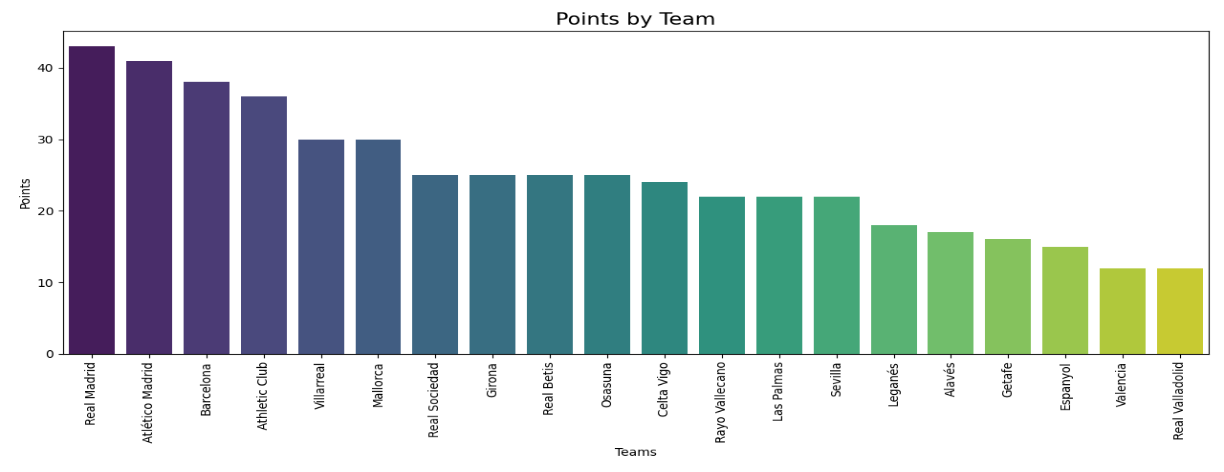
## **Data Cleaning**

* Duplicate Removal: Duplicate records are removed.
* Handling Missing Values: Any missing values in the dataset are filled with 0 ( if exists ).
* Data Type Conversion: Columns such as GP, W, D, L, F, A, GD, and P are converted to numeric types for analysis.

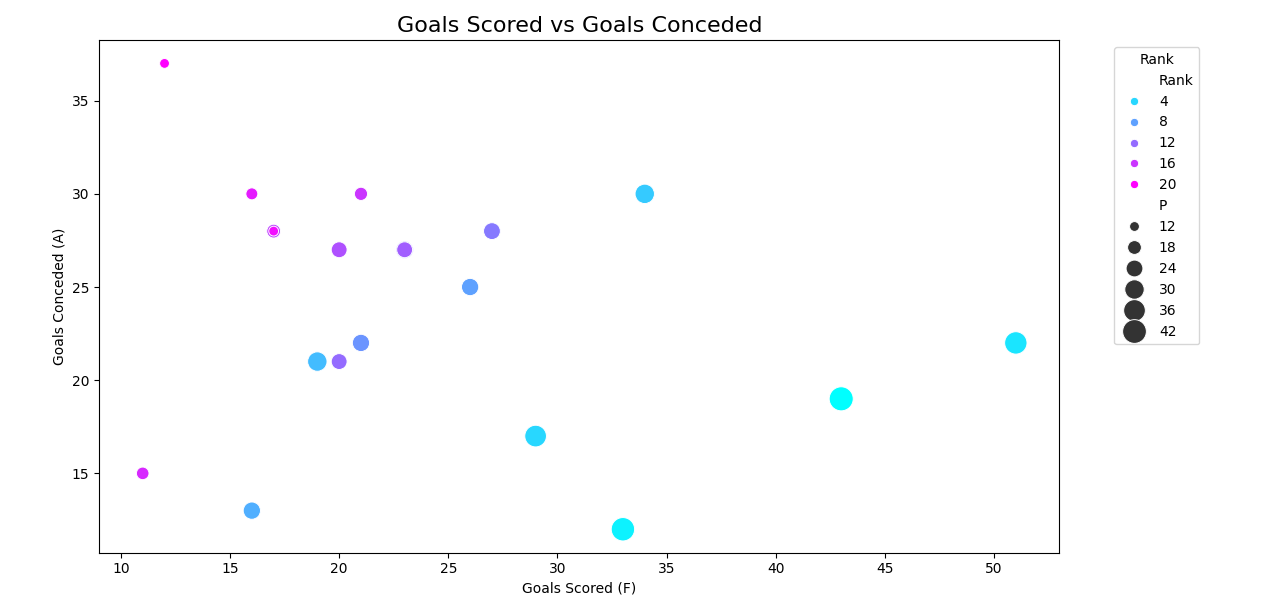
## **Data Visualization**

The following visualizations help us study team performance:

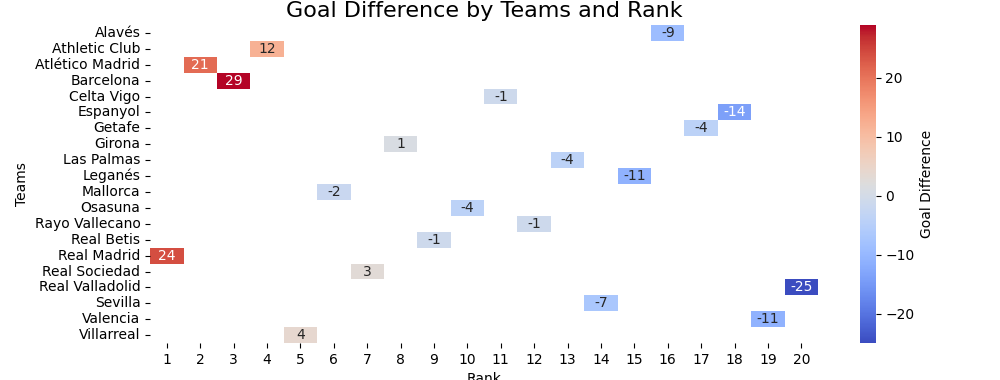
* Bar Chart: Total points each team has accumulated so far.



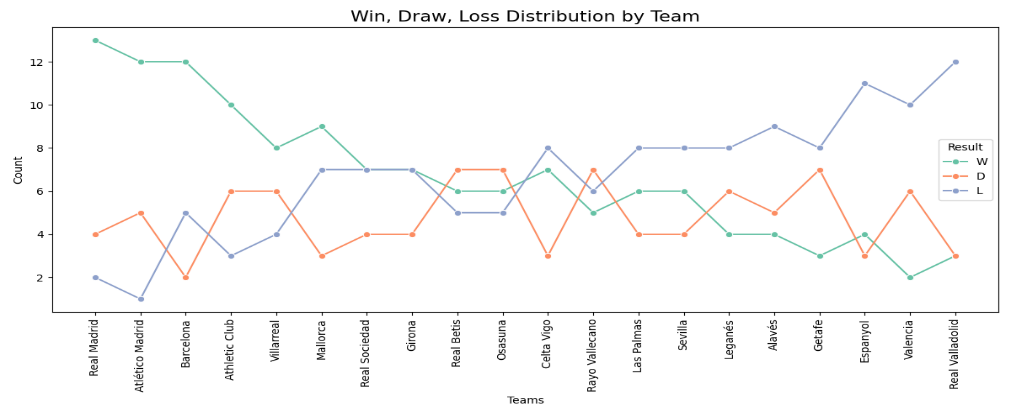
* Scatter Plot: Goals scored vs goals conceded.



* Heatmap: Differences in the number of goals by teams and their rank.



* Line Plot: Distribution of Wins, Draws and Losses by Team.



## **Machine Learning Model**

A linear regression model is created to predict total points earned by teams relative to:

* Wins (W), Draws (D), Losses (L)
* Goals Scored (F), Goals Conceded (A)
* Goal Difference (GD)

**Steps:**

* Data is split into training and testing sets using an 80-20 split.
* The model is trained using LinearRegression from Scikit-learn.
* Predictions are made on the test data, and the Mean Squared Error (MSE) is calculated.

## **Prediction of Final Points**

It is designed to make predictions about the total points of a team at the end of 38 games. For this purpose, a function named predict\_points\_for\_row has been created. It takes the average corresponding performance metrics and scales them to estimate the expected final points.

# **Results**

## **Data Insights**

* Top Performing Teams: Severally analyzed in terms of aggregate points and goal scored.
* Weak Teams: Poor teams are identified in terms of their performance regarding the number of losses and goals conceded.

## **Predictions**

* Winner: The team predicted to have the highest points at the end of the season. Along with this, it predicts the final points for a team based on their current statistics.



* Relegated Teams: Bottom three teams are relegated based on their predicted points.

A black screen with white text

Description automatically generated

# **Code Functionality**

## **Functions**

* Predict\_points\_for\_row: Predicts the final points for a team based on their current statistics.
* Data Visualization: Several plots are generated to aid in understanding team performance.

## **Outputs**

* Cleaned and processed dataset saved as laliga\_table.csv.
* Predicted final points for each team added to the dataset.

# **Conclusion**

The project aptly demonstrates the capabilities to extract, clean and analyze football data and produces insights into team performance. By visualizing and leveraging machine learning models, this analysis identifies probable winners and relegated teams for the season. It also scales up the use of predictive modeling and advanced techniques for team statistics. This study highlights the potential of data-driven approaches in sports analytics, enabling stakeholders to make informed decisions based on current performance metrics.

# **Future Enhancements**

The further improvements recommended to this project are as follows:

* High-end modeling: Integrating Ridge or Lasso regression for overfitting and reliability in prediction.
* Feature expansion: Dynamic features would include recent form from the team comprising of player statistics, and performance at home/away, to increase the robustness of prediction.
* Realtime: Automating the scraping of data and processing pipelines would live continuous monitoring of team performances.
* Interactive Visuals: Using tools like Plotly or Dash would give rise to dynamic and user-friendly dashboards.
* Scenario Simulation: development of module for simulation of various scenarios, such as high-impact key player injuries or assumed results from the match which would lead to strategic insights.

## **References**

* ESPN Staff. "La Liga Standings Table." *ESPN*, 2025. Retrieved from <https://www.espn.in/football/table/_/league/esp.1>.
* McKinney, W. (2010). "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, 51-56. Retrieved from [https://pandas.pydata.org](https://pandas.pydata.org/)
* Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, 9(3), 90-95. Retrieved from <https://matplotlib.org>.
* Waskom, M. et al. (2021). "Seaborn: Statistical Data Visualization." *Journal of Open-Source Software*, 6(60), 3021. Retrieved from <https://seaborn.pydata.org>.
* Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from <https://scikit-learn.org>.
* Python Software Foundation. (2023). "Python Language Reference, version 3.10." Retrieved from <https://www.python.org>.
* Plotly Team. (2023). "Interactive Data Visualization with Plotly." Retrieved from <https://plotly.com>.
* Dash Developers. (2023). "Dash: A Python Framework for Building Web Applications." Retrieved from <https://dash.plotly.com-=o0i> 76r5e