Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

# Stacked Ensemble-Based Framework for Predicting Market and Tactical Fit in Football Transfers

**Submitted in partial fulfillment of the requirements of the degree of**

## Bachelor of Technology

**In**

## Artificial Intelligence (AI) and Data Science

**By,**

**Vedant Bhawnani** **60018210069**
**Shubh Harde** **60018220135**

**Under the guidance of**
**Prof. Shruti Dodani**
**Dwarkadas J. Sanghvi College of Engineering**

**UNIVERSITY OF MUMBAI**

**ARTIFICIAL INTELLIGENCE (AI) AND DATA SCIENCE DEPARTMENT**

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

# CERTIFICATE

This is to certify that, the project entitled "**Stacked Ensemble-Based Framework for Predicting Market Value and Tactical Fit in Football Transfers**" is a bonafide work of Vedant Bhawnani(60018210069) and Shubh Harde(60018220135) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.Tech. in Artificial Intelligence (AI) and Data Science.

Prof. Shruti Dodani

**Internal Guide** **External Examiner**

Dr. Pratik Kanani Dr. Hari Vasudevan

**Head of Department** **Principal**

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

# Project Report Approval

This project report entitled "**Stacked Ensemble-Based Framework for Predicting Market Value and Tactical Fit in Football Transfers**" by Vedant Bhawnani, Shubh Harde is approved for the degree of B.Tech. in Artificial Intelligence (AI) and Data Science Engineering.

**Examiners**

**1) ……..**

**2) ……..**

**Date:**

**Place:**

# Declaration

We declare that, this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that, We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1) Shubh Harde – 60018220135

2) Vedant Bhawnani - 60018210069

**Date:**

**Abstract**

The transfer markets in football are a dynamic and pivotal period which entails a state of volatility where teams and clubs strategize and aim to strengthen their teams. This is more often than not preceded by in-depth analysis of data related to hundreds of promising and prospective players to find the perfect replacement for the outgoing player. This analysis currently includes labour-intensive tasks of manual video reviews and scouts going to games to watch a prospect play. This research introduces the Football Player Replacement Finder, a novel approach to reduce the complexity and time required for scouting and acquiring impactful talents by using advanced machine learning models and automated data scraping pipelines. Our system employs supervised models for gauging the performance and price of football players along with clustering techniques for player profiling, enabling stat-by-stat comparison of players. By integrating advanced metrics along with appealing visualisations, our system empowers decision-makers to streamline their scouting process and uncover valuable talents effectively.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

# Chapter 1

# Introduction

## 1.1 Description

One of the most important periods for any football team trying to bolster their roster is the transfer window, particularly in January and July. Clubs put a lot of effort into scouting possible additions during these months, while simultaneously monitoring players who might depart. Finding qualified replacements without going over budget is crucial for clubs since maintaining financial stability is a top concern. A thorough examination of a number of variables, including player statistics, market trends (such as inflation and volatility), team dynamics when adding a new player, and even possible conflicts in player personalities, is necessary to scout individuals while making sure they match the budget.

This procedure has historically taken a long time and involved reviewing a lot of video footage, compiling subjective reports, and manually comparing statistics. It's getting harder for contemporary football teams to stay up to date with these outdated techniques as more data becomes available and measurements get more complex.

## 1.2 Problem Formulation

The issue is that many football teams continue to employ antiquated and time-consuming scouting techniques. Despite the abundance of player data and statistics provided by systems such as FBref and Sofifa, the present method of identifying appropriate substitutes still mostly depends on manual comparisons and subjective judgements. These conventional techniques just cannot keep up with the volume and complexity of contemporary player scouting as more data becomes available.

Furthermore, depending solely on subjective reports may result in biases in player evaluations and the omission of important information regarding player performance. Clubs also face the onerous task of identifying replacements who meet the team's tactical requirements and budget, which is hard to handle by hand.

## 1.3 Proposed Solution

This study aims to streamline and modernize the recruitment and replacement process, by incorporating machine learning and automating data analytics. The system can find players who are statistically comparable to the one being replaced and anticipate player performance using machine learning methods, such as supervised machine learning and clustering algorithms. To make sure that the comparison between players is as precise and pertinent as feasible, it makes use of similarity metrics including cosine similarity, Pearson correlation, and Euclidean distance.

Implementing automated data scraping lets us quickly collect player performance data from online sources, including FBref and TransferMarkt. The system focuses specifically on the statistical performance of a player, things like expected goals, assists, and passing accuracy. This process removes the manual efforts and potential biases found in conventional scouting, allowing scouts to focus on other things. It ensures player comparisons are faster, more objective, and grounded purely in performance indicators.

Clubs can use this technique to make better-informed decisions regarding possible acquisitions, guaranteeing that new players meet the team's long-term objectives as well as its immediate tactical requirements.

## 1.4 Scope of the Project

This project aims to deliver a one-stop tool for scouting and team management. Its purpose is to assist in uncovering hidden talent and discovering players suited to a team's style. The system will also use similarity measures on online data to evaluate and find replacements for departing players. Emphasis is placed solely on on-field metrics that align directly with a team's requirements.

The project will not, however, address the broader aspects of football management, such as contract negotiations, off-field player conduct, or psychology. Unless there is a direct impact of an action on a player's performance, such as injury records, the system will not take non-performance characteristics into consideration. In the end, the technology is aimed to offer precise player comparisons and recruiting forecasts, leaving the club's management staff to gauge how well a player fits in their club dynamics.

# Chapter 2

# Review of Literature

The estimation of football players' market value has been an area of extensive research, with scholars utilizing a variety of methods, datasets, and algorithms to improve the how accurately the models can predict these values . This section provides a detailed literature review focusing on the application of machine learning and deep learning techniques for player valuation. Several studies have identified the key factors that most influence the market value. Among these, age emerges as a crucial deciding factor, as younger players are often valued higher due to their developmental potential. Additionally, a player's popularity can elevate their market worth, as it is commonly associated with increased fan engagement and commercial appeal [1]. The player's on-field position is another crutial factor. Attackers, in particular, are generally attributed higher market values [2].

The use of regression models to determine the factors influencing football transfer fees was first introduced by Carmichael and Thomas (1993) [3]. Building on their approach and research, many further studies adopted the regression-model in the sports and analysis domain. Building on this tradition, Mustafa and Al-Asadi utilized FIFA data as a benchmark for actual market values and applied both linear and non-linear models to estimate player prices [4]. Their methodology incorporated nine distinct parameters, including international reputation and the "weak foot" attribute, among others.

Stanojevic and Gyarmati conducted a study utilizing statistical methods, sourcing their data from the sports analytics firm InStat and the publicly available platform Transfermarkt [5]. Adopting a conventional methodology, their research aimed to estimate the market value of 12,858 football players based on various performance indicators. They employed clustering techniques to interpret player performance data and developed a model using 45 predictor variables. The resulting estimates demonstrated better accuracy compared to the widely accepted market values provided by Transfermarkt, highlighting the potential of advanced data analytics and granular performance metrics in enhancing valuation precision [6].

Müller et al. [1] proposed a data-driven methodology to tackle the limitations of crowd-sourced market value estimates. Their study utilized data from the top five European football leagues and constructed a comprehensive dataset incorporating player-specific attributes such as age, position, and nationality. Notably, their approach also integrated supplementary data from external sources, including Wikipedia, Facebook, and Google metrics, reflecting public interest and

online presence. A linear regression model was employed to estimate player market values, and the results aligned closely with existing crowd-sourced valuations, demonstrating the viability of their approach.

Dobson et al. [7] examined the influence of player-specific metrics on transfer fees and observed significant volatility in transfer valuations, even within the same competitive league. Expanding on this line of inquiry, more recent work by Depken II and Globan employed linear regression analysis to demonstrate that English football clubs tend to pay a premium for players in the transfer market compared to their counterparts from other European nations.

Yigit et al. introduced an innovative methodology for estimating player market values by utilizing a broad spectrum of player attributes, including on-field performance metrics, demographic characteristics, and market-related factors. Their dataset encompassed 5,316 players from 11 prominent football leagues in Europe and South America. The study achieved market value predictions close to the actual transfer market values by integrating data from the Football Manager simulation game alongside transfer value data sourced from Transfermarkt.[8].

In a distinct approach, Behravan et al. applied Particle Swarm Optimization (PSO) to predict player market values, using data from the FIFA 20 dataset, where the in-game player value was considered the true market value. They employed an automatic clustering algorithm to categorize players into four clusters based on their positions. Their method demonstrated superior performance, with RMSE and MAE values of 2,819,286 and 711,029,413, respectively, compared to the results obtained by Müller et al. [6], which had RMSE and MAE values of 5,793,474 and 3,241,733. These results underscore the effectiveness of their approach in predicting market values more accurately [9].

Ian et al. [10] used machine learning techniques to estimate football transfer fees, using data from sofifa.com and transfermarkt.com. They trained both linear regression and XGBoost models on a variety of performance metrics, including data from Instat and GIM performance ratings. Their results showed that the XGBoost model outperformed the linear regression model in predicting transfer fees. This study underscores the potential of machine learning to enhance transfer decision-making, specifically in answering the question, "What is the expected transfer fee of a player based on their previous performance?" The authors also propose further research to evaluate the "reasonableness" of transfer fees by considering post-transfer performance, publically available data sources, and potentially applying similar machine learning methods.

# Chapter 3

# System Requirements Specification

## 3.1 Introduction

### 3.1.1 Aim

This section lays out the blueprint for the Football Player Replacement Finder system. Detailed here is what the system needs to do (its functions) and how well it needs to do it (its qualities). The core idea is to give football clubs, their scouts, and analysts a one-stop tool. By utilizing data analytics and machine learning, this approach aims to increase the accuracy and efficiency of finding possible replacement players.

### 3.1.2 What this covers

The Football Player Replacement Finder system is designed to handle several key tasks. It automatically gathers player performance statistics, market information, and player demographics from public websites like FBref and Transfermarkt. The system then combines this data, cleaning up inconsistencies to build a single, consistent player dataset. It generates detailed player profiles, including performance numbers, estimated market value, and personal details. These profiles are enhanced by adding feature-engineered metrics specific to different on-field roles. When a player leaves a club, the system finds and ranks potential replacements by evaluating statistical similarity, playing style, and other criteria set by the user. Finally, the system presents all this information through an easy-to-use interface featuring helpful visualizations, aiding scouts in decision-making.

Our primary focus is on a player's on-field performance and how it aligns with a team's needs. We are not including areas like contract terms, in-depth player psychology, off-field conduct, or comprehensive injury logs in the system. These are outside the system's scope, unless an injury directly impacts the core on-field statistics obtained from our sources. Similarly, the system does not provide real-time match analysis or use live data feeds; it updates its data periodically by checking the defined online sources.

### 3.1.3 Overview

This remainder of this section is structure as follows: Section 3.2 provides the an overview – its main functions, who'll be using it, and significant limitations or dependencies. Then, Section 3.3

details exactly what the system must do, how well it must perform, and how it interacts with the outside world.

## 3.2 System Overview

### 3.2.1 Product Goal

The Football Player Replacement Finder is aimed to be a dedicated tool to aid decision-making within football scouting departments. While it operates as a self-contained system, it depends heavily on the data drawn from publicly accessible online platforms. In its current form, it's not designed to plug directly into existing club management software, but rather to deliver actionable intelligence that can then be manually woven into a club's operational workflows. Though it stands on the shoulders of established data analytics concepts and machine learning approaches tailored for the football world, it is a fresh product aimed to be an aid to the scouts in the decision-making process.

### 3.2.2 Core Functions

The Football Player Replacement Finder will handle several key responsibilities:

- Automatically fetch data from the defined football statistics websites.
- It's designed to build up thorough player profiles, looking at both raw stats and underlying playing styles.
- A core capability will be its intelligent suggestions for similar players, especially useful when a replacement is needed.
- The system will also contain predictive models to give an estimate of a player's market value.
- All this will be accessible through a user-friendly interface designed for easy data exploration and clear visualization of football analytics.

We'll break these down in much more detail in Section 3.3.1.

### 3.2.3 Target Users

This system is designed while keeping the following roles and people in mind:

- **Football Scouts and Analysts:** These are the frontline professionals tasked with assessing players. We expect them to know football inside out, but their tech-savviness might vary.

The system needs to be approachable even for those not deeply versed in complex data tools.

- **Team Managers/Coaches:** They might use the system to get a clearer picture of player profiles or to double-check scouting insights.
- **Sports Data Aficionados/Researchers (Secondary Audience):** We also foresee individuals using the system for academic pursuits or personal explorations into player data.

Providing only a web interface is intentional since the target audience might not be tech-savvy, hence a web interface with provide the least resistance to use.

### 3.2.4 Operational Boundaries and Limitations

The quality of the system's insights defend on the accuracy and ongoing availability of data. This said, in the future the model should be retrained to take into account the inflation and market trends. The data gathering should be done systematically, following the guidelines of the website being scraped, and pipelines being run on the said data.

The initial goal of the project is focused only on the men's professional football leagues, though the use case of this project can be extended into different sports with relevant and high quality data. The system is built to run on everyday desktops and laptops, as detailed in section 3.3.4.

### 3.2.5 Key Assumptions and Dependencies

We're proceeding with a few assumptions in mind:

- That the public football data sites we rely on (FBref, Transfermarkt, etc.) will continue to be accessible, and their basic structure won't change so drastically as to completely break our data collection methods without warning.
- Users will need a stable internet link for the system to fetch fresh data.

## 3.3 Detailed System Requirements

### 3.3.1 Functional Requirements

Here, we spell out the precise actions and capabilities of the system. Each function gets a unique tag (FR.x) for easy reference.

#### 3.3.1.1 FR.1: Managing the Data Flow

- **FR.1.1 (Fetching Data Automatically):** The system is required to automatically pull player statistics, market information (including past values if gettable), and demographic details from our specified web sources (FBref, Transfermarkt).
- **FR.1.2 (Combining Data Together):** The system must be capable of merging data from these diverse sources. This involves cleaning it up, transforming it as needed, and sorting out any discrepancies (like different ways player names are spelled) to produce one consistent player dataset.
- **FR.1.3 (Maintaining Data Standards):** There needs to be a way for the system to update its player database with the newest information from our sources, either when a user triggers it or on a set schedule.

#### 3.3.1.2 FR.2: Analysing and Profiling Players

- **FR.2.1 (Building Player Sections):** For every player, the system will construct a thorough profile. This section will show key performance indicators (KPIs), data about their position(s), contract details (where available), and basic demographic info.
- **FR.2.2 (Metrics by Position):** The system needs to calculate and show specialized, derived stats that are tailored to different roles on the field (e.g., how many Shots are Saved per 90 minutes for Keepers, or Progressive Passes per 90 for Midfielders).

#### 3.3.1.3 FR.4: Forecasting Market Value

- **FR.4.1 (Model Readiness):** The system will provide a Stacked Ensemble model, trained on the data available during the last training of the model to predict the market value of a player.
- **FR.4.2 (Showing Predictions):** For every player, the system will display its predicted market value. This will sit alongside their actual market value (if we have it from our sources) and, some indication of its typical error margin, using the $R^2$ score.

#### 3.3.1.4 FR.5: Interface and Visuals

- **FR.5.1 (Dynamic Dashboards):** The system will feature interactive dashboards that show player data in an intuitive fashion. This means things like radar charts for comparing player attributes, and showing calculated metrics to deep dive into a player's football psychic.

- **FR.5.2 (League-Level Insights):** Users should be able to pick a league and then see visual summaries of league standings, team stats, and leader boards for top players (like top scorers or assist providers), among other relevant league-wide information.
- **FR.5.3 (Easy Navigation):** The User Interface (UI) has to be straightforward. Users should find it easy to search for players, apply filters and navigate the other features the UI offers.
- **FR.5.4 (System Feedback):** The UI needs to keep the user in the loop, especially when data is being loaded or processed, by being responsive and providing clear feedback.

### 3.3.2 How the System Should Perform: Non-Functional Requirements

This part specifies the quality benchmarks for the system.

#### 3.3.2.1 NFR.1: Speed and Responsiveness

- **NFR.1.1 (Quick Lookups):** Basic player searches and pulling up a player's profile should happen fast.
- **NFR.1.2 (Efficient Data Gathering):** Full cycles of scraping data or updating it should be done in a sensible amount of time(a few hours, this will depend heavily on the network and deployment server constraints) and must be "polite" to the source websites (e.g., not hammering them with too many requests too quickly).

#### 3.3.2.2 NFR.2: Ease of Use

- **NFR.2.1 (Short Learning Curve):** Someone new to the system, but who understands football, should be able to get the hang of core tasks (like finding a replacement or viewing a profile) with very little fuss – possibly with no formal training.
- **NFR.2.2 (Clear Visuals):** All charts, graphs, and any other visual ways data is presented must have clear labels and be easy to understand at a glance.
- **NFR.2.3 (Preventing and Handling Mistakes):** Where possible, the system should stop users from entering invalid data. When mistakes do happen, it needs to provide clear messages that help the user fix the problem.

#### 3.3.2.3 NFR.3: Dependability

- **NFR.3.1 (Resilient Data Scrapers):** The bits of code that grab data from websites should be built to cope with small changes in how those websites are laid out. Big changes might

still need code updates, though.

- **NFR.3.2 (Availability - If Online):** If we deploy this as a web service, we'd aim for it to be up and running 99.5% of the time. (This doesn't apply if it's just a desktop app).
- **NFR.3.3 (Keeping Data Safe):** The system must make sure the data it stores is kept intact, preventing it from getting corrupted or accidentally lost.

#### 3.3.2.4 NFR.4: Maintainability and Scope of Growth

- **NFR.4.1 (Built in Modules):** The system should be put together in a modular way (e.g., data scraping, machine learning models, and the user interface should be somewhat separate parts).
- **NFR.4.2 (Well-Commented Code):** Informational comments should be added to key parts of the code, especially the tricky sections like algorithms and how data is processed, to explain what is going on.
- **NFR.4.3 (Easy to Tweak):** Things like website addresses for data sources, settings for the machine learning models (where it makes sense), and how often data is scraped should be adjustable without having to rewrite the code itself.

#### 3.3.2.5 NFR.5: Accuracy

- **NFR.5.1 (Data Quality):** The data scraped should be a true reflection of what's on the source websites at the moment it was collected.
- **NFR.5.2 (Good Predictions):** The model that predicts market values should hit a certain target for accuracy – for instance, an R-squared value better than 0.85 when tested on data it hasn't seen before, and a minimal difference when comparing the R-squared value between the train and test datasets.

### 3.3.3 Interacting with the Outside World: External Interfaces

#### 3.3.3.1 User Interfaces

The system will feature a Graphical User Interface (GUI). The look and feel, along with key user journeys, are shown in our UI design screenshots (you can find these in Chapter 4, under the User Interface Design section).

#### 3.3.3.2 Connections to Hardware

No specialized or custom hardware connections are needed. The system is designed to run on standard desktop or laptop computers.

#### 3.3.3.3 Connections to Other Software

- For data scraping, the system will connect to external websites (FBref, Transfermarkt, Sofifa) using standard web protocols (HTTP/HTTPS).
- The system will lean on several Python libraries for its heavy lifting: Pandas and NumPy for data wrangling, Scikit-learn and XGBoost for machine learning, and Matplotlib, Seaborn, or Plotly for creating visualizations.

#### 3.3.3.4 Network requirements

An active internet connection is a must for the system to scrape data from online sources. It will use the usual web protocols (HTTP/HTTPS) to communicate between the user's browser and the server where it is deployed.

### 3.3.4 Specific Requirements

We are targeting an audience that may or may not own specialized hardware, and have kept the necessary requirements at a minimum. A everyday computer or laptop with a decent network connection should be able to access websites. To replicate the work denoted here, the minimum requirements are a laptop with atleast 8GB RAM, sufficient hardware space, and a fast network connection. Having a dedicated GPU would be beneficial in training the machine learning model.

## 3.4 Use Case Description

This section breaks down the primary ways users will interact with the system.

### 3.4.1 Target Audience

- **Scout/Analyst:** This is our intended target audience. They'll be interacting with the system to assess players and find suitable replacements.

### 3.4.2 Use Case Descriptions

#### 3.4.2.1 UC.1: Examining a Player's Detailed Profile

**Name:** Review Player Dossier

**Goal in Context:** The Scout/Analyst aims to gain a comprehensive understanding of a specific player by viewing their detailed statistical information, performance metrics, and associated visualizations.

**Preconditions:**

- The system has player data available.
- The Scout/Analyst has selected a specific player for review.

**Trigger:** The Scout/Analyst navigates to a player's profile page or selects a player from a list.

**Main Success Scenario:**

1. The Scout/Analyst selects a player.
2. The system retrieves and displays the player's detailed profile, including:
    - Basic information (age, nationality, position, club).
    - In-depth performance statistics (goals, assists, passing accuracy, defensive actions, etc., relevant to their position).
    - Advanced metrics (xG, xA, progression stats).
    - Visualizations of performance (e.g., radar charts, performance graphs over time).
    - Comparison with similar players or league averages (if applicable).
3. The Scout/Analyst reviews the information.

**Extensions (Alternative Flows):**

- **2a. Player Data Incomplete:** If some specific metrics are unavailable for the player, the system displays available data and indicates missing information.
- **2b. Comparison Requested:** The Scout/Analyst initiates a direct comparison with another player from the profile view.

**Postconditions:**

- The Scout/Analyst has a detailed understanding of the player's capabilities and performance profile.
- The Scout/Analyst may add the player to a shortlist or take notes.

#### 3.4.2.2 UC.2: Assessing League-Wide Data and Trends

**Name:** Analyze League Landscape

**Goal in Context:** The Scout/Analyst wants to review aggregated statistics, current standings, and identify top-performing players within a chosen football league.

**Preconditions:**

- The system has league data available (standings, team stats, player stats).
- The Scout/Analyst has access to the league analysis section.

**Trigger:** The Scout/Analyst selects a specific league to analyze.

**Main Success Scenario:**

1. The Scout/Analyst chooses a football league from the available options.
2. The system retrieves and displays a league dashboard, including:
    - Current league standings table.
    - League-wide aggregated statistics (e.g., average goals per game, possession stats).
    - Lists of top performers in key categories (top scorers, top assisters, most clean sheets).
    - Visualizations of team performance comparisons (e.g., attack vs. defense scatter plots).
3. The Scout/Analyst reviews the league overview and identifies trends or standout teams/players.

**Extensions (Alternative Flows):**

- **3a. Drill-down to Team/Player:** The Scout/Analyst clicks on a team or player from the league dashboard to view their detailed profile (linking to UC.2 or a similar team use case).
- **3b. Filter/Sort Data:** The Scout/Analyst applies filters (e.g., by date range, specific stats) or sorts leader boards.

**Postconditions:**

- The Scout/Analyst has a better understanding of the competitive landscape of the selected league.
- Potential scouting targets or areas of interest within the league may be identified.

# Chapter 4

# Analysis Modeling

## 4.1 Time Line Chart

The proposed system was developed in the final year of B.Tech, covering the span of August '24 - April '25. This includes the time taken for project ideation and project execution. A detailed chart showing the timeline of the project is shown below in figure :

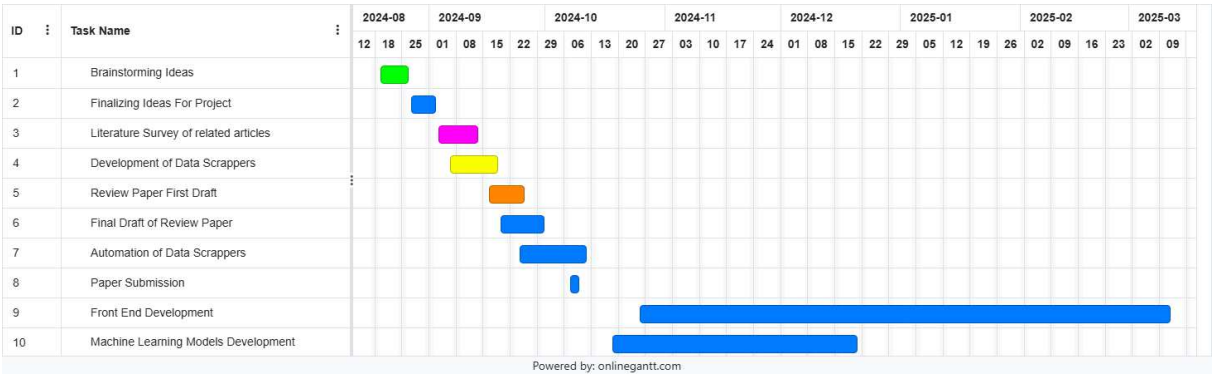



**Figure 1:** Timeline chart

# Chapter 5

# Design

## 5.1 Proposed Design

The proposed system uses advanced on-field player performance metrics, domestic and continental performance of football clubs, players perceived market value, player injury data and domestic competition metrics like standings and top performers. The data is collected from FBref and Transfermarkt using automated web scraping scripts. The collected data is put through a data processing pipeline which cleans and transforms the data to match the needs of the system. The preprocessed data is used by machine learning models for calculating consolidated metrics for comparing players Fig. 1. Proposed Methodology and teams, determining similar players, formation fit analysis and market price prediction. Figure 1 shows the proposed methodology.

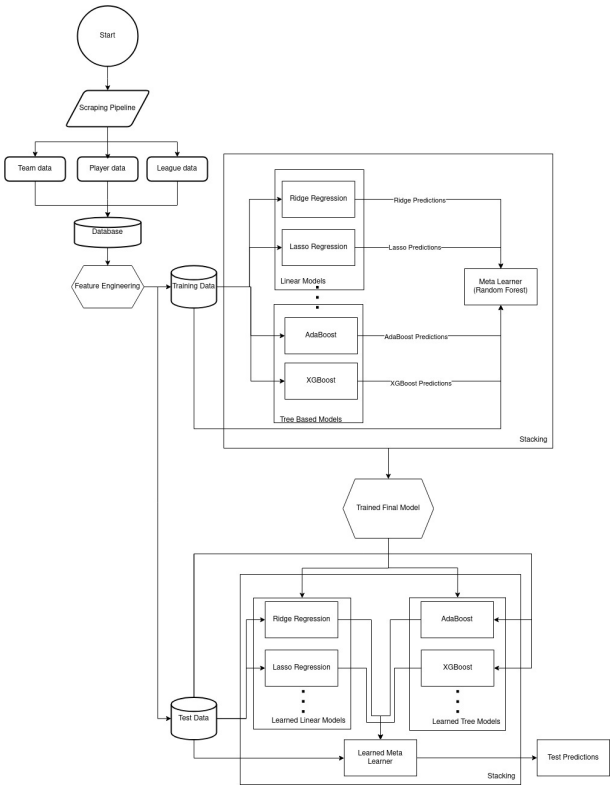



**Figure 2:** Proposed Model Architecture

## 5.2 User Interface Design

This section details the design and workflow of the application's user interface (UI). The UI was designed to facilitate intuitive navigation and efficient data exploration for users. The process is presented through a series of screenshots, demonstrating a typical user journey from the application's entry point to key functional areas.

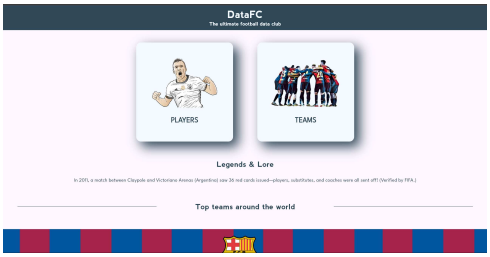
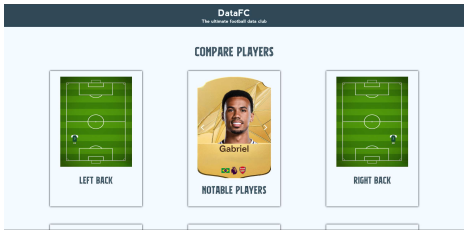
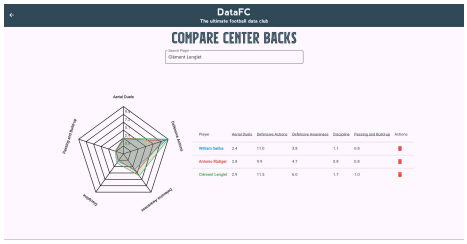



**Figure 3:** UI Home Page

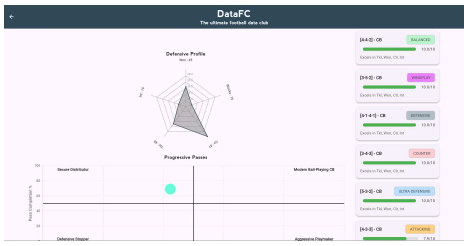
The initial interaction occurs at the **UI Home Page** (Figure 3). This serves as the central navigation hub, providing access to the core functionalities of the application, including player selection and league analysis. From this page, users can proceed to explore individual player statistics and comparisons, or explore their favourite leagues and teams.



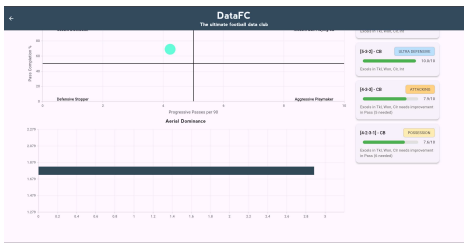

**(a)** Players section - Select player position




**(b)** Choose player to compare




**(c)** Player Visualization Example 1



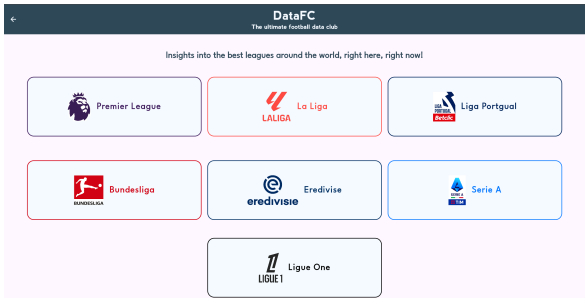

**(d)** Player Visualization Example 2

**Figure 4:** Player Selection and Visualization

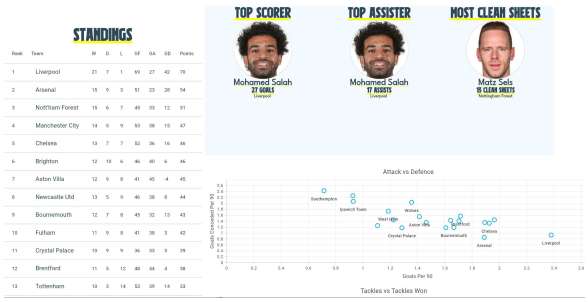
The subsequent stage, depicted in Figure 4, involves **Player Selection and Visualization**. The workflow commences with the user selecting a specific player position, narrowing the scope of potential candidates (Figure 4a). Subsequently, the user can choose a particular player for comparative analysis (Figure 4b). The application then generates visualizations illustrating key

performance indicators (KPIs) for the selected players, as shown in Figures 4c and 4d. These visualizations aim to provide a comprehensive overview of individual player performance.

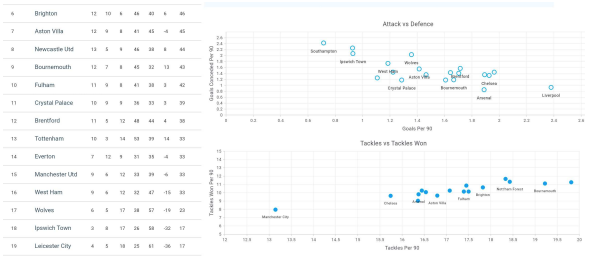
The final segment of the UI flow focuses on league-level analysis, as illustrated in Figure 5. This process begins with the user selecting a specific league of interest (Figure 5a). Following league selection, the application presents a series of dashboards providing an aggregate overview of league performance, as demonstrated in Figures 5b and 5c. These dashboards incorporate key metrics and visualizations to facilitate comprehensive league analysis.




**(a)** League Selection




**(b)** Premier League Dashboard 1




**(c)** Premier League Dashboard 2

**Figure 5:** League Selection and Dashboards

The overall design of the UI prioritizes ease of use and clarity of information presentation. The screenshot walkthrough presented in this section demonstrates a typical user journey and highlights the key functionalities of the application.

# Chapter 6

# Implementation

## 6.1 Data Pipeline

The data pipeline serves as the backbone of the system, ensuring that the data is systematically collected, cleaned and stored for efficient use for accurate analysis. The pipeline automates the extraction and preparation of data to be used in machine learning models.

## 6.2 Data Collection

The first step in the data pipeline is collection of data from external sources. The data is collected primarily from Fbref and Transfermarkt. Fbref provides data ranging from basic statistics for players like nationality, height, advanced on-field metrics such as expected goals and expected assists covering domestic leagues over forty countries. Transfermarkt gives detailed data about player transfers, including transfer fees, contract duration and clubs involved. Utilizing data from both the sources has offered a comprehensive perspective to the system. Multiple automated custom web scrapers were developed to systematically extract data related to players, teams, and the top five European domestic leagues. The scrapers are as follows:

1. **League Standings:** The system incorporates a web scraper tailored to retrieve the standings table of the top five domestic competitions in Europe allowing automated collection of team level data such as position, points, wins, losses and goal scoring statistics. Algorithm 1 shows the algorithm for the League Standings Scraper.

2. **Squad Stats Scraper:** The automated squad stats scraper ex- tracts team level data from Europe's top five leagues. The extracted data includes a vast set of tables which cover various performance dimensions such as standard stats (matches played, goals, xG), goalkeeping and advanced goalkeeping (save rates, post shot xG, etc.), shooting (shot volume, conversion), passing and pass types (progressive passes, pass lengths), goal and shot creation (SCAs, GCAs), defensive actions (tackles, interceptions), possession (carries, take-ons), playing time (minutes played, starts), and miscellaneous metrics (fouls, aerial duels, cards). This scraper enables the system to work with detailed and comprehensive statistical squad-level performance evaluation. Algorithm 2 shows the algorithm for the Squad Stats Scraper.

3. **Top Performers Scraper:** The Top Performers Scraper extracts data related to the top

performing players from Europe's top five leagues. The data includes information about the player with the most goals (number of goals, team name, player image link), the player with the most assists provided (number of assists provided, team name, player image link) and the goal keeper with the most clean sheets kept (number of clean sheets, team name, player image link). Algorithm 3 shows the algorithm for the Top Performers Scraper.

4. **Player Stats Scraper:** The Player Stats Scraper extracts player performance data. The data includes a wide range of performance metrics distributed in various categories. These include Standard Stats (appearances, minutes played, goals, and assists), Shooting (number of shots, shots accuracy, xG, xGoT), Passing (total passes, completion percentages, key passes), Pass Types (e.g. long passes, through balls, switches), Goal and Shot Creation metrics that capture both direct and indirect contributions to scoring opportunities, and Defensive Actions (tackles, interceptions, blocks). Possession-related (carries, touches, dribbles) as well as playing time (minutes per appearance, starting/substitution patterns) are also retrieved. This collection of scraper data allows the system to perform a detailed analysis of the contribution of a player throughout the season. Algorithm 4 shows the algorithm for the Player Stats Scraper.

5. **TransferMarkt Scraper:** Using the Transfermarkt scraper, the system is served with crucial off-field player metrics such as market value, injury status, and contract length, which are essential metrics consumed by the machine learning models for predicting the market prices of the players. Algorithm 5 shows the algorithm for the Transfermarkt Scraper.

## 6.3 Data preprocessing

Data preprocessing is an integral part of the methodology as it provides the system with clean and accurate data. The data preprocessing pipeline was implemented as a multistep approach to ensure the reliability and usability of the data. First, data from Fbref.com and Transfermarkt.com were consolidated based on the player's last name and age. In instances where the players' last names were the same, the positions of players were considered for correct aggregation. Secondly, several values were missing or zero by default (e.g., xG for most goalkeepers). To tackle such discrepancies, any row with more than eight-five percent missing values was dropped, ensuring robustness of the dataset. In the remaining dataset, numerical values were imputed using the median where the absence of values was insignificant. As the final step, the data on age and market values was put through the statistical technique called winsorization to limit the impact of outliers in the data. It was found that winsorization had an impact on the results and normalizing the data for players too old or too young aided in standardizing the data.

## 6.4 Feature Engineering

As part of the feature engineering process, the dataset was divided into ten classes: goalkeepers (GK), left backs (LB), right backs (RB), center backs (CB), center defensive midfielders (CDM), center midfielders (CM), center attacking midfielders (CAM), left wingers (LW), right wingers (RW), and center forwards (CF). For each class, customized metrics were developed using the preprocessed data. The metrics were as follows:

### 6.4.1 Goalkeepers (GK)

The following metrics quantify the on-field performances for goalkeepers.

#### 6.4.1.1 Shots Saved

The shot-stopping ability of goalkeepers per 90 minutes was calculated using the formula:

$$\text{Shots Saved} = \frac{\text{Saves}}{90} \tag{6.1}$$

#### 6.4.1.2 Expected Goals Prevention (EGP)

This custom metric quantifies the goals prevention performance of goalkeepers per 90 minutes played.

$$\text{Expected Goals Prevention (EGP)} = \frac{\text{PSxG} + \text{GA}}{90} \tag{6.2}$$

#### 6.4.1.3 Cross and Aerial Control (CAC)

Shows how well the goalkeeper performs at catching or punching crosses coming into the 16-yard box per 90 minutes played.

$$\text{Cross and Aerial Control (CAC)} = \frac{\text{Stp}}{90} \tag{6.3}$$

#### 6.4.1.4 Sweeper Keeper Activity (SKA)

Quantifies the goalkeeper's ability to perform sweeping actions outside the 16-yard box per 90 minutes played.

$$\text{Sweeper Keeper Activity (SKA)} = \frac{\text{OPA}}{90} \tag{6.4}$$

#### 6.4.1.5 Distribution Ability

Shows how capable the goalkeeper is at distributing the ball with their feet. Calculated per 90 minutes played.

$$\text{Distribution Ability} = \frac{\text{Cmp} + \text{KP} + \text{FinalThird}}{90} \tag{6.5}$$

**Table 1:** Goal Keepers

| Name | Shot Stopping | EGP | CAC | SKA | Distribution |
|---|---|---|---|---|---|
| Alisson | 2.20 | -0.01 | 0.33 | 1.87 | 26.26 |
| G. Donnarumma | 2.52 | -0.03 | 0.43 | 0.55 | 25.21 |
| J. Oblak | 2.42 | -0.08 | 0.61 | 0.81 | 20.73 |
| M. Maignan | 2.51 | -0.17 | 0.74 | 1.89 | 35.85 |
| T. Courtois | 1.91 | -0.06 | 0.43 | 0.57 | 29.52 |

### 6.4.2 Center Backs (CB)

The following metrics quantify the on-field performances for center backs. Table II displays a subset of the center backs dataset.

#### 6.4.2.1 Defensive Actions

Custom metric showing the center back's defensive contribution on the field per 90 minutes played.

$$\text{Defensive Actions} = \text{Defensive Contribution} \tag{6.6}$$

$\text{Defensive Contribution} =$

#### 6.4.2.2 Aerial Ability

Quantifies the aerial solidity of the center back per 90 minutes played.

$$\text{Aerial Ability} = \frac{\text{Won}}{90} \tag{6.7}$$

#### 6.4.2.3 Passing Ability

Shows how well the center back passes the ball and progresses the ball upfield per 90 minutes played.

$$\text{Passing Ability} = \frac{\text{Cmp} + \text{KP} + \text{PrgP}}{90} \tag{6.8}$$

#### 6.4.2.4 Positioning and Defensive Awareness

Quantifies the positional awareness of the center back on the field per 90 minutes played.

$$\text{Positioning and Defensive Awareness} = \frac{\text{Blocks} + \text{Clr}}{90} \tag{6.9}$$

#### 6.4.2.5 Disciplinary Record

Shows how disciplined the center back is across the game. Calculated per 90 minutes played.

$$\text{Disciplinary Record} = \frac{\text{CrdY} + \text{CrdR} + 2\text{CrdY} + \text{Fouls}}{90} \tag{6.10}$$

**Table 2:** Center Backs

| Name | Def. Actions | Aerial Duels | Passing | Def. Aware. | Discipline |
|---|---|---|---|---|---|
| Marquinhos | 12.43 | 4.07 | 13.85 | 4.80 | 1.15 |
| P. Cubarsí | 8.11 | 2.72 | 11.35 | 3.37 | 0.70 |
| P. Torres | 8.02 | 2.32 | 9.25 | 3.60 | 0.40 |
| V. van Dijk | 11.55 | 2.45 | 8.02 | 5.68 | 0.51 |
| W. Saliba | 10.98 | 2.43 | 8.08 | 3.83 | 1.10 |

### 6.4.3 Full Backs (LB and RB)

The following custom metrics have been used to quantify the performances of left backs and right backs. Table III displays a subset of the fullbacks' dataset.

#### 6.4.3.1 Defensive Duties

$$\text{Defensive Duties} = \frac{\text{Def 3rd} + \text{Int} + \text{Blocks} + \text{Clr} + \text{Recov}}{90} \tag{6.11}$$

#### 6.4.3.2 Offensive Contributions

$$\text{Offensive Contributions} = \frac{\text{PrgC} + \text{PrgP} + \text{KP} + \text{xA}}{90} \tag{6.12}$$

#### 6.4.3.3 Final Third Play

This custom metric shows how well the full-back makes themselves available to contribute in the final third, per 90 minutes played.

$$\text{Final Third Play} = \frac{\text{Crs} + \text{SCA} + \text{CPA} + \text{PPA}}{90} \tag{6.13}$$

#### 6.4.3.4 Possession Play

Quantifies how well the full-back takes care of the ball on their feet, per 90 minutes played.

$$\text{Possession Play} = \frac{\text{Att 3rd possession} + \text{TotDist}}{90} \tag{6.14}$$

#### 6.4.3.5 Dribbling Accuracy

Measures how well the player dribbles through the opposition's press.

$$\text{Dribbling Accuracy} = \frac{\text{Succ}}{90} \tag{6.15}$$

**Table 3:** Full Backs

| Name | Att. Contributions | Final Third | Possession | Dribbling |
|---|---|---|---|---|
| A. Balde | 9.24 | 8.29 | 29.83 | 0.43 |
| A. Robertson | 10.69 | 9.68 | 32.86 | 0.04 |
| D. Udogie | 9.48 | 3.56 | 26.86 | 0.11 |
| F. Dimarco | 7.41 | 14.59 | 35.74 | 0.03 |
| F. Mendy | 4.61 | 0.71 | 21.42 | 0.03 |

### 6.4.4 Central Defensive Midfielders (CDM)

The following custom metrics have been used to quantify the performances of center defensive midfielders.

#### 6.4.4.1 Defensive Contributions

$$\text{Defensive Contributions} = \frac{\text{Tkl} + \text{Int} + \text{Blocks} + \text{Clr} + \text{Recov}}{90} \tag{6.16}$$

#### 6.4.4.2 Passing Ability

$$\text{Passing Ability} = \frac{\text{Cmp}}{90} \tag{6.17}$$

#### 6.4.4.3 Build-Up Play

$$\text{Build-Up Play} = \frac{\text{xA} + \text{xAG} + \text{Ast} + \text{PrgDist}}{90} \tag{6.18}$$

#### 6.4.4.4 Ball Recovery & Defensive Work

$$\text{Ball Recovery \& Defensive Work} = \frac{\text{Recov} + \text{Int}}{90} \tag{6.19}$$

#### 6.4.4.5 Line Breaking Passes

$$\text{Line Breaking Passes} = \frac{\text{KP} + \text{PrgP} + (1/3)\ \text{passing}}{90} \tag{6.20}$$

**Table 4:** Center Defensive Midfielders

| Name | Def. Work | Passing | Build-Up | Recoveries | Line Breaking |
|---|---|---|---|---|---|
| B. Guimarães | 10.26 | 3.07 | 22.80 | 5.98 | 13.75 |
| Casemiro | 17.54 | 6.53 | 24.28 | 6.80 | 11.14 |
| G. Xhaka | 9.33 | 3.71 | 37.14 | 5.57 | 23.84 |
| J. Neves | 13.41 | 5.73 | 26.61 | 7.46 | 16.13 |
| Y. Bissouma | 12.95 | 7.32 | 21.00 | 6.63 | 10.24 |

### 6.4.5 Central Midfielders (CM)

The following custom metrics have been used to quantify the performances of center midfielders.

#### 6.4.5.1 Passing and Vision

Quantifies how well the center midfielders pass the ball to contribute to offensive phases of the play, per 90 minutes played.

$$\text{Passing and Vision} = \frac{\text{PrgP} + (1/3)\ \text{passing}}{90} \tag{6.21}$$

#### 6.4.5.2 Dribbling

Shows how well the center midfielder takes care of the ball and dribbles past opponents, per 90 minutes played.

$$\text{Dribbling} = \frac{\text{Succ} + \text{PrgC} + \text{CPA}}{90} \tag{6.22}$$

#### 6.4.5.3 Defensive Work

Explains the contribution of the center midfielder in defence, per 90 minutes played.

$$\text{Defensive Work} = \frac{\text{Tkl} + \text{Int} + \text{Blocks} + \text{Clr} + \text{Recov}}{90} \tag{6.23}$$

#### 6.4.5.4 Chance Creation

Quantifies the creative qualities of the center midfielder.

$$\text{Chance Creation} = \frac{\text{SCA} + \text{xG} + \text{xA} + \text{xAG}}{90} \tag{6.24}$$

#### 6.4.5.5 Possession Retention

Shows the ability of the center midfielder to retain the ball and not concede possession to the opposition, per 90 minutes played.

$$\text{Possession Retention} = \frac{\text{Cmp} + \text{KP} + (1/3)\ \text{passing} + \text{Succ}}{90} \tag{6.25}$$

**Table 5:** Center Midfielders

| Name | Passing | Dribbling | Def. Work | Chance Creation | Possession |
|---|---|---|---|---|---|
| D. Rice | 11.26 | 4.13 | 9.56 | 4.17 | 54.04 |
| F. Valverde | 13.30 | 2.67 | 10.72 | 2.91 | 67.67 |
| Pedri | 18.64 | 4.20 | 12.43 | 4.92 | 75.93 |
| Vitinha | 2.19 | 4.66 | 8.66 | 3.16 | 15.71 |

### 6.4.6 Central Attacking Midfielders (CAM):

The following custom metrics have been used to quantify the performances of central attacking midfielders.

**a. Creativity and Playmaking:** Quantifies the creativity of the central attacking midfielder, per 90 mintutes played.

$$\text{Playmaking} = \frac{xA + SCA + 1/3\_\text{passing}}{90s} \tag{6.26}$$

**b. Ball Progression (BP):** Numerifies the ability of the central attacking midfielder to move the ball up-field, per 90 minutes played.

$$\text{Ball Progression} = \frac{PrgP + PrgC}{90s} \tag{6.27}$$

**c. Final Third Impact (FTI):** Shows how much the central attacking midfielders impacts the game in the final (attacking) third, per 90 minutes played.

$$\text{Final Third Impact} = \frac{Att3rd\_possession + CPA + PPA}{90s} \tag{6.28}$$

**d. Goal Threat:** This metric shows the central attacking midfielder's ability to score goals, per 90 minutes played.

$$\text{Goal Threat} = \frac{xG + npxG + Gls}{90s} \tag{6.29}$$

**e. Final Ball Efficiency (FBE):** The ability of the central attacking midfielder to deliver the final ball, per 90 minutes played.

$$\text{Final Ball Efficiency} = \frac{xA + xAG + PPA}{90s} \tag{6.30}$$

**Table 6:** Central Attacking Midfielders

| Name | Playmaking | BP | FTI | Goal Threat | FBE |
|---|---|---|---|---|---|
| D. Olmo | 7.76 | 9.11 | 31.22 | 1.4 | 0.79 |
| F. Wirtz | 9.52 | 11.44 | 51.39 | 1.09 | 1.15 |
| Isco | 13.09 | 10.2 | 34.3 | 1.22 | 0.6 |
| Ju. Bellingham | 8.44 | 10.10 | 27.15 | 1.11 | 0.58 |
| X. Simons | 7.69 | 9.62 | 29.75 | 0.75 | 0.78 |

### 6.4.7 Wingers (LW and RW):

The following custom metrics have been used to quantify the performances of wingers.

**a. Dribbling and Ball Carrying:** Shows how well the winger is at dribbling the ball past opponents, per 90 minutes played.

$$\text{Dribbling} = \frac{Succ + PrgC + CPA}{90s} \tag{6.31}$$

**b. Crossing and Playmaking (CAP):** This metric quantifies the playmaking and crossing ability

of the winger, per 90 minutes played.

$$\text{Crosses and Playmaking} = \frac{xA + xAG + Crs}{90s} \tag{6.32}$$

**c. Goal Threat (GT):** This metric shows the winger's ability to score goals, per 90 minutes played.

$$\text{Goal Threat} = \frac{xG + npxG + Gls}{90s} \tag{6.33}$$

**d. Final Third Involvement (FTI):** Shows how much the winger impacts the game in the final (attacking) third, per 90 minutes played.

$$\text{Final Third Impact} = \frac{Att\_3rd\_possession + CPA + PPA}{90s} \tag{6.34}$$

**Table 7:** Wingers

| Name | Dribbling | CAP | Goal Threat | FTI |
|---|---|---|---|---|
| Bukayo Saka | 10.07 | 7.82 | 0.89 | 42.62 |
| Ousmane Dembélé | 12.51 | 5.56 | 3.07 | 48.25 |
| Rodrygo | 9.39 | 4.02 | 0.72 | 41.80 |
| Raphinha | 6.39 | 7.80 | 1.73 | 38.85 |

### 6.4.8 Center Forwards (CF):

The following custom metrics have been used to quantify the performances of center forwards.

**a. Goal Threat (GT):** This metric shows the center forward's ability to score goals, per 90 minutes played.

$$\text{Goal Threat} = \frac{xG + npxG + Gls}{90s} \tag{6.35}$$

**b. Chance Conversion :** Quantifies how efficient is the center forward at converting chances, per 90 minutes played

$$\text{Chance Conversion} = \frac{G - PK + xG}{90s} \tag{6.36}$$

**c. Link-up Play (LUP):** Shows how well the center forward links up with the team through passing the ball, per 90 minutes played.

$$\text{Link-Up Play} = \frac{PrgR + xA + PPA}{90s} \tag{6.37}$$

**d. Shooting Accuracy:** Shows how accurately does the center forward shoot the ball on goal, per 90 minutes played.

$$\text{Shooting Accuracy} = \frac{SoT + Sh}{90s} \tag{6.38}$$

**e. Penalty Box Presence (PBP):** Quantifies the how present the center forward is in the penalty box, per 90 minutes played.

$$\text{Penalty Box Presence} = \frac{Att\_Pen}{90s} \tag{6.39}$$

**Table 8:** Center Forwards

| Name | GT | Ch.Conv. | LUP | Shooting Accuracy | PBP |
|---|---|---|---|---|---|
| E. Haaland | 2.18 | 0.05 | 4.12 | 2.06 | 6.20 |
| H. Kane | 2.36 | 0.07 | 7.05 | 1.73 | 6.07 |
| K. Mbappé | 2.02 | 0.05 | 14.00 | 2.25 | 9.70 |
| L. Martínez | 1.35 | 0.03 | 7.54 | 1.32 | 5.56 |
| R. Lewandowski | 2.76 | 0.08 | 5.66 | 1.65 | 5.96 |

## 6.5 Machine Learning Models

This section presents the proposed architecture of the machine learning model. The parameters and rationale behind the selected models are also explained.

### 6.5.1 Dimensionality Reduction

To avoid over-fitting, which is often introduced with high-dimensional data, PCA was performed after cleaning and feature engineering. This preprocessing reduced some dimensionality and highly correlated data, ensuring that the principal components identified by PCA were robust and effectively captured the variance.

- The Elbow Method (Scree Plot) explains the variance versus the number of components in a range (x, y).

To better understand the workings behind PCA, the mathematical formulas governing PCA are provided below.

Given a standardized dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $n$ is the number of samples and $d$ is the number of features, the covariance matrix is computed as:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X} \tag{6.40}$$

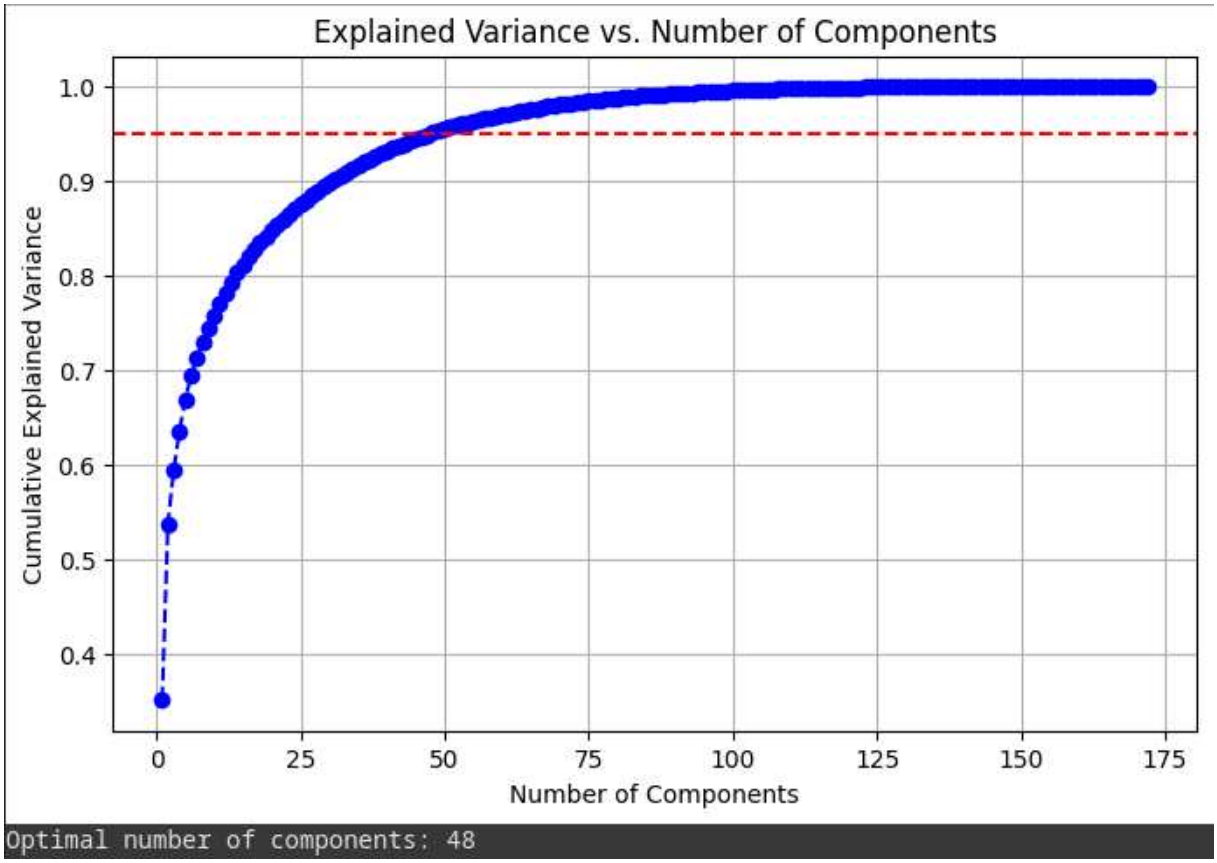



**Figure 6:** PCA Scree Plot

Eigenvalue decomposition is then performed on the covariance matrix:

$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{6.41}$$

where $\lambda_i$ is the eigenvalue corresponding to the eigenvector $\mathbf{v}_i$. The eigenvectors are sorted in descending order of their eigenvalues.

Let $\mathbf{V}_k \in \mathbb{R}^{d \times k}$ be the matrix of the top $k$ eigenvectors. The data is then projected onto the new $k$-dimensional subspace as[11]:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}_k \tag{6.42}$$

Here, $\mathbf{Z} \in \mathbb{R}^{n \times k}$ is the transformed feature matrix in the reduced-dimensional space.

Following PCA, the transformed data are passed into an SEM, a stacked ensemble model, which performs the prediction of the market value of a player.

### 6.5.2 Stacked Ensemble Modeling

In recent times, multiple meta-heuristic learners and optimization algorithms have been doctored in the domain of football analytics[12]. This study presents a novel approach among those to predict the market value of a football player, based on real data, compared to the FIFA values used by many other studies.

Stacking is an ensemble approach that uses a 2-level approach, level-0 base learners and a level-1 meta learner. The level-1 meta learner is an aggregator that receives the output from single-based learners.

#### 6.5.2.1 Base Learners

The base learners used are selected to capture linear and non-linear relationships in the data. Both parametric and non-parametric models were selected.

**Linear Learners**

- **Ridge Regression**: Also called Tikhonov regularization, it is used in ill-posed problems, useful to mitigate the problem of multicollinearity in regression problems, caused by a high dimensionality. It is useful in this study due to large number of principal components (>40). It employs $L_2$ regression to control multilinearity. Ridge regression minimizes the following loss function[13]:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_2^2 \right\} \tag{6.43}$$

- **Lasso Regression**: Lasso Regression is a statistical operator that penalizes the model to prevent overfitting and enhance accuracy. It does so by shrinking some coefficients to zero, effectively excluding them from the model. It employs $L_1$ regression for automation feature selection.

  Lasso regression introduces an $L_1$ penalty to promote sparsity[14]:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \right\} \tag{6.44}$$

- **ElasticNet**: ElasticNet combines both, Lasso and Ridge regression, which improves its ability with regards with reconstruction. ElasticNet combines both $L_1$ and $L_2$ penalties[15]:

$$\hat{\beta} = \arg\min_{\beta} \{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \} \tag{6.45}$$

**Tree-Based Learners**

- **Random Forest Regressor**: Random forest regression is a machine learning technique that uses multiple decision trees to predict continuous values using the bootstrap averaging method on the outputs given by each tree. It is included for its ability in smoothing out the errors from other learners, in the overall stacking architecture, while effectively capturing non-linear dependencies.
- **XGBoost**: By leveraging a magnitude of decision trees, it focuses on creating a powerful predictive model by using an iterative process that focuses on minimizing errors. XGBoost is proved to be unbeatable for handling structured data, along with its ability to address class imbalance, which was crucial in our study, due to the limited data of goalkeepers compared to other positions.
- **LightGBM**: LightGBM is designed specifically for large-scaled data, and employs a leaf-wise growth strategy, opposed to the level-based growth strategy used by other tree based algorithms. This allows it to have deeper trees and better predictive ability.
- **Gradient Boosting Regressor**: Similar to XGBoost, it builds the model iteratively.
- **AdaBoost Regressor**: AdaBoost is often used in conjunction with other machine learning algorithms to improve performance. The output of multiple weak learners is combined into a weighted sum that represents the final output of the model[16].
- **Support Vector Regression (SVR)**: SVR tries to find a function that is able to accurately predict the continuous output value for any given output value. It uses both linear and non linear kernels.

### 6.5.3 Meta Learner

The meta learner, also known as the Level-1 learner, is the final model that receives the output from all the base models, and predicts the market value by using the stacking ensemble algorithm. This study uses the Random Forest Regressor as the meta learner. A common strategy to train a stacking model is to use a hold-out set, where the dataset is split into two parts - the first layer is trained using the first part of the dataset, and the second layer is given the second set of data. The output from the meta model is used to create a new dataset, which makes this new dataset a 3D dataset. This new dataset ensures that the model learns the target value, given the inputs

from the first layer. The rationale behind using the Random Forest in both base learners and as a meta learner is to use its ability to generalize well, and smoothen the results, and helps capture the complex relationships. As a base learner, it introduces diversity, and as a meta learner, its insensitivity to multicollinearity makes it a strong candidate to be used in both the layers.

The mathematical representation for the proposed model is given as:

$$\hat{y} = f_{\text{meta}}\left(f_1(\mathbf{Z}), f_2(\mathbf{Z}), \ldots, f_m(\mathbf{Z})\right) \tag{6.46}$$

**Where:**

- $\hat{y}$: Final prediction (market value)
- $f_1, f_2, \ldots, f_m$: Base learners (e.g., Ridge, Lasso, RF, XGBoost, etc.)
- $\mathbf{Z} \in \mathbb{R}^{n \times k}$: PCA-reduced feature matrix with $n$ samples and $k$ components
- $f_{\text{meta}}$: Meta learner (Random Forest Regressor)

# Chapter 7

# Testing

To evaluate the performance of the final SEM, multiple regression metrics were computed on both training and test sets. The rationale behind computing both training and testing sets was to check for overfitting. The difference between the training and test sets is a great metric to understand the state of the model. The higher the difference between these, the more the model overfits or underfits. The model achieved an $R^2$ score of 0.9464 on the training set and 0.9457 on the test set. The metrics are shown in Table 9.

The cross-validation $R^2$ score being close to the $R^2$ scores of the train and test sets indicates that the model generalizes well and avoids significant overfitting. Furthermore, the chosen combination of base learners and meta learner appears suitable for the type of data this study addresses.

**Table 9:** Evaluation metrics for the Stacking Ensemble Model

| **Metric** | **Train** | **Test** |
|---|---|---|
| $R^2$ Score | 0.9464 | 0.9457 |
| Mean Squared Error (MSE) | $9.73 \times 10^{13}$ | $9.69 \times 10^{13}$ |
| Root Mean Squared Error (RMSE) | 3,119,656.30 | 3,112,937.64 |
| Mean Absolute Error (MAE) | 2,199,469.47 | 2,259,460.70 |
| Cross-Validation $R^2$ (Mean $\pm$ Std) | | $0.9383 \pm 0.0029$ |

# Chapter 8

# Results and Discussions

This section details the findings of our research, specifically the performance evaluation of the developed Stacked Ensemble Model (SEM) for predicting player market values in football. The analysis encompasses key performance indicators, a thorough examination of potential overfitting issues, and a discussion of the broader implications of the model's results within the football ecosystem.

## 8.1 Overall Performance

The SEM's efficacy was gauged using a variety of established metrics, assessed on both the training and test datasets:

- **R-squared ($R^2$):** The SEM yielded an $R^2$ of 0.9464 when applied to the training dataset and a closely comparable $R^2$ of 0.9457 on the test dataset. These results denote that the model effectively elucidates approximately 94.64% of the variance inherent in the training data and 94.57% of the variance observed in the test data.
- **Mean Squared Error (MSE):** The MSE registered at 9.73 x $10^{13}$ for the training set and 9.69 x $10^{13}$ for the test set. These values represent the average of the squared discrepancies between the model-projected and actual market valuations.
- **Root Mean Squared Error (RMSE):** Quantitatively, the RMSE was found to be 3,119,656.30 and 3,112,937.64 for the training and test sets, respectively. Measured in the same unit as the target variable (Euros), the magnitude of the model's average prediction error becomes more readily apparent. The value of 3,112,937.64 implies the model, on average, estimates market values within approximately 3.1 million Euros of the true value.
- **Mean Absolute Error (MAE):** An MAE of 2,199,469.47 was seen on the training and 2,259,460.70 on the test set. The MAE provides a robust indication of the magnitude of the average error, at roughly 2.26 million euros.
- **Cross-Validation $R^2$:** Cross-validation yielded a mean $R^2$ of 0.9383, with a standard deviation of ± 0.0029. This indicates the model's robust ability to generalize and the overall stability and reliability of the data utilized.

As can be seen in Table 6.1, these metrics suggest that the SEM exhibits strong predictive

performance and generalization ability.

## 8.2 Overfitting Analysis

We explicitly explored the possibility of overfitting to ensure the reliability of our model's predictive capacity. As a gauge for overfitting potential, we compared the $R^2$ metric between our training and test sets. Here, a relatively small gap was observed (0.9464 and 0.9457, respectively) indicating the model's ability to generalize to unseen data.

The results from our cross-validation exercise ($0.9383 \pm 0.0029$) corroborate this finding. A close alignment between the mean cross-validated $R^2$ and the $R^2$ calculated for the test set further suggest the model is not overly reliant on the nuances of the training data, demonstrating good out-of-sample performance.

## 8.3 Justification for Model Architecture and Component Choices

The SEM architecture was thoughtfully chosen to represent both linear and non-linear associations within the player market data. We chose the Random Forest Regressor as our meta-learner due to the following characteristics:

- **Generalization:** Random Forest's ability to generalize improves overall model robustness and accuracy.
- **High Volume Handling**: It effectively handles the large feature space while facilitating easier identification and prediction.

## 8.4 Limitations

Several limitations should be specified to guide the model's improvement in future iterations. These limitations include:

- Inherent biases and potential incompleteness in the data sourced from FBref and Transfermarkt, primarily due to their reliance on scouting subjective inputs.
- Exclusion of qualitative factors such as contract negotiations or player psychology data.

Future improvements include:

- Expanding the input data with additional sources that encompass player psychology and other factors such as contract negotiations.

- Additional consideration of complex machine learning algorithms for higher performance accuracy.
- Further investigation and analysis on feature importances for model improvement.

# Chapter 9

# Conclusion

By efficiently automating data extraction and processing of advanced player statistics from robust and vast data sources such as Fbref and Transfermarkt, the proposed system uses the extracted data to not only automate the manual and subjective process of scouting football players, but also drive the decision making process of coaches and managers. The system leverages complex machine learning models for predicting the market value predictions, along with clustering algorithms to identify players that are similar both, tactically and stylistically. The system includes role-specific metrics and performs position-aware analysis. Additionally, the usage of detailed visualizations such as radar charts and other performance graphs, makes the data analysis interpretable and actionable for coaches.

The system provides a robust framework to find replacements for football players, there are several aspects that exist for future enhancements. In the future, we would like to work on integrating match-by-match real time performance data which would offer dynamic tracking of players on the field. Additionally, building an all- inclusive analysis hub by including video analytics and sentiment analysis of players from media sources. Finally, expanding our databases to retired players, women's leagues and youth leagues would open underexplored doors.

# References


[1] O. Müller, A. Simons, and M. Weinmann, "Beyond crowd judgments: Data-driven estimation of market value in association football," *European Journal of Operational Research*, vol. 263, no. 2, pp. 611–624, 2017.

[2] Q. Yi, M.-A. Gómez, H. Liu, B. Gao, F. Wunderlich, and D. Memmert, "Situational and positional effects on the technical variation of players in the uefa champions league," *Frontiers in Psychology*, vol. 11, p. 1201, 2020.

[3] F. Carmichael and D. Thomas, "Bargaining in the transfer market: Theory and evidence." *Applied Economics*, vol. 25, pp. 1467–76, 12 1993.

[4] M. A. Al-Asadi and S. Tasdemır, "Predict the value of football players using fifa video game data and machine learning techniques," *IEEE Access*, vol. 10, pp. 22 631–22 645, 2022.

[5] Transfermarkt, "Transfermarkt - football transfers and market values," https://www.transfermarkt.com, 2024, accessed: 2024-04-11.

[6] R. Stanojevic and L. Gyarmati, "Towards data-driven football player assessment," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 167–172.

[7] S. Dobson, B. Gerrard, and S. Howe, "The determination of transfer fees in english nonleague football," *Applied Economics*, vol. 32, no. 9, pp. 1145–1152, 2000.

[8] A. T. Yiğit, B. Samak, and T. Kaya, "Football player value assessment using machine learning techniques," in *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*, 2020, pp. 289–297.

[9] I. Behravan and S. M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," *Soft Computing*, vol. 25, no. 3, pp. 2499–2511, 2021.

[10] I. G. McHale and B. Holmes, "Estimating transfer fees of professional footballers using advanced performance metrics and machine learning," *European Journal of Operational Research*, vol. 306, no. 1, pp. 389–399, 2023.

[11] I. Jolliffe, "Principal component analysis springer verlag," 01 2002.

[12] S. Buyrukoğlu and S. Savaş, "Stacked-based ensemble machine learning model for positioning footballer," *Arabian Journal for Science and Engineering*, vol. 48, no. 3, pp. 1371–1383, 2023.

[13] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. [Online]. Available: http://www.jstor.org/stable/1267351

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[15] H. Zou and T. Hastie, "Zou h, hastie t. regularization and variable selection via the elastic net. j r statist soc b. 2005;67(2):301-20," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301 – 320, 04 2005.

[16] S. Rosset, H. Zou, and T. Hastie, "Multi-class adaboost," *Statistics and its interface*, vol. 2, 02 2006.

# Publications

Harde, S., Bhawnani, V., & Savant, S. (2025). Comparative Analysis of Data Driven Techniques to Predict Transfer Prices of Football Players. *International Journal of Innovative Science and Research Technology*, *10*(3), 735–739. Published by IJISRT.

# Acknowledgements

We would like to express our gratitude to the personnels and institutes that helped us throughout the course of this project and whose guidance was invaluable for the completion of this project.

First and foremost, we would like to thank our project guide, Prof. Shruti Dodani for her constant support, astute insights and constructive feedback through every stage of the project. Her expertise was instrumental in shaping up this project and bringing it to fruition. As Brad Henry once said, *"A good teacher can inspire hope, ignite the imagination, and instill a love of learning."*

We are also grateful to the AI & DS Department, DJSCE, for providing the necessary resources and an academic environment fostering growth and inquisitiveness.

Having an idea and having it turn into a project is as hard as it sounds. We would like to thank all those around us, our friends and family, for their patience, understanding and constant encouragement throughout our journey.

We started this journey hoping to bring a positive impact, and the support and encouragement we received from everyone has been crucial in pursuing that aspiration.