

AI Football Performance Analyzer

Members:

- **Adrien Schuttig**, ECE Paris, adrienschuttig@gmail.com
 - **Maxime Laurent**, ESILV Nantes, maxime.laurent@edu.devinci.fr
 - **Louis Le Forestier**, ESILV Nantes, louis.le_forestier@edu.devinci.fr
-

Table of Contents

1. [Introduction](#)
 2. [Datasets](#)
 3. [Methodology](#)
 4. [Evaluation & Analysis](#)
 5. [Related Work](#)
 6. [Conclusion](#)
-

I. Introduction

Motivation: Why are we doing this?

Football is a data-rich sport where analyzing player performance can significantly impact scouting, training, and match strategy. We aim to leverage machine learning to move beyond subjective observation and provide data-driven insights into player capabilities and potential.

What do we want to see at the end?

We aim to build a system that can:

1. **Analyze** current player statistics to evaluate their overall performance.
2. **Predict** a player's future development trajectory (e.g., will they improve, stay stable, or decline?).
3. **Visualize** these insights in an accessible way for coaches and analysts.

II. Datasets

We are using the **Football Players Data** dataset from Kaggle for this study.

Source and Description

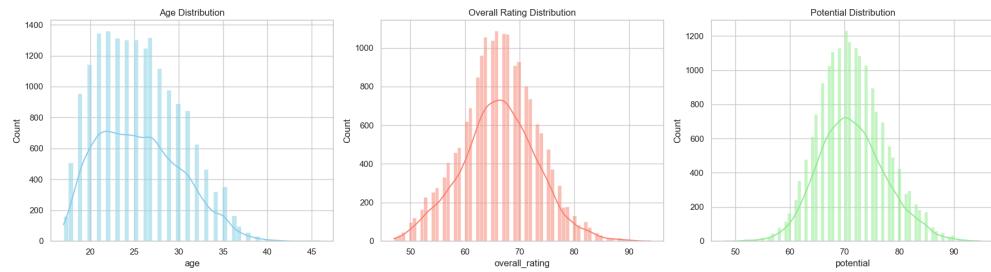
Attribute	Details
Source	Kaggle Link: Football Players Data
Size	17,954 rows (players)
Columns	51 attributes (physical, technical, mental, ratings)

Exploratory Data Analysis (EDA)

The EDA helped in understanding the data structure and identifying key relationships for modeling.

1. Skill Distribution

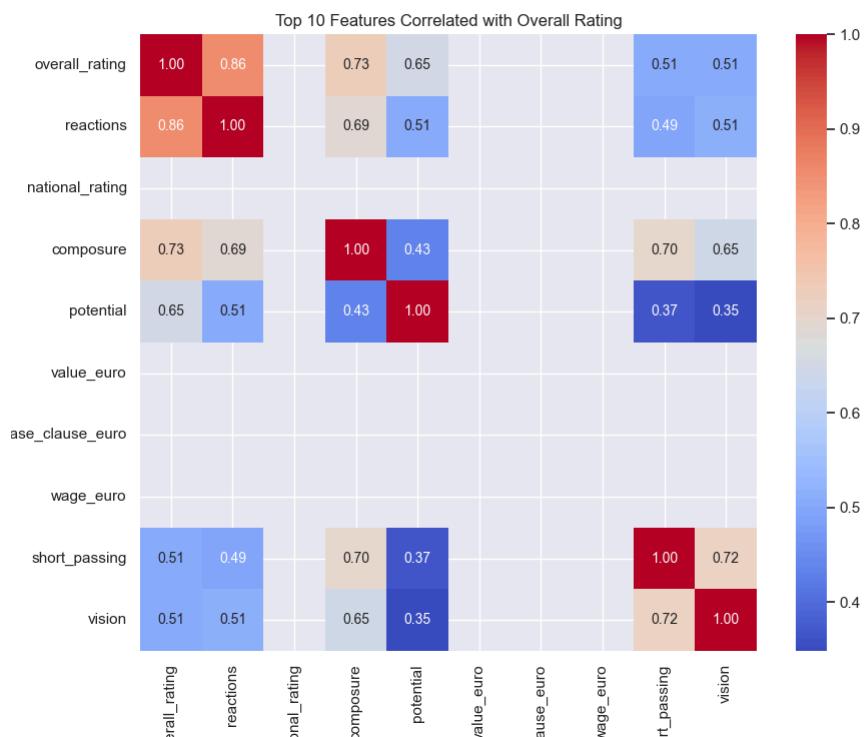
The distribution of ratings (`overall_rating` and `potential`) is **strongly skewed to the right**, indicating that the majority of players fall into the low to medium rating categories.



2. Correlation with Performance (`overall_rating`)

The correlation matrix highlights the attributes that most influence the overall rating.

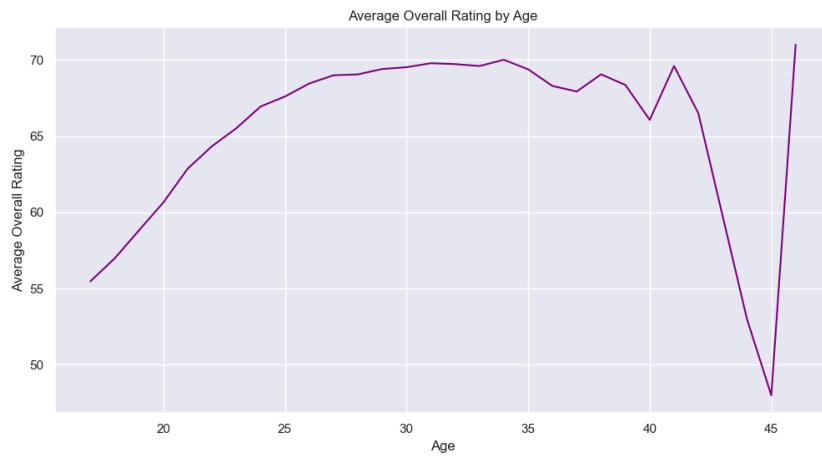
Attribute	Correlation	Observation
<code>reactions</code>	0.84	The most predictive attribute, emphasizing the importance of rapid decision-making.
<code>potential</code>	0.69	Indicates a strong influence of future prospects on the current assessment.
<code>composure</code>	0.68	A key mental factor for performance.
<code>short_passing</code>	0.59	Core technical skills are crucial.



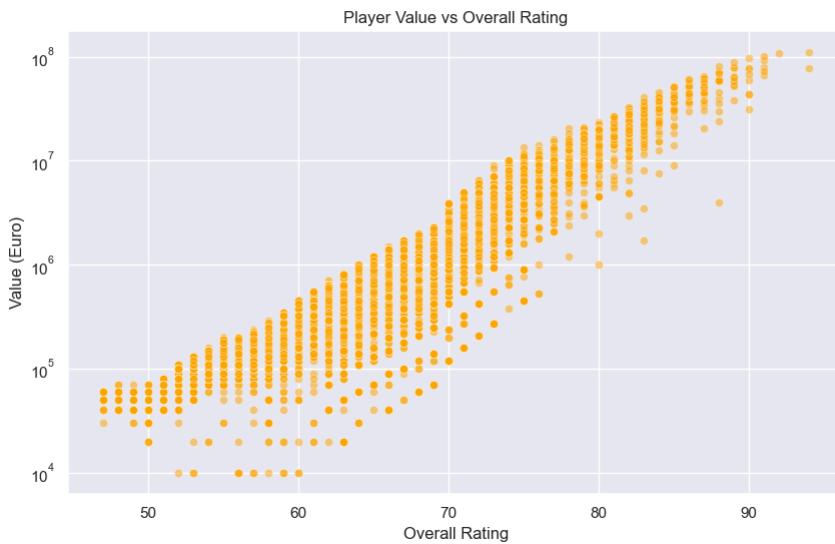
3. Age, Value, and Nationality Relationship

Age analysis confirms that the average player level (`overall_rating`) **peaks between 27 and 31 years** before declining.

- **Age vs Rating:**



- **Value vs Rating:** Market value (`value_euro`) is highly correlated with the overall rating, increasing exponentially, as shown on the logarithmic scale.



Critical Data Assessment

Category	Positive Points	Negative Points
----------	-----------------	-----------------

Feature Richness	✓ 51 varied attributes (physical, technical, mental) for comprehensive analysis.	✗ High Multicollinearity (several attributes are highly correlated with each other), which will necessitate rigorous feature selection.
Modeling Objective	✓ potential is an excellent indicator for future growth prediction.	✗ The <code>overall_rating</code> relies heavily on the <code>reactions</code> variable, which could bias the model if overused, at the expense of other skills.

III. Methodology

Choice of Algorithms

We employ two primary machine learning models using `scikit-learn`:

1. Linear Regression (for Current Performance):

- **Goal:** Predict a player's `overall_rating`.
- **Why:** The target variable is continuous. Linear regression provides a clear interpretation of how each feature (like speed or shooting) contributes to the overall score.

2. Logistic Regression (for Future Development):

- **Goal:** Classify a player's future growth into categories: `high_growth`, `likely_improve`, `stable`, or `decline`.
- **Why:** This is a classification problem. We created a custom target variable based on the gap between `potential` and `overall_rating` and the player's `age`. Logistic regression allows us to estimate the probability of a player falling into each category.

Features

We selected a subset of key attributes that most strongly correlate with performance:

- **Physical:** Age, Height, Weight, Acceleration, Sprint Speed, Stamina, Strength.
- **Technical:** Finishing, Dribbling, Short Passing.

IV. Evaluation & Analysis

Regression Results (Predicting Overall Rating)

- **Model:** Linear Regression
- **Metrics:**
 - **MSE (Mean Squared Error):** Measures the average squared difference between estimated values and the actual value.
 - **R² Score:** Indicates how well the data fit the regression model.
- **Visualization:** We generated a scatter plot (`reg_true_vs_pred.png`) comparing true ratings vs. predicted ratings. A tight clustering around the diagonal indicates high accuracy.

Classification Results (Predicting Future Growth)

- **Model:** Logistic Regression (Multinomial)
- **Classes:**
 - `high_growth` : Young players with a large potential gap.
 - `likely_improve` : Players with significant room for improvement.
 - `stable` : Players near their peak.

- `decline` : Players whose potential is lower than their current rating.
- **Metrics:** We use Precision, Recall, and F1-Score to evaluate the classifier's performance across all classes.

V. Related Work

- **Libraries Used:**
 - `pandas` : For data manipulation and cleaning.
 - `scikit-learn` : For implementing Linear and Logistic Regression models.
 - `matplotlib / seaborn` : For data visualization.
- **References:**
 - Scikit-learn Documentation: <https://scikit-learn.org/>
 - Kaggle Dataset: [Football Players Data](#)

VI. Conclusion: Discussion

This project demonstrates that standard player attributes can effectively predict both current ability and future potential.

- **Findings:** Physical stats combined with technical skills like passing and dribbling are strong predictors of a player's overall rating.
- **Future Work:** We could enhance the model by:
 - Incorporating match performance data (goals, assists per game).
 - Using more complex models like Random Forests or Neural Networks for non-linear relationships.
 - Building a web interface (Streamlit) to allow users to input player stats and get real-time predictions.