TRƯỜNG ĐẠI HỌC BÁCH KHOA - ĐẠI HỌC ĐÀ NẴNG

KHOA CÔNG NGHỆ THÔNG TIN


# Dự đoán kết quả bóng đá

## Khoa học dữ liệu

**NHÓM 5**

**Nguyễn Tiến Trọng 102200291**

**Thân Văn Hồng Sơn 102200285**

# Danh sách thành viên và phân công nhiệm vụ

| Sinh viên thực hiện | Nhiệm vụ | Đánh giá |
|---|---|---|
| Nguyễn Tiến Trọng | • Tiền xử lý<br>• Phân tích đặc trưng<br>• Dự đoán bằng model Regression | • Hoàn thành<br>• Hoàn thành<br>• Hoàn thành |
| Thân Văn Hồng Sơn | • Crawl dữ liệu<br>• Làm sạch dữ liệu<br>• Dự đoán bằng model Classification | • Hoàn thành<br>• Hoàn thành<br>• Hoàn thành |

# Mục tiêu:

Xây dựng mô hình dự đoán kết quả bóng đá

# Giải pháp:

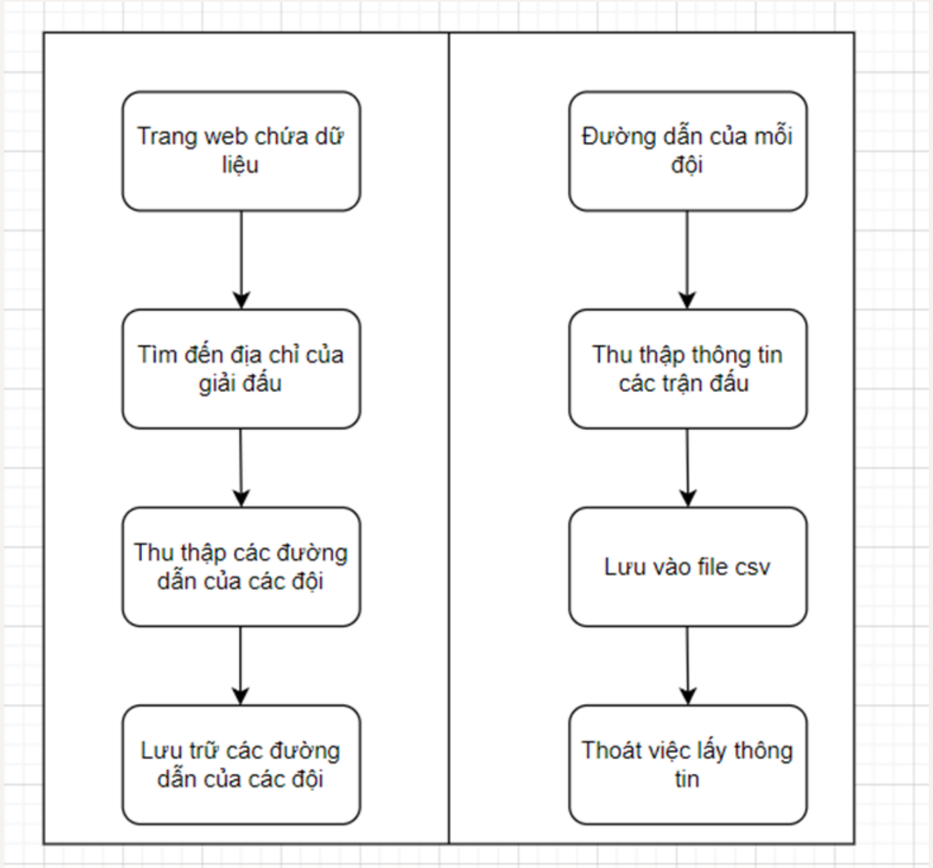
Ứng dụng kiến thức đã được học về khoa học dữ liệu để sử dụng các mô hình học máy phù hợp từ đó huấn luyện rồi dự đoán.

# Nội dung

1. Thu thập dữ liệu
2. Trích xuất đặc trưng
3. Mô hình dự đoán
4. Kết luận và hướng phát triển

# I. Thu thập dữ liệu

Dữ liệu được thu thập từ nguồn: https://fbref.com/en/comps




Thu thập dữ liệu

| | So_Cot | Kieu_Du_Lieu | So_Gia_Tri_Null |
|---|---|---|---|
| date | 1496 | datetime64[ns] | 0 |
| time | 1496 | object | 0 |
| comp | 1496 | object | 0 |
| round | 1496 | object | 0 |
| day | 1496 | object | 0 |
| venue | 1496 | object | 0 |
| result | 1496 | object | 0 |
| gf | 1496 | float64 | 0 |
| ga | 1496 | float64 | 0 |
| opponent | 1496 | object | 0 |
| xg | 1496 | float64 | 0 |
| xga | 1496 | float64 | 0 |
| poss | 1496 | float64 | 0 |
| captain | 1496 | object | 0 |
| formation | 1496 | object | 0 |
| referee | 1496 | object | 0 |
| sh | 1496 | float64 | 0 |
| sot | 1496 | float64 | 0 |
| dist | 1496 | float64 | 0 |
| fk | 1496 | float64 | 0 |
| pk | 1496 | float64 | 0 |
| pkatt | 1496 | float64 | 0 |
| season | 1496 | int64 | 0 |
| team | 1496 | object | 0 |
| target | 1496 | int64 | 0 |
| venue_code | 1496 | int64 | 0 |
| opp_code | 1496 | int64 | 0 |
| hour | 1496 | int64 | 0 |
| day_code | 1496 | int64 | 0 |

Số lượng dữ liệu

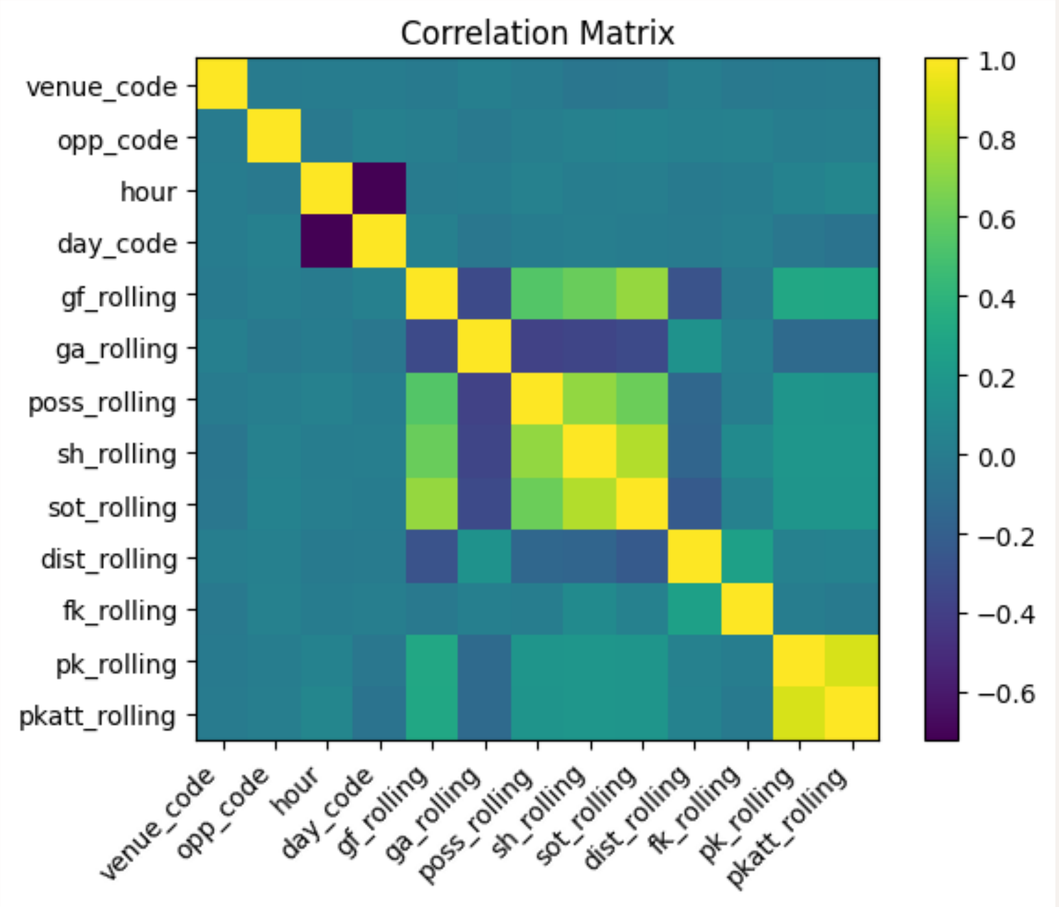
# II. Trích xuất đặc trưng

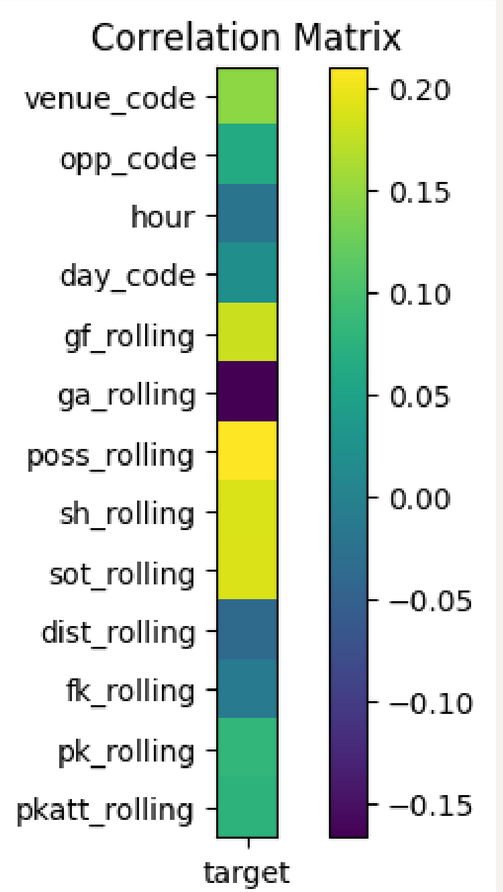
**Tính toán phong độ 5-6 trận gần nhất của đội bóng tại thời điểm đang xét**
**Những đặc trưng dùng để thống kê phong độ bao gồm:**

- **Số bàn thắng ghi được**
- **Số bàn thua**
- **Tỉ lệ kiểm soát bóng**
- **Số cú sút tạo ra**
- **Số cú sút trúng đích tạo ra**
- **Số quả phạt đền được thực hiện**
- **Số quả phạt đền được thực hiện thành công**

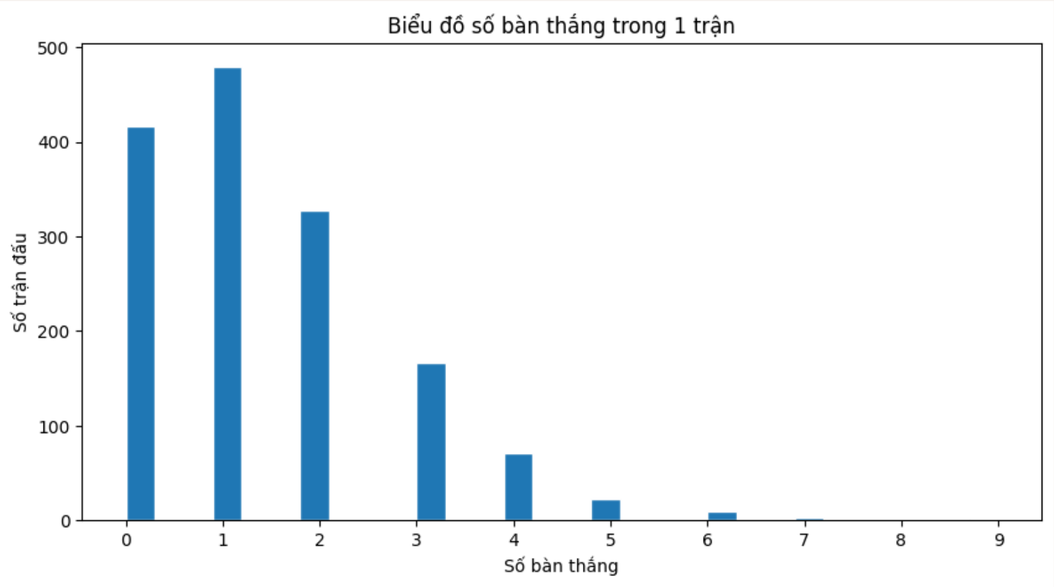
# II. Trích xuất đặc trưng

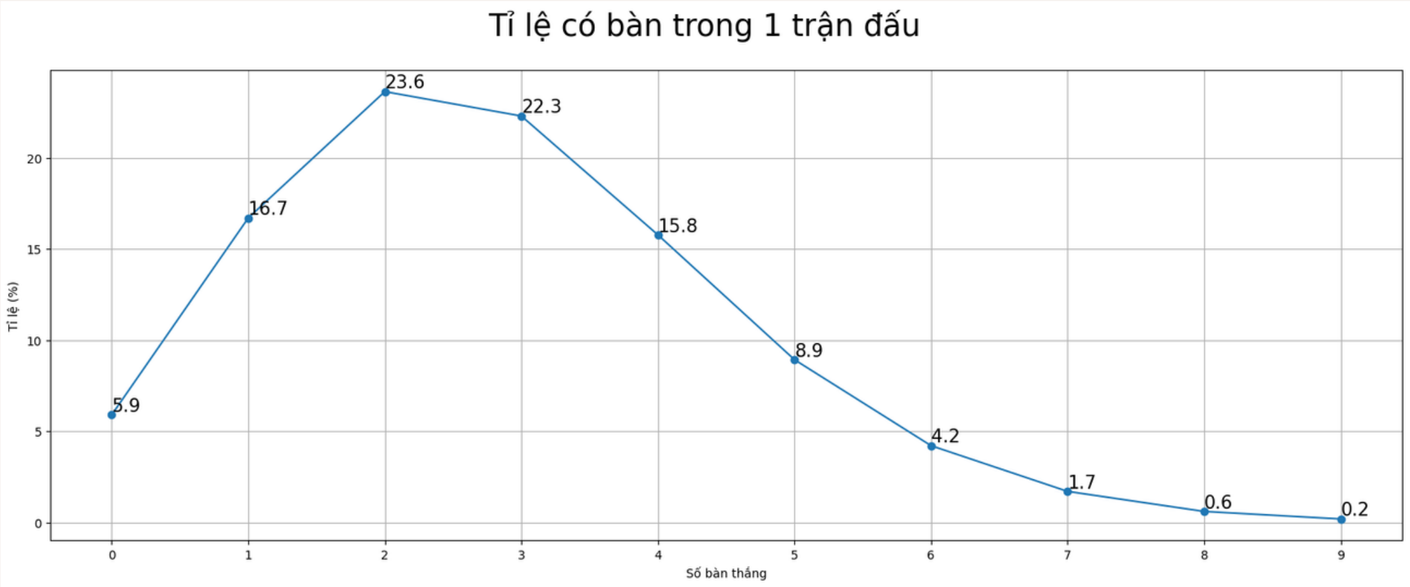
**Chúng ta sẽ lựa chọn những đặc trưng có sự tương quan lớn với kết quả của trận đấu (target)**

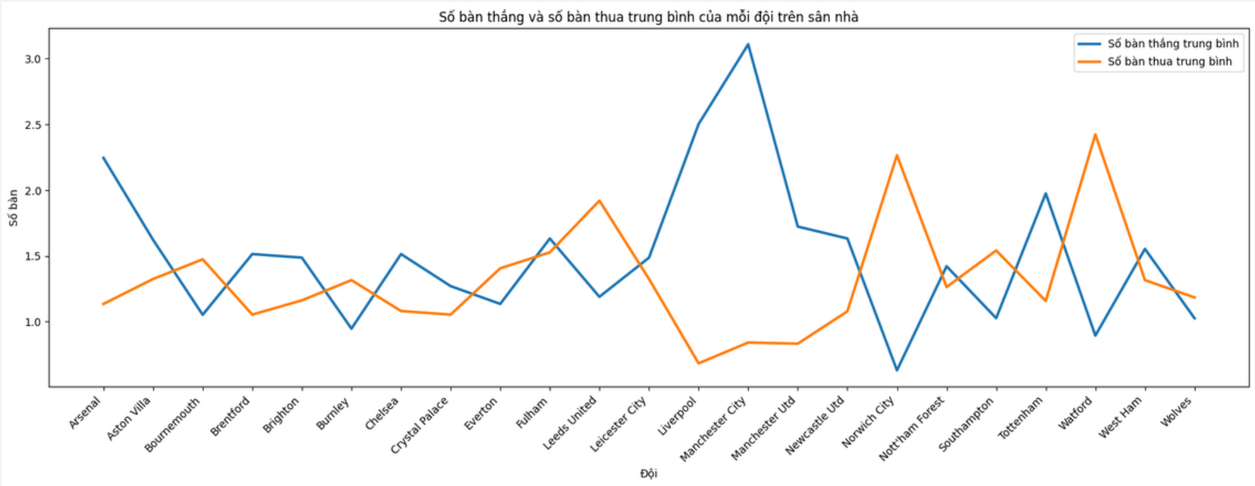


# II. Trích xuất đặc trưng



# II. Xử lí đặc trưng

**Tiến hành xử lí đặc trưng bằng các phương pháp:**

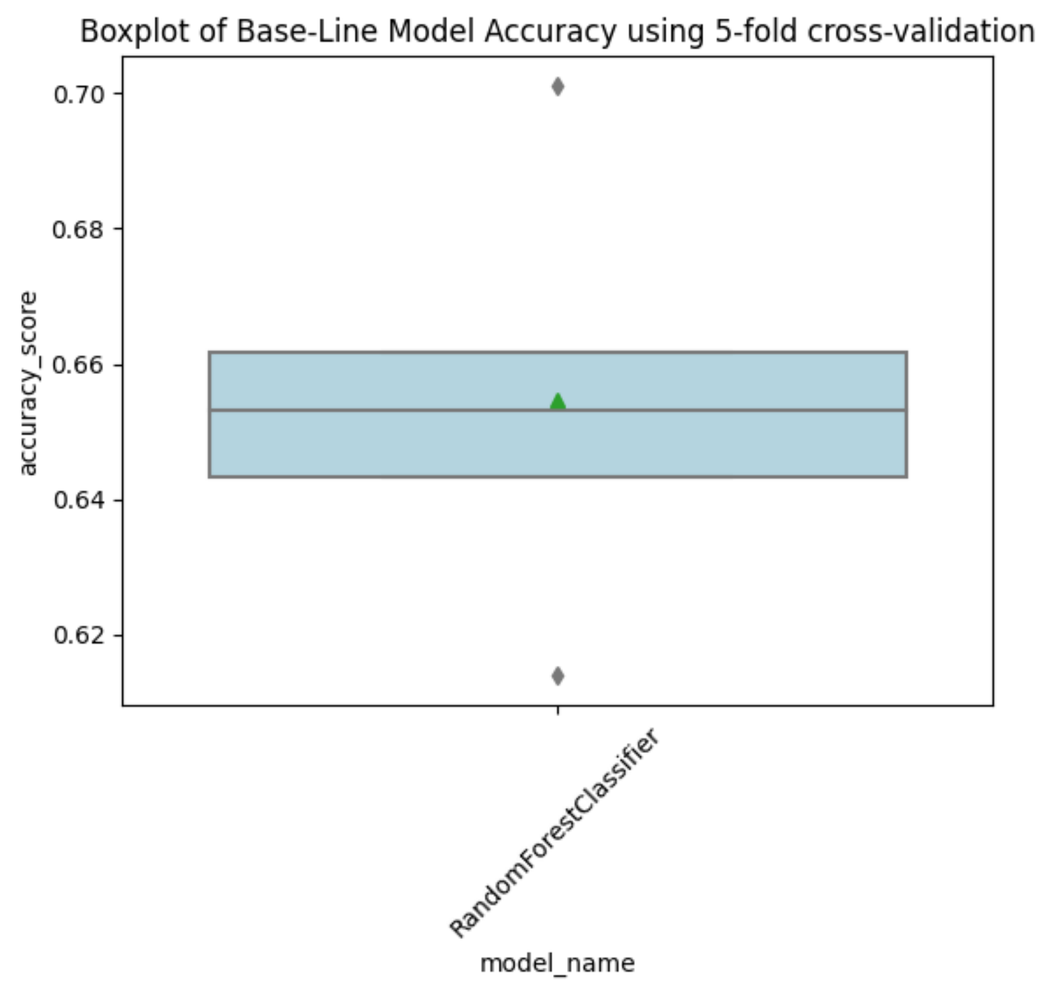
- **Outlier: gaussian, skewed**
- **Transform: Normalize**
- **Feature Selection: SelectKBest, VarianceThresold**

**Kết quả thu được không làm tăng tỉ lệ chính xác của mô hình**

# III. Mô hình dự đoán

- Nhóm đã sử dụng 2 mô hình là Random Forest Classification và Poisson Regression để dự đoán kết quả bóng đá.
- Trong mô hình nhóm sử dụng metrics accruracy score và mean squared error
- Thử nghiệm dự đoán với 1 số đặc trưng quan trọng của mô hình

# III. Mô hình dự đoán(RandomForest)




| | Mean | Standard Deviation |
|---|---|---|
| model_name | | |
| RandomForestClassifier | 0.654672 | 0.031595 |

**Siêu tham số:**

- **n_estimators=50**
- **min_samples_split=10**

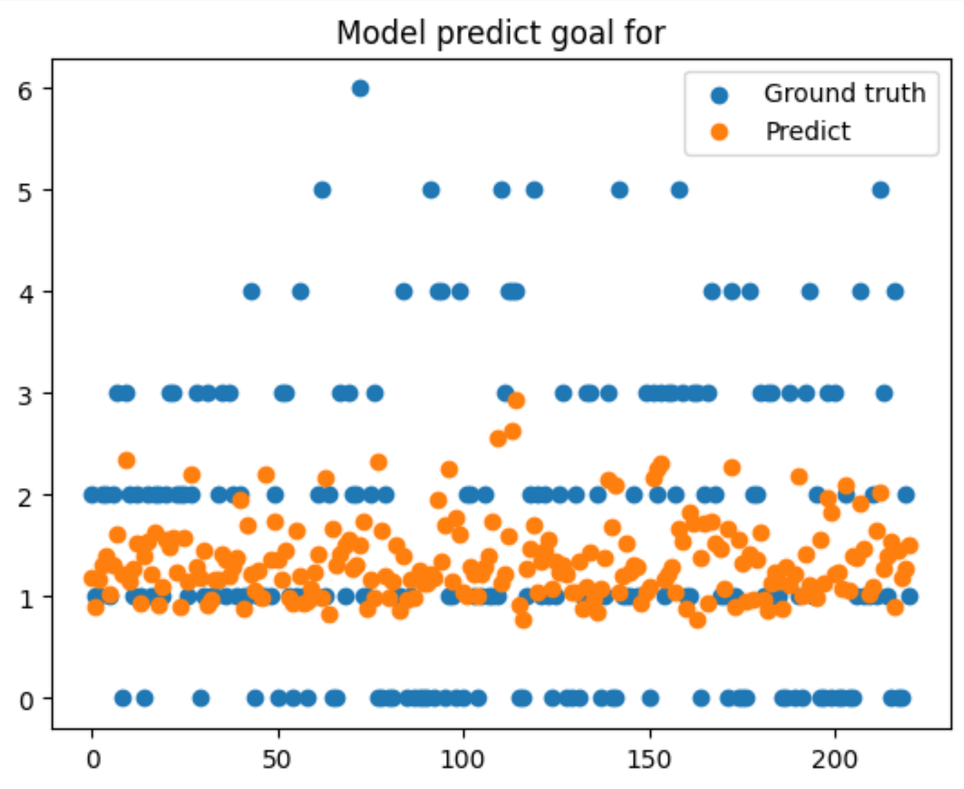
**Độ chính xác mô hình đã tăng hơn 1.3%**

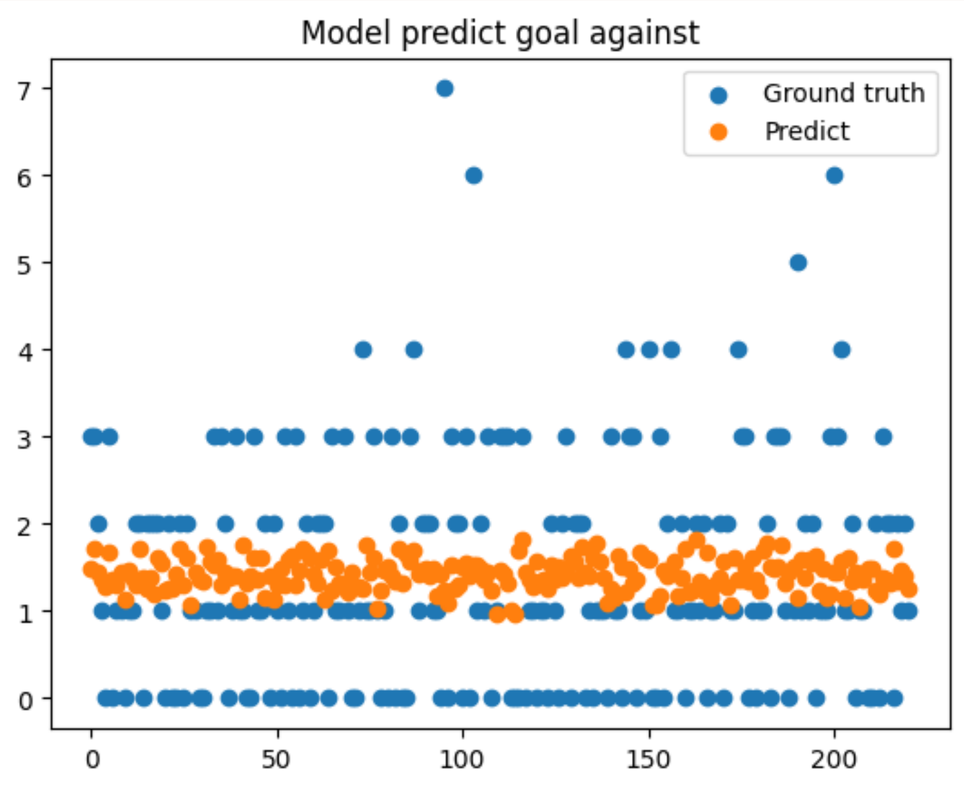
# III. Mô hình dự đoán(RandomForest)

```
Predict: Liverpool Win against Bournemouth
Predict: Liverpool Win against Manchester City
Predict: Liverpool draw or lose against Chelsea
Predict: Liverpool draw or lose against Arsenal
Predict: Liverpool draw or lose against Leeds United
Predict: Liverpool Win against Nott'ham Forest
Predict: Liverpool Win against West Ham
Predict: Liverpool Win against Tottenham
Predict: Liverpool Win against Fulham
Predict: Liverpool Win against Brentford
Predict: Liverpool Win against Leicester City
Predict: Liverpool Win against Aston Villa
```

**Dự đoán kết quả của CLB Liverpool từ 11/3/2023 - 20/5/2023**

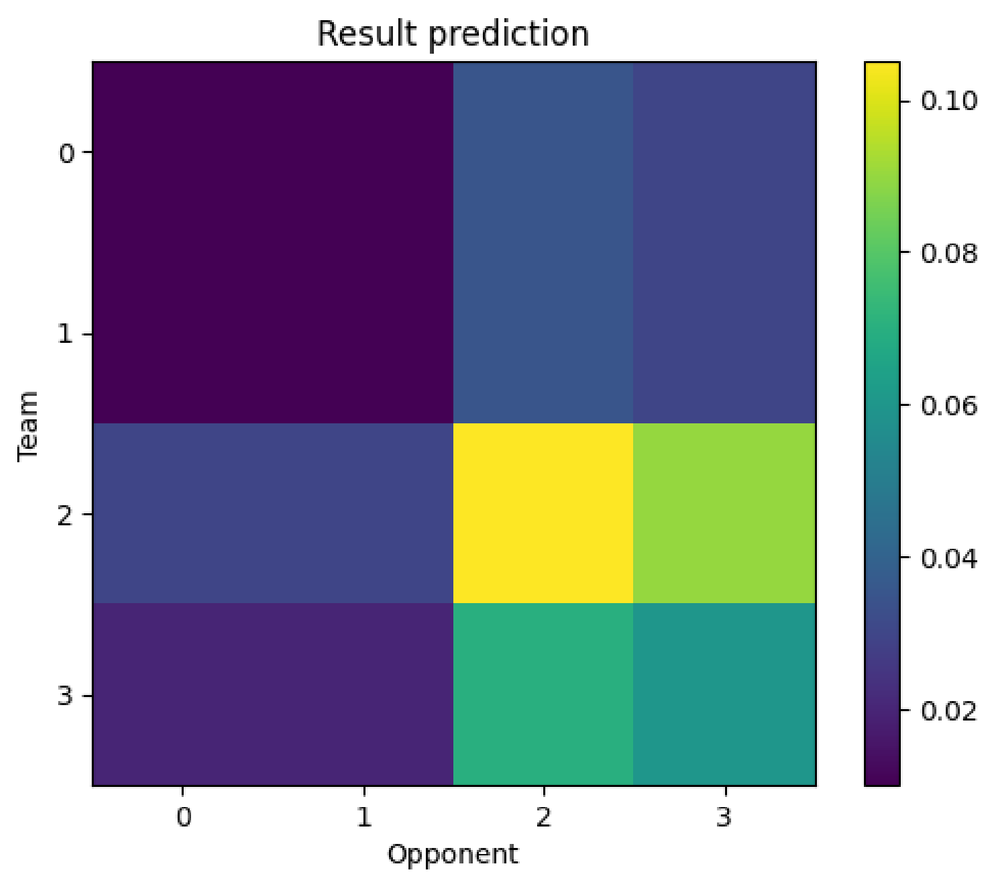
# III. Mô hình dự đoán (Poisson)




**Xây dựng 2 mô hình Poisson hồi quy, dự đoán số bàn thắng ghi được và số bàn thua phải nhận**

# III. Mô hình dự đoán (Poisson)




Từ đó, chúng ta có thể biết được xác suất xuất hiện của từng tỉ số trong trận đấu.

Khi cộng tất cả các xác suất có tỉ số có lợi cho đội thì ta nhận được tỉ lệ thắng của đội bóng đó.

Mô hình có độ chính xác: 60%

# Nhận xét

- Cả 2 mô hình đều cho độ chính xác trong khoảng 60-70%
- Điều đó có thể hiểu được vì bóng đá hiện đại đề cao tính chiến thuật, tỉ số có thể không được định đoạt bởi những thống kê, mà đơn thuần chỉ là số bàn thắng ghi được.

# IV. Kết luận và hướng phát triển

- Thông qua đề tài lần này, các thành viên trong nhóm hiểu rõ hơn về môn Khoa học dữ liệu, đặc biệt là quá trình thu thập dữ liệu, trực quan hóa dữ liệu để lựa chọn các đặc trưng phù hợp nhất và sử dụng các thông số để đánh giá mô hình, từ đó đưa ra được kết quả tốt nhất qua các lần kiểm thử.
- Qua quá trình thực hiện ở trên, có thể thấy được tất cả các bước trên đều quan trọng trong việc đưa ra mô hình dự đoán tối ưu nhất có thể. Có nhiều yếu tố ảnh hưởng đến tốc độ huấn luyện và độ chính xác làm mô hình còn một số hạn chế:

- Việc thu thập dữ liệu trên các trang web có thể gặp khó khăn do việc bảo mật của trang web hoặc do cùng một dữ liệu về giá kim cương thì thông tin chi tiết giữa các trang web lại khác nhau nên có thể ảnh hưởng đến việc áp dụng cho nhiều trang web.

- Đối với tập dữ liệu nhỏ hơn thì độ chính xác của mô hình thấp hơn so với khi sử dụng với tập dữ liệu lớn hơn. Tuy nhiên, trong một số trường hợp thì độ chính xác tăng không đáng kể mà còn làm cho mô hình trở nên phức tạp hơn và ảnh hưởng đến thời gian huấn luyện.

Giải pháp và hướng phát triển:

- Thu thập nhiều dữ liệu từ nhiều nguồn khác nhau để có tập dữ liệu tối ưu nhất.
- Sử dụng các phương pháp tiền xử lý khác nhằm tăng độ chính xác.
- Sử dụng thêm nhiều mô hình nhằm tăng kết quả dự đoán.

# Thank you !