ĐẠI HỌC QUỐC GIA TP.HCM

Trường Đại học Khoa học Tự nhiên

Báo cáo tiến độ đồ án

Thành viên nhóm:

Trương Tiến Đạt - 22120060

Phạm Văn Hoàng Nam - 22120220

Tống Trọng Tâm - 22120322

Vũ Châu Minh Trí - 22120456

Học phần: Nhập môn Khoa học dữ liệu

GV Lý thuyết: Lê Ngọc Thành

GV Hướng dẫn Thực hành: Lê Nhực Nam

TP.HCM - 2024

Mục lục

Chương 1

Phân công công việc

MSSV	Họ và tên	Phân công
22120060	Trương Tiến Đạt	Tiền xử lý, khám phá và trực quan hóa dữ liệu
22120220	Phạm Văn Hoàng Nam	Cào và tổng hợp dữ liệu, làm sạch dữ liệu
22120322	Tống Trọng Tâm	Cài đặt và đánh giá mô hình dự đoán kết quả
22120456	Vũ Châu Minh Trí	Cài đặt và đánh giá mô hình dự đoán kết quả

Chương 2

Kế hoạch thực hiện (dự kiến)

Giai đoạn	Công việc	Thời gian	Kỳ vọng
1. Thu thập và khám phá dữ liệu	Thu thập dữ liệu	3/11 - 10/11	Thu thập được dữ liệu các thống kê về toàn bộ trận đấu tại Ngoại hạng Anh từ mùa giải 2017/2018 đến hiện nay (2760 trận)
	Tiền xử lý dữ liệu	10/11 - 17/11	Làm sạch dữ liệu
2. Tiền xử lý dữ liệu, thử nghiệm với mô hình đơn giản để đánh giá	Khám phá dữ liệu, trực quan hóa dữ liệu	17/11 - 24/11	Hiểu rõ đặc điểm, mối quan hệ và xu hướng trong dữ liệu
	Thử áp dụng dự đoán bằng mô hình đơn giản (Random Forest)	17/11 - 24/11	Đánh giá nhanh tính khả thi của dữ liệu trong việc dự đoán kết quả, rút ra được việc cần làm ở giai đoạn tiếp theo

3. Thu thập thêm dữ liệu	Thu thập thêm dữ liệu cần thiết	24/11 - 2/12	Hoàn thiện bộ dữ liệu cần thiết cho việc xây dựng mô hình cuối cùng
4. Xây dựng mô hình và đánh giá	Dùng các mô hình phức tạp hơn, dựa trên nguồn dữ liệu tổng hợp từ hai lần thu thập dữ liệu	2/12 - 10/12	Xây dựng được hệ thống có khả năng dự đoán kết quả các trận bóng tại Ngoại hạng Anh với độ chính xác cao
5. Hoàn thiện	Kiểm tra và hoàn thiện project, viết báo cáo	11/12 - 14/12	Hoàn thiện project

Chương 3

Tổng quan nội dung giữa kì

3.1. Chủ đề: Dự đoán bóng đá

- Bóng đá là môn thể thao phổ biến nhất trên thế giới, với hàng tỷ người hâm mộ và tác động sâu rộng đến văn hóa, kinh tế và xã hội. Các giải đấu lớn như World Cup, Champions League, và đặc biệt là Ngoại hạng Anh luôn thu hút sự chú ý đặc biệt từ khán giả toàn cầu.
- Ngoại hạng Anh được đánh giá là giải đấu hấp dẫn nhất thế giới nhờ tính cạnh tranh cao, sự góp mặt của những đội bóng hàng đầu và những cầu thủ xuất sắc nhất. Kết quả các trận đấu không chỉ mang ý nghĩa giải trí mà còn cung cấp cơ sở cho các phân tích chiến thuật, quản lý đội bóng và nhiều ứng dụng khác.

3.2. Giới thiệu:

- Hệ thống này tập trung vào việc dự đoán kết quả các trận đấu tại Ngoại hạng Anh (thắng/hòa/thua) bằng cách sử dụng các mô hình học máy, học sâu.
- Trước mắt đang thử nghiệm mô hình `Random Forest` để kiểm tra khả năng dự đoán chính xác.

3.3. Dữ liệu:

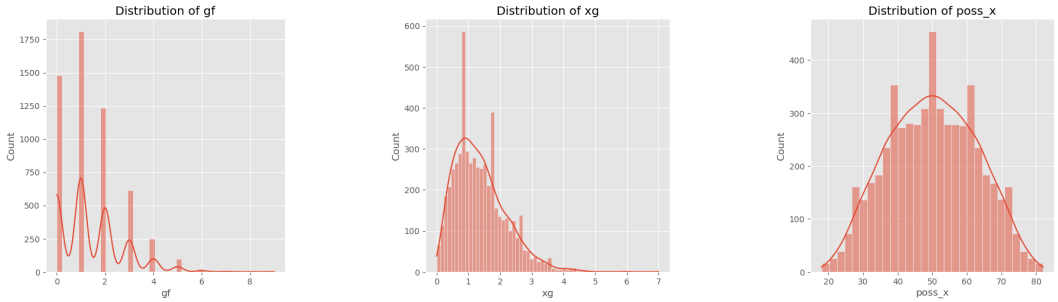
- Dữ liệu được thu thập từ toàn bộ các trận đấu tại Ngoại hạng Anh trong giai đoạn từ mùa giải 2017/2018 đến hết vòng 10 của mùa giải 2024/2025, với tổng cộng 5520 dòng dữ liệu.
- Dữ liệu được tổng hợp từ hai nguồn đáng tin cậy: https://fbref.com và https://www.football-data.co.uk.

- Hai nguồn dữ liệu này được lựa chọn vì đảm bảo tính cập nhật, độ chính xác cao, và cung cấp đầy đủ các thông tin quan trọng như kết quả trận đấu, thông số thống kê chi tiết, cũng như lịch sử đối đầu.

3.4. Trực quan hóa dữ liệu

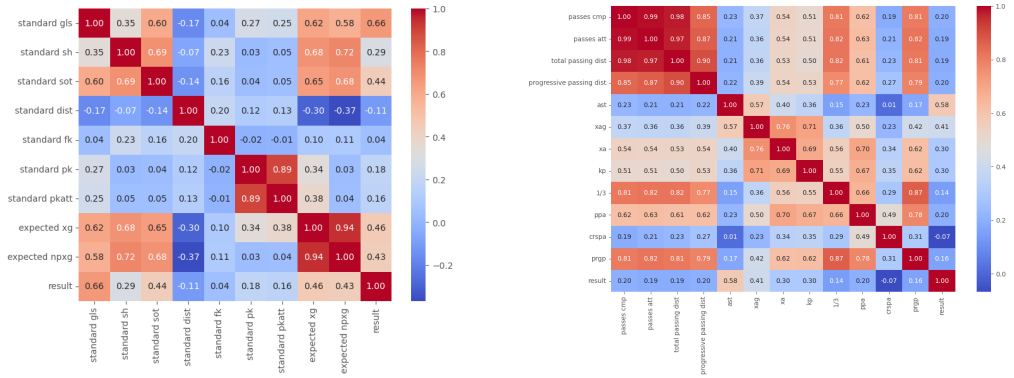
- Mục tiêu của việc trực quan hóa là khám phá các đặc điểm quan trọng trong dữ liệu, hiểu rõ mối quan hệ giữa các đặc trưng và kết quả trận đấu, từ đó hỗ trợ xây dựng và tối ưu hóa mô hình dự đoán.

- Phân phối của một số đặc trưng như: số bàn thắng ghi được (`GF`), bàn thắng kỳ vọng (`xG`) và tỷ lệ kiểm soát bóng (`Possession`).



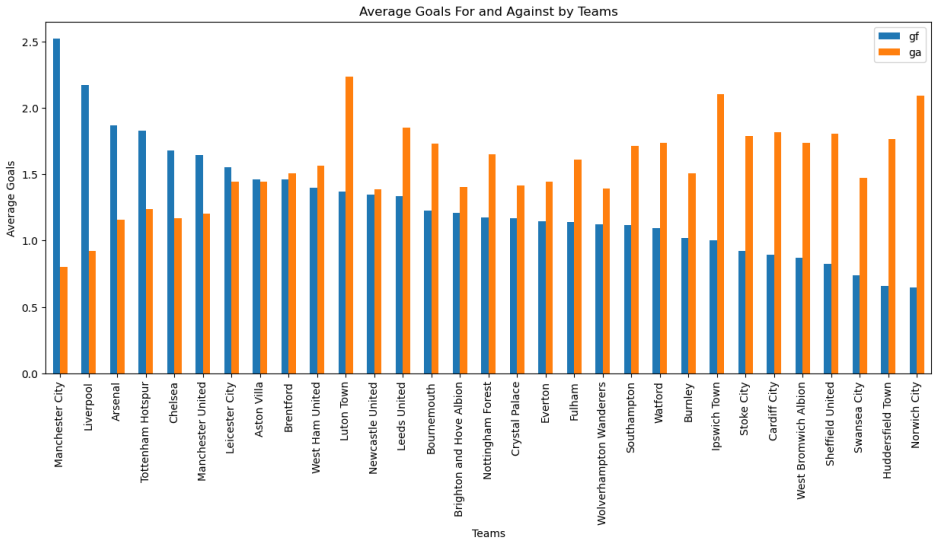
Hình 3.1: Một số phân phối của các đặc trưng

- Ma trận tương quan để đánh giá mối quan hệ giữa các đặc trưng (ví dụ: `shooting`, `passing`) và kết quả trận đấu (`results`).



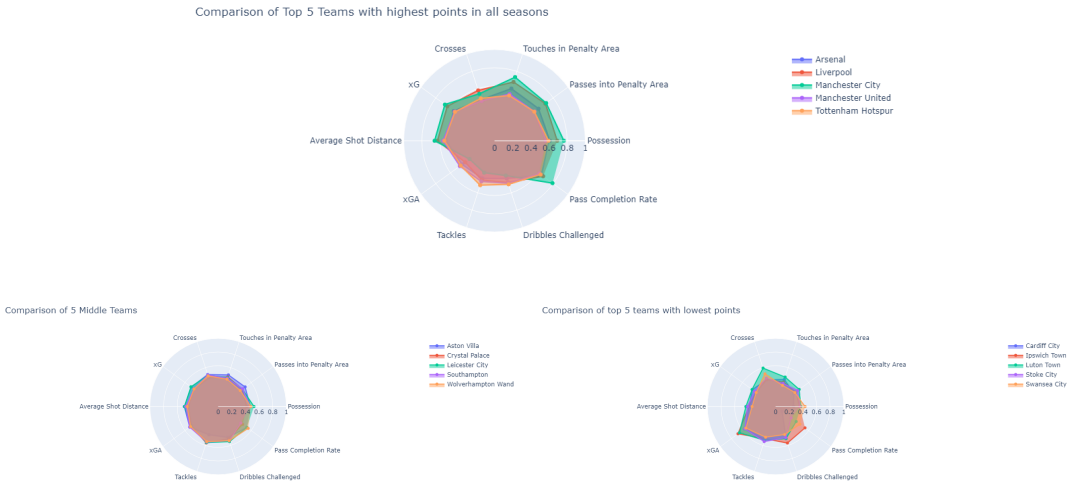
Hình 3.2: Ma trận tương quan của một số nhóm đặc trưng với kết quả trận đấu

- Trực quan hóa cụ thể một số đặc trưng:
  - Trung bình bàn thắng kỳ vọng (`xG`) cho từng đội qua các mùa giải.



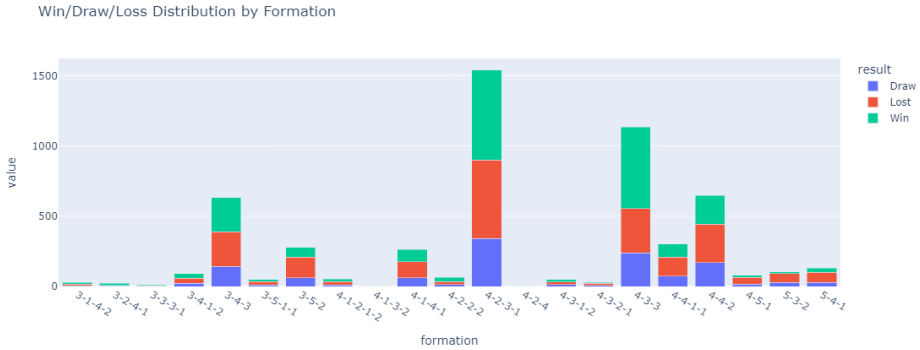
Hình 3.3: Trung bình bàn thắng và bàn thua của mỗi đội

– So sánh các nhóm đội bóng (mạnh - trung bình - yếu) ở một vài chỉ số để rút ra mối liên hệ giữa sức mạnh đội bóng và lối chơi (chẳng hạn như các đội càng mạnh thì các chỉ số về bàn thắng, tấn công và kiểm soát bóng có xu hướng tăng còn các đội càng yếu thì các chỉ số về phòng ngự, bàn thua càng tăng).



Hình 3.4: Khác biệt giữa các nhóm đội bóng ở những chỉ số cơ bản

– Phân bố kết quả trận đấu (thắng/hòa/thua) dựa trên sơ đồ chiến thuật.



Hình 3.5: Sự ảnh hưởng của sơ đồ chiến thuật đến kết quả trận đấu

3.5. Tiền xử lý và chuẩn bị dữ liệu

3.5.1. Mã hóa các đặc trưng dạng category

Các đặc trưng như đội bóng (`team`), sơ đồ chiến thuật (`formation`), đội trưởng (`captain`), và trọng tài (`referee`) được mã hóa thành các giá trị số để sử dụng trong mô hình. Đây là các thông tin luôn có trước khi trận đấu diễn ra, đảm bảo tính khả dụng cho dự đoán.

3.5.2. Xử lý các đặc trưng dạng số

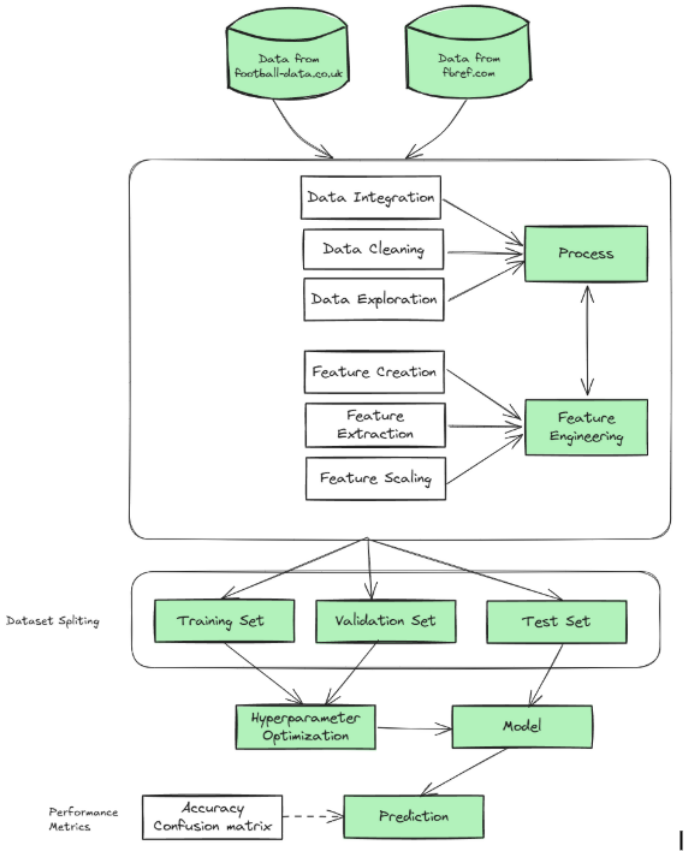
Khác với các đặc trưng trên, các đặc trưng dạng số chỉ có được sau trận đấu (vd: số bàn thắng, tỷ lệ kiểm soát bóng), các giá trị này được tính trung bình trong 4 trận gần nhất của mỗi đội để mô phỏng phong độ gần đây, tăng khả năng phản ánh thực tế trong dự đoán.

3.5.3. Chia dữ liệu thành các tập

- `Tập huấn luyện (train set)`: Bao gồm các trận đấu từ mùa giải 2017/2018 đến hết mùa 2020/2021, dùng để huấn luyện mô hình.
- `Tập kiểm định (validation set)`: Bao gồm các trận đấu trong mùa giải 2021/2022, dùng để đánh giá hiệu quả và điều chỉnh siêu tham số.
- `Tập kiểm tra (test set)`: Gồm các trận đấu từ mùa 2022/2023 đến hiện tại, dùng để kiểm tra độ chính xác và khả năng tổng quát hóa của mô hình trên dữ liệu mới.

Cách chia này đảm bảo tính liên tục theo thời gian, giúp mô hình học được xu hướng thay đổi qua từng mùa giải và giảm thiểu khả năng rò rỉ dữ liệu.

3.6. Áp dụng mô hình vào dự đoán

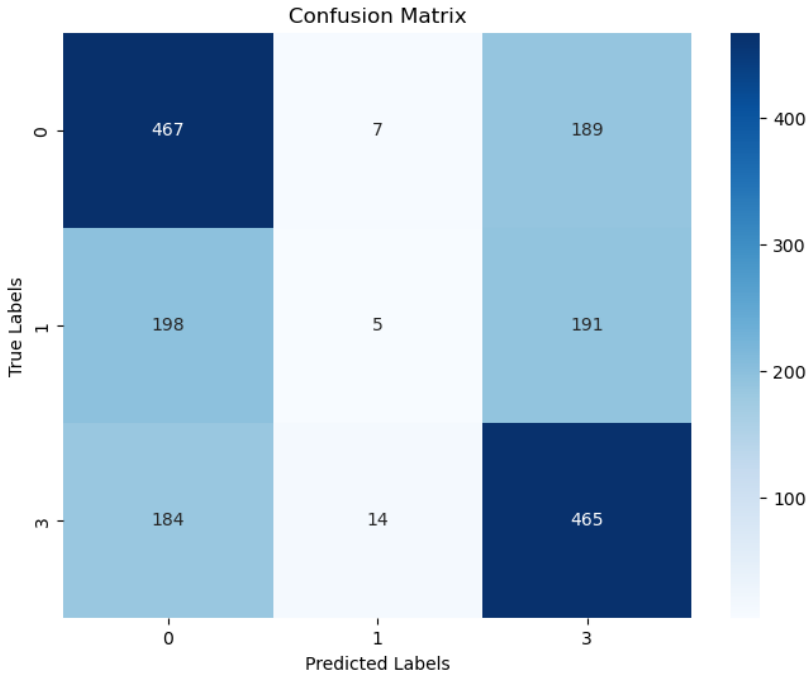


Hình 3.6: Pipeline train model

- Xây dựng mô hình: Lựa chọn các đặc trưng quan trọng có ảnh hưởng lớn đến kết quả trận đấu để tối ưu hóa hiệu suất dự đoán.
- Điều chỉnh siêu tham số: Tinh chỉnh các tham số của mô hình nhằm đạt được hiệu quả cao nhất, tối ưu độ chính xác và khả năng tổng quát hóa.

3.7. Kết quả

- `Accuracy`: 0.564767
- `Confusion matrix`



Độ chính xác của mô hình dự đoán khá thấp, cần phải xem xét tiếp tục cải thiện.

3.8. Việc cần làm

1. Áp dụng các kỹ thuật nâng cao để cải thiện hiệu suất mô hình, bao gồm tối ưu hóa siêu tham số, sử dụng `Bootstrap Sampling` và các phương pháp `Boosting`.

2. Thử nghiệm với các mô hình khác như RNN hoặc GRU để đánh giá hiệu quả so sánh với `Random Forest`.

3. Tìm kiếm và tích hợp thêm dữ liệu từ các nguồn khác để bổ sung các đặc trưng phong phú hơn, chẳng hạn như phong độ của từng cầu thủ, ảnh hưởng của huấn luyện viên, hoặc chiến thuật đội hình.

4. Tiến hành xử lý và phân tích dữ liệu mới, kiểm tra sự tương quan giữa các đặc trưng để đảm bảo tính hữu ích trong dự đoán.