# ĐỒ ÁN II

Phạm Phước Bảo Tín and Trịnh Quốc Dân

April 2025

## 1 Tên dataset

Premier League Player Stats (2015-2025)
Link mã nguồn để crawl dataset: Sofascore Football Premier league Crawl[1].

## 2 Mô tả dataset

### 2.1 Giới thiệu Dataset

Dataset "Premier League Player Stats (2015-2025)"được crawl từ trang Sofascore[2] tại thời điểm ngày 2-4-2025, chứa thông tin thống kê về hiệu suất thi đấu của các cầu thủ tại giải Ngoại hạng Anh qua các mùa giải 2015-2025. Dữ liệu bao gồm số bàn thắng, số pha rê bóng thành công, số pha tắc bóng, kiến tạo, tỉ lệ chuyền chính xác và điểm trung bình trên thang điểm mười theo SofaScore. Riêng mùa giải **24-25** chưa kết thúc nên cập nhật đến vòng đấu thứ 30[3].

### 2.2 Cấu trúc Dataset

Dataset gồm 9 cột:

- **Mùa giải**: Mùa bóng mà dữ liệu được ghi nhận (VD: 24/25 cho mùa 2024-2025).
- **Tên đội**: Đội bóng mà cầu thủ đang thi đấu cho mùa giải tương ứng.
- **Tên cầu thủ**: Tên của cầu thủ trong mùa giải tương ứng.
- **Số bàn thắng**: Tổng số bàn thắng ghi được trong mùa giải.
- **Số pha rê bóng thành công**: Số lần rê bóng thành công của cầu thủ.
- **Số pha tắc bóng**: Số lần tắc bóng thành công của cầu thủ.
- **Số pha kiến tạo**: Số lần chuyền bóng thành công giúp đồng đội ghi bàn.
- **Tỉ lệ chuyền chính xác (%)**: Tỉ lệ đường chuyền thành công của cầu thủ trong mùa giải.
- **Điểm TB (SofaScore)**: Điểm trung bình cầu thủ đạt được theo đánh giá của SofaScore.

## 3 Link Dataset

Link Dataset: Premier League Player Stats (2015-2025)[4]. Dataset này được lưu dưới định dạng CSV, có thể tải về và xử lý dễ dàng bằng Python, R hoặc Excel. Dung lượng khoảng 300KB.

[1] Nguồn mã: `https://github.com/BAOTIN2004/sofascore-football-crawler`
[2] Trang web: `https://www.sofascore.com`
[3] Dữ liệu cập nhật theo từng vòng đấu tại: `https://www.sofascore.com/tournament/football/england/premier-league/17#id:61627`
[4] Link tải dataset: `https://drive.google.com/file/d/1H3tkR8zTUWfr2VEJ_4Ib81GujOTRr9NY/view`

# 4 Phương pháp thu thập dữ liệu

## 4.1 Phương pháp thu thập dữ liệu

Dữ liệu được thu thập từ Sofascore[5] thông qua **Web Scraping** bằng **Selenium** và **BeautifulSoup**.

- **Selenium**: Điều hướng trang web, mô phỏng hành vi người dùng.
- **BeautifulSoup**: Phân tích và trích xuất dữ liệu từ HTML.
- **Pandas**: Lưu dữ liệu vào DataFrame, xuất CSV.

## 4.2 Các bước tổng quát

### 4.2.1 Truy cập website và chờ trang tải

- Dùng **Selenium** để mở trang Premier League 2024/2025[6].
- Chờ trang tải hoàn toàn.

### 4.2.2 Lấy danh sách mùa giải

- Trích xuất danh sách mùa giải.
- Mở dropdown chọn mùa giải.

### 4.2.3 Xác định và thu thập dữ liệu

- Cuộn trang đến vị trí bảng dữ liệu.
- Dùng **BeautifulSoup** để lấy HTML trang hiện tại.
- Xác định vị trí các cột cần lấy.
- Trích xuất cầu thủ, đội, bàn thắng, rê bóng, tắc bóng, kiến tạo, chuyền chính xác, điểm SofaScore.

### 4.2.4 Điều hướng giữa các trang

- Tìm nút chuyển trang.
- Nếu có thể click, chuyển sang trang tiếp theo.
- Nếu bị vô hiệu hóa → kết thúc mùa giải hiện tại.

### 4.2.5 Lưu dữ liệu và xuất CSV

- Dữ liệu được lưu vào DataFrame với cột tiêu đề có thể dịch sang tiếng Việt.
- Xuất ra CSV.

## 4.3 Khó khăn gặp phải và cách giải quyết trong quá trình thực hiện

1. Website tải dữ liệu động:

- **Vấn đề**: Sofascore không tải toàn bộ dữ liệu ngay từ đầu mà tải dần khi người dùng tương tác.
- **Giải pháp**: Cuộn trang xuống và chờ tải dữ liệu (window.scrollBy, WebDriverWait).

2. Xác định cột dữ liệu không cố định:

- **Vấn đề**: Ban đầu sử dụng chỉ số cố định (cols[1], cols[2],...), nhưng 2-3 mùa gần đây có thêm cột "bàn thắng kỳ vọng".
- **Giải pháp**: Xác định cột bằng tiêu đề thay vì chỉ số. Ví dụ: `headers.index("Goals")`.

[5]Trang web: `https://www.sofascore.com`

[6]Trang thống kê giải đấu theo mùa: `https://www.sofascore.com/tournament/football/england/premier-league/17#id:61627`