**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**BỘ MÔN LẬP TRÌNH NGÔN NGỮ PYTHON**

****

**BÁO CÁO VỀ BÀI TẬP LỚN**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **: KIM NGỌC BÁCH** |
| **Lớp** | **: D23CQCE06-B** |
| **Họ Tên**  **MSV** | **: LÊ ĐĂNG HIỆP**  **: B23DCDT086** |
|  |  |

***Hà Nội – 2025***

**LỜI MỞ ĐẦU**

Trong thời đại dữ liệu phát triển mạnh mẽ như hiện nay, việc thu thập, xử lý và phân tích dữ liệu đóng vai trò then chốt trong nhiều lĩnh vực, đặc biệt là trong thể thao, nơi mà mỗi con số đều có thể tác động đến chiến thuật, giá trị chuyển nhượng hay thành tích của đội bóng. Với mong muốn vận dụng những kiến thức đã học về lập trình Python vào thực tiễn, em đã thực hiện bài tập lớn với chủ đề **“Phân tích dữ liệu thống kê cầu thủ Premier League mùa giải 2024–2025”**.

Bài báo cáo này trình bày quá trình xây dựng một hệ thống thu thập và phân tích dữ liệu cầu thủ từ trang FBref.com — một trang web cung cấp số liệu thống kê bóng đá uy tín. Dữ liệu thu thập được sẽ được xử lý, phân tích thống kê cơ bản, trực quan hóa bằng biểu đồ và ứng dụng một số phương pháp học máy để dự đoán giá trị chuyển nhượng cầu thủ.

Cụ thể, bài tập được chia thành 4 bài chính:

**Bài 1:** Tự động thu thập dữ liệu thống kê cầu thủ Premier League có thời gian thi đấu trên 90 phút bằng Selenium và BeautifulSoup.

**Bài 2:** Phân tích Top 3 cầu thủ có chỉ số cao nhất và thấp nhất ở từng hạng mục, tính toán thống kê mô tả và trực quan hóa phân phối dữ liệu bằng biểu đồ.

**Bài 3:** Phân tích toàn bộ các chỉ số dạng số, xác định các chỉ số hợp lệ, tính toán thống kê và đánh giá đội bóng có phong độ tốt nhất dựa trên các chỉ số tấn công.

**Bài 4:** Thu thập dữ liệu giá trị chuyển nhượng từ các trang chuyên về bóng đá, chuẩn hóa dữ liệu và ứng dụng mô hình Random Forest để dự đoán giá trị chuyển nhượng của các cầu thủ dựa trên các chỉ số thống kê chuyên môn.

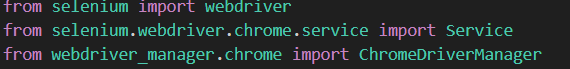
Thông qua bài tập này, em có cơ hội vận dụng kiến thức về **web scraping**, **pandas**, **matplotlib**, **machine learning** và các kỹ năng lập trình xử lý dữ liệu thực tế. Đồng thời, bài tập cũng giúp em hiểu rõ hơn về cách tổ chức và phân tích dữ liệu thể thao một cách khoa học, hỗ trợ cho việc ra quyết định trong bóng đá hiện đại.

Em xin chân thành cảm ơn thầy **Kim Ngọc Bách** đã tận tình hướng dẫn và tạo điều kiện để em thực hiện bài tập này.

# Bài 1: Lấy dữ liệu cầu thủ từ trang [FBref](https://fbref.com/en/" \t "_new)

**Bước 1: Khởi động trình duyệt và truy cập trang web**

- Đầu tiên cần cài đặt thư viện selenium và khởi tạo trình duyệt tự động Chrome. Ở đây em dùng ChromeDriverManager để tự động tải đúng version trình điều khiển phù hợp với máy.





-> Lệnh này khởi động một trình duyệt Chrome mới bằng Selenium

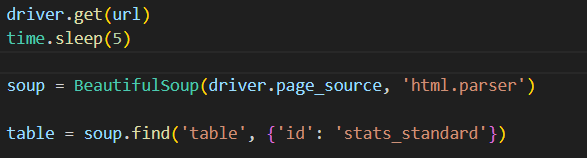
Tiếp theo em gán địa chỉ trang web Premier League Stats vào biến url:



-> Đây là trang chứa bảng thống kê mình cần thu thập.

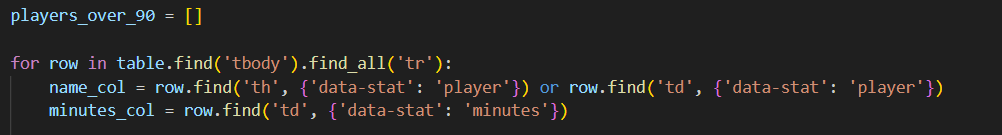
**Bước 2: Truy cập trang và lấy HTML**

- Để trình duyệt tự động truy cập trang web và lấy nội dung HTML sau khi trang đã load em sử dụng lệnh driver.get(url) để truy cập trang web. Sau đó em dùng BeautifulSoup để trích xuất source HTML của trang. Cuối cùng em tạo biến table để tìm bảng thống kê có id: stats\_standard.



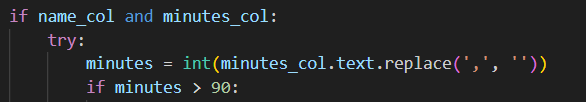
**Bước 3: Duyệt từng hàng cầu thủ trong hàng.**

- Duyệt qua từng hàng <tr> trong hàng, lấy thông tin tên cầu thủ và số phút thi đấu. Em dùng .find(‘tbody’).find\_all(‘tr’) để lấy từng dòng dữ liệu. Sau đó lấy tên cầu thủ ở cột data-stat=’player’ và lấy số phút ở cột data-stats=’minutes’



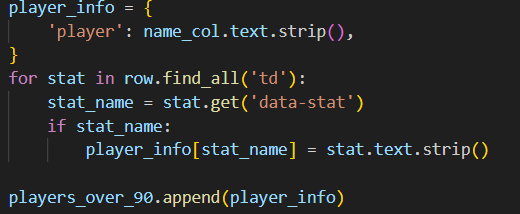
**Bước 4: Lọc cầu thủ thi đấu trên 90 phút.**

- Chỉ giữ lại những cầu thủ có thời gian thi đấu trên 90 phút. Phần này em sẽ kiểm tra nếu có dữ liệu tên và phút thì chuyển giá trị phút từ string sang int và so sánh điều kiện thi đấu trên 90 phút.



**Bước 5: Lấy toàn bộ chỉ số chuyên môn của cầu thủ.**

- Nếu cầu thủ đủ điều kiện, tiếp tục lấy toàn bộ chỉ số thống kê khác của cầu thủ đó. Tạo dictionary player\_info sau đó duyệt tất cả các cột <td> còn lại. Gán tên chỉ số và giá trị vào dictionary. Thêm cầu thủ vào danh sách player\_over\_90.



**Bước 6: Tạo DataFrame và đổi tên cột.**

- Sau khi duyệt xong tất cả cầu thủ, chuyển danh sách player\_over\_90 sang DataFrame để dễ xử lý tiếp.



-> Tạo DataFrame từ list các dictionary.

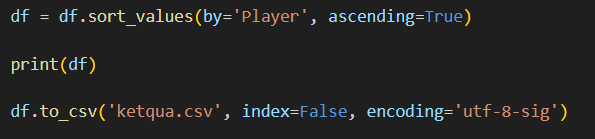
- Tiếp theo em chuẩn hoá tên các cột về dạng quen thuộc để dễ phân tích.



-> Đổi tên các cột quan trọng về đúng định dạng mình muốn.

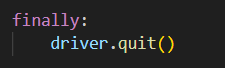
**Bước 7: Sắp xếp dữ liệu và lưu vào CSV.**

- Để dễ đọc và xử lý tiếp, em sắp xếp danh sách cầu thủ theo tên và lưu vào file CSV. Trong file em sắp xếp theo tên cầu thủ và lưu dữ liệu ra file ketqua.csv(dùng utf-8-sig để đọc tiếng Việt chuẩn).



**Bước 8: Đóng trình duyệt.**

- Tắt trình duyệt tự động sau khi hoàn thành nhiệm vụ.



# Bài 2: Phân tích top cầu thủ, thống kê mô tả và trực quan hoá.

**Bước 1: Thiết lập các hằng số và đọc dữ liệu.**

- Khai báo các hằng số tên file, số lượng cầu thủ top cần tìm, các chỉ số cần phân tích và đọc dữ liệu từ file CSV kết quả của Bài 1.

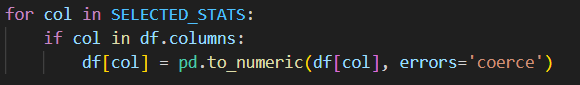






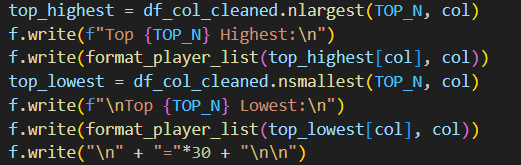
**Bước 2: Chuyển đổi dữ liệu chỉ số về dạng số.**

- Các cột chỉ số trong bảng có thể đang ở dạng string. Cần chuyển về dạng số để tính toán được. Chuyển đổi từng cột trong SELECTED\_STATS sang numeric, nếu lỗi thì trả về NaN.



**Bước 3: Tìm top 3 cầu thủ cao nhất và thấp nhất từng chỉ số.**

- Lọc ra top N cầu thủ có chỉ số cao nhất và thấp nhất cho từng thống kê đã chọn.



-> Lệnh này lấy ra top 3 cầu thủ có chỉ số cao nhất/thấp nhất trong từng cột.

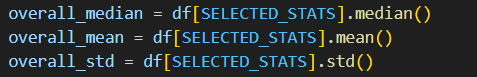
- In ra file top\_3.txt theo cầu trúc dễ đọc:



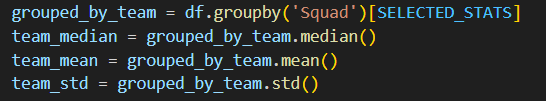
-> Ghi kết quả từng chỉ số vào file text.

**Bước 4: Tính thống kê mô tả (trung vị, trung bình, độ lệch chuẩn)**

- Tính các thông số mô tả cơ bản cho từng chỉ số: Median (Trung vị), Mean (Trung bình), Standard Deviation (Độ lệch chuẩn).



-> Tính cho toàn bộ giải.

Ta tiếp tục tính cho từng đội:  


-> Gom nhóm theo đội, tính thông số tương ứng.

**Bước 5: Lưu kết quả thống kê vào file CSV.**

- Gộp các kết quả tính toán lại thành DataFrame và lưu vào results2.csv để dùng cho bước sau hoặc tham khảo.

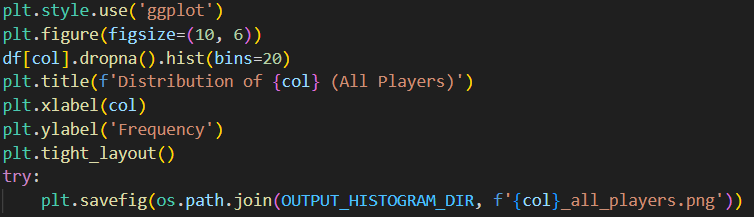


**Bước 6: Vẽ biểu đồ phân phối.**

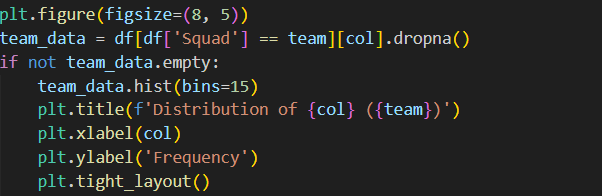
- Để trực quan hoá phân phối từng chỉ số:

+ Vẽ histogram cho toàn giải.

+ Vẽ histogram riêng cho từng đội.



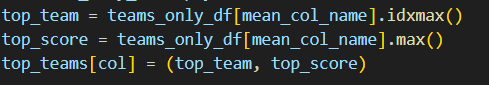
-> Tạo biểu đồ cho từng chỉ số toàn giải.

- Vẽ tiếp cho từng đội:  


-> Tạo histogram riêng cho từng đội, từng chỉ số.

**Bước 7: Xác định đội dẫn đầu từng chỉ số.**

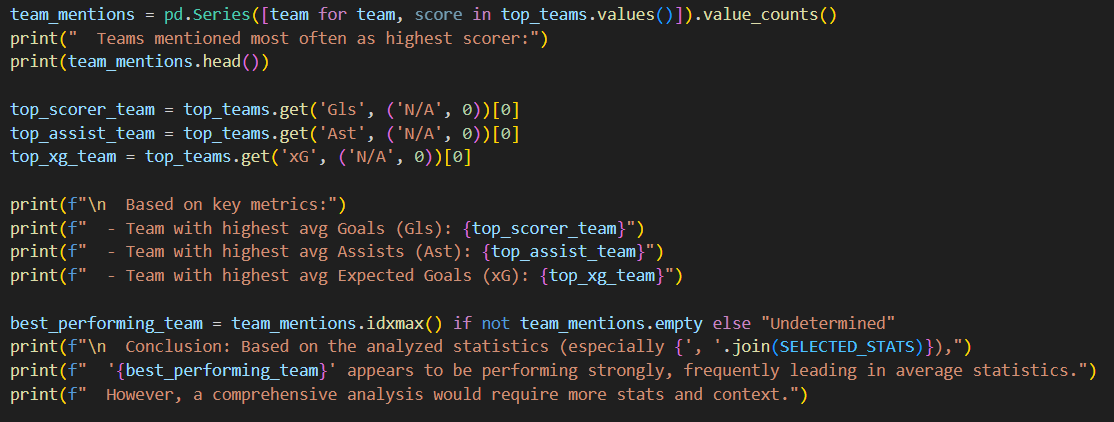
- Tìm đội có trung bình cao nhất cho từng chỉ số, để biết đội nào mạnh ở mục nào.



-> Lấy tên đội dẫn đầu và giá trị trung bình cao nhất.

**Bước 8: Đánh giá tổng thể.**

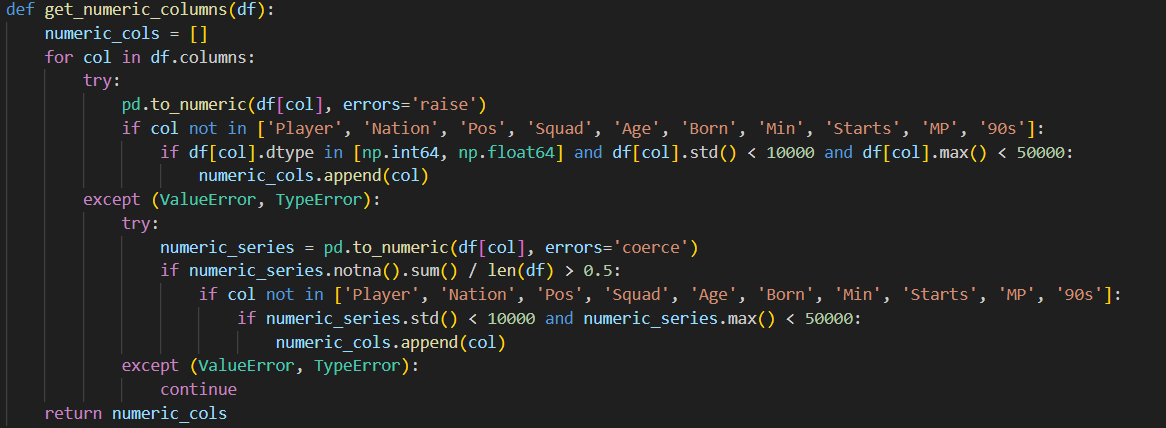
- Đếm số lần mỗi đội dẫn đầu chỉ số để xác định đội thi đấu nổi bật nhất.



# Bài 3: Phân tích toàn bộ các chỉ số số học.

**Bước 1: Đọc dữ liệu và xác định cột số hợp lệ.**

- Đọc file ketqua.csv chữa dữ liệu cầu thủ và xác định những cột nào thực sự là số (loại bỏ các cột chữ như Player, Squad,…)







-> Hàm này kiểm tra từng cột, nếu chuyển hết được giá trị sang số mà giá trị không quá lớn thì giữ lại.

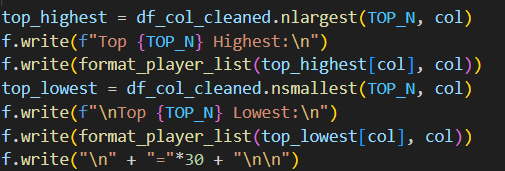
**Bước 2: Chuyển các cột số về numeric.**

- Đảm bảo các cột số được chuyển hoàn toàn sang kiểu số (float/int) để tránh lỗi tính toán.



**Bước 3: Tìm top 3 cao nhất và thấp nhất mỗi chỉ số.**

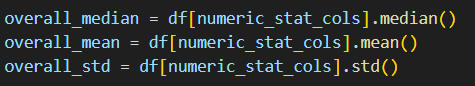
- Với mỗi cột số, tìm top N cầu thủ có chỉ số cao nhất và thấp nhất, ghi vào file top\_3.txt.

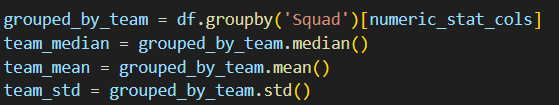


-> Tìm top 3 cao nhất và thấp nhất từng cột số.

**Bước 4: Tính thống kê mô tả.**

- Tính các thông số mô tả cơ bản: Median, Mean, Standard Deviation.

Cho toàn bộ cầu thủ:  


Cho từng đội bóng:  


**Bước 5: Xuất kết quả thống kê ra file.**

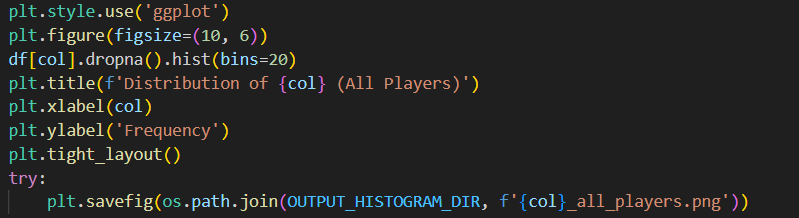
- Tổng hợp các giá trị thống kê lại thành một bảng DataFrame và lưu vào file results2.csv.



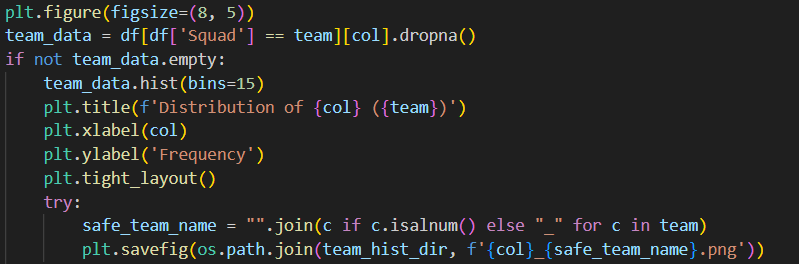
**Bước 6: Vẽ biểu đồ phân phối.**

- Vẽ histogram để trực quan hoá phân phối từng chỉ số toàn giải và từng đội.

Toàn giải:

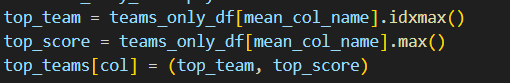


Từng đội:



**Bước 7: Tìm đội dẫn đầu từng chỉ số tấn công quan trọng.**

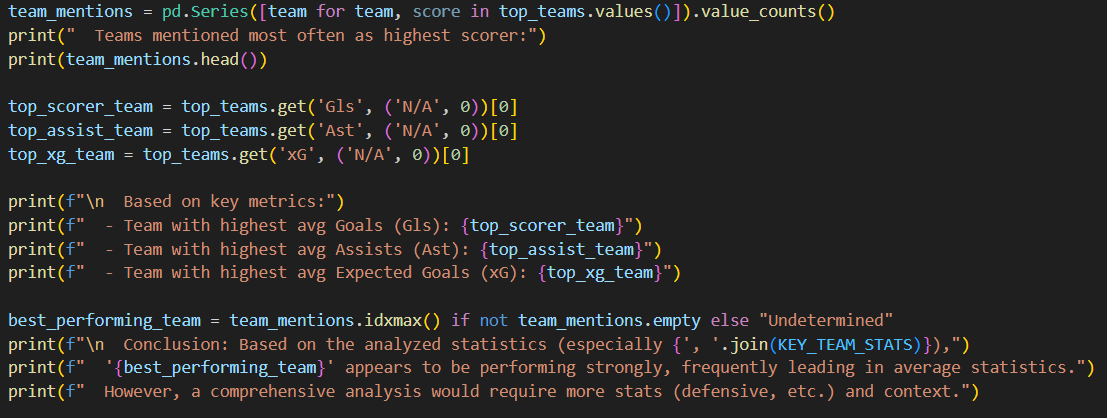
- Với các chỉ số tấn công chính (Gls, Ast, G+A, xG, xAG, PrgC, PrgP), tìm đội có trung bình cao nhất.



-> Xác định đội dẫn đầu ở mỗi chỉ số.

**Bước 8: Xác định đội thi đấu nổi bật nhất.**

- Đếm số lần từng đội dẫn đầu các chỉ số để xác định đội mạnh nhất toàn giải.

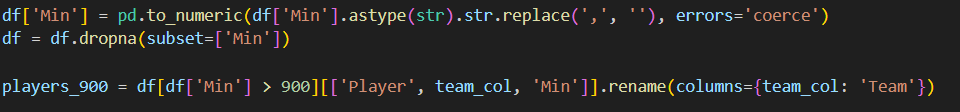


# Bài 4: Dự đoán giá trị chuyển nhượng cầu thủ bằng Machine Learning.

**Bước 1: Đọc dữ liệu và lọc cầu thủ thi đấu trên 900 phút.**

- Đọc dữ liệu từ file ketqua.csv, chuyển số phút thành số thực và chỉ giữ cầu thủ có số phút thi đấu lớn hơn 900.





-> Chỉ lọc những cầu thủ đáng tin cậy cho dự đoán (thi đấu đủ nhiều).

**Bước 2: Lấy giá trị chuyển nhượng của từng cầu thủ.**

- Dùng Selenium vào Transfermarkt và FootballTransfers để lấy giá trị chuyển nhượng. Nếu trang đầu tiên lỗi, chuyển sang trang thứ hai. Nếu cả hai thất bại thì gán ‘N/a’.



-> Crawl dữ liệu web theo từng cầu thủ.

Kết quả được lưu vào file transfer\_values.csv

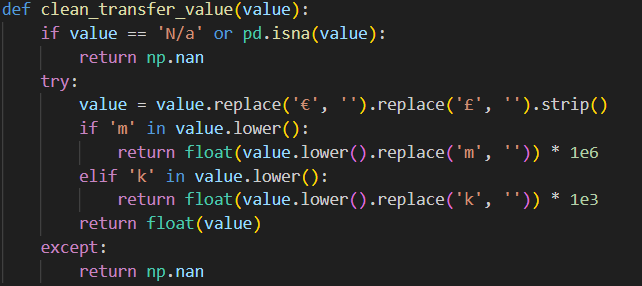
**Bước 3: Gộp dữ liệu giá trị với bảng cầu thủ.**

- Ghép giá trị chuyển nhượng lấy được với danh sách cầu thủ theo tên.



**Bước 4: Chuyển giá trị tiền về dạng số.**

- Giá trị chuyển nhượng thu được có thể ở dạng “€35m” hoặc “£15k”, cần chuyển thành số.



-> Chuyển hoá giá trị về dạng số để đưa vào model.

**Bước 5: Xác định các đặc trưng (feature) cho dự đoán.**

- Chọn các chỉ số thống kê chuyên môn phù hợp để làm input cho moddel dự đoán giá trị cầu thủ.



Và lọc những cột có đủ dữ liệu:  


**Bước 6: Chuẩn hoá dữ liệu.**

- Dữ liệu số học phải chuẩn hoá về cùng thang đo (trung bình = 0, độ lệch chuẩn = 1) để model hoạt động ổn định hơn.



**Bước 7: Chia dữ liệu train - test.**

- Chia ngẫu nhiên dữ liệu thành 80% train và 20% test để kiểm tra độ chính xác dự đoán.



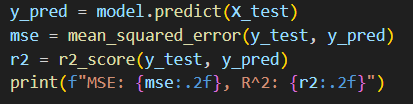
**Bước 8: Huấn luyện mô hình Ramdom Forest.**

- Dùng RamdomForestRegressor - một thuật toán machine learning mạnh, ít cần tinh chỉnh, dự đoán tốt với dữ liệu nhỏ và vừa.



**Bước 9: Dự đoán và tính sai số.**

- Dùng model để dự đoán giá trị cầu thủ, kiểm tra độ chính xác bằng sai số trung bình MSE và hệ số R².



**Bước 10: Dự đoán toàn bộ và lưu kết quả.**

- Dự đoán giá trị chuyển nhượng cho toàn bộ cầu thủ và lưu vào file transfer\_predictions.csv.

