TRUNG TÂM NORDIC CODER

Python for Data Analysis

BÁO CÁO CUỐI KHÓA

Đề tài:

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KẾT QUẢ TRẬN ĐẤU BÓNG ĐÁ TRÊN TẬP DỰ LIỆU 11 GIẢI VÔ ĐỊCH QUỐC GIA CHÂU ÂU

Giải viên : thầy Nguyễn Ngọc Tú

Học viên thực hiện : Trần Đức Trung

TP HỒ CHÍ MÌNH 12/2019

Mục Lục

Danh Sách Hình

CHƯƠNG I: Tổng Quát

1.01 Kết quả mong muốn

Xây dựng một mô hình giúp dự đoán kết quả của trận đấu bóng đá dựa trên một tập dữ liệu có sẵn.

1.02 Tập dữ liệu

Nguồn: https://www.kaggle.com/hugomathien/soccer

1.03 Giả thuyết

So sánh tỷ lệ phần trăm (%) thắng trên sân nhà của đội nhà (**A**) và tỷ lệ phần trăm (%) thắng trên sân khách của đội khách (**B**):

- Nếu A > B, đội nhà thắng.
- Nếu A < B, đội khách thắng.
- Nếu A = B, trận đấu đó sẽ hòa.

Tỷ lệ phần trăm (%) thắng trên sân nhà: là phần trăm số trận thắng trên các trận sân nhà đã đầu của một đội.

Tỷ lệ phần trăm (%) thắng trên sân khách: là phần trăm số trận thắng trên các trận sân khách đã đầu của một độ

CHƯƠNG II: Phân Tích Dữ Liệu

2.01 Lọc dữ liệu quan trọng

Dataframe Match:

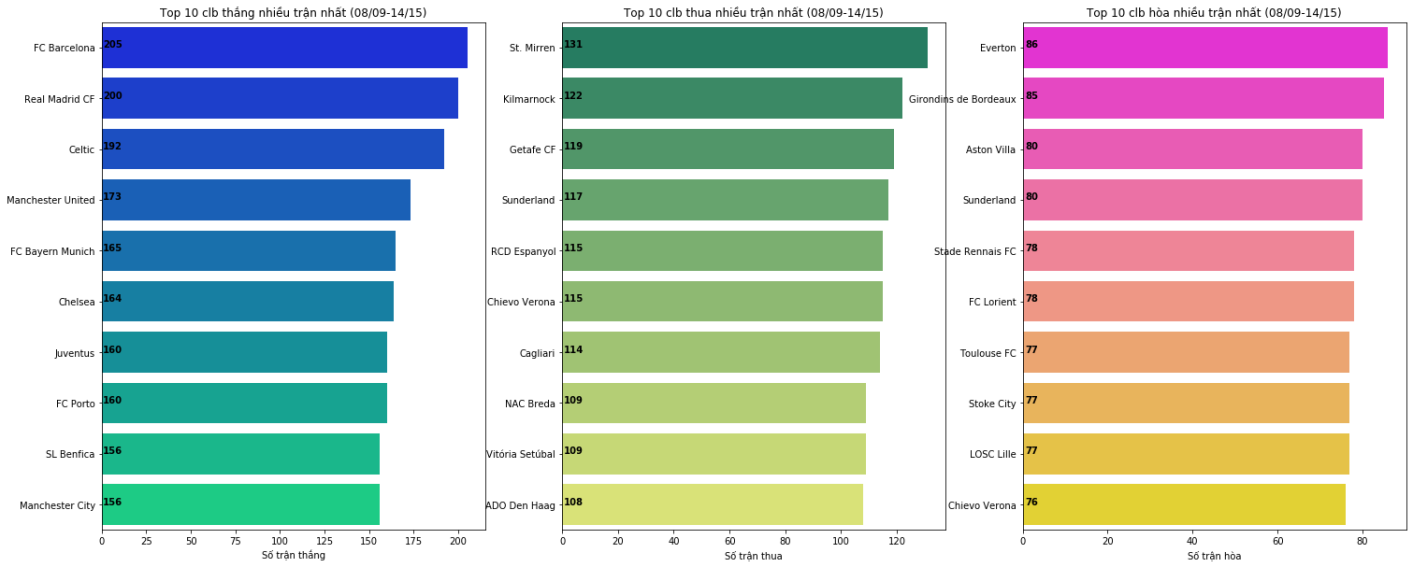
- ‘season’: Mùa giải tổ chức thi đấu, trong tập dữ liệu từ mùa 2008/2009).
- ‘home_team_api_id’: Số nhận diện đội bóng chủ nhà.
- ‘away_team_api_id’: Số nhận diện đội khách.
- ‘home_team_goal’: Số bàn thắng đội nhà.
- ‘away_team_goal’: Số bàn thắng đội khách.

Dataframe Team:

- ‘team_api_id’: Số nhận diện đội bóng.
- ‘team_long_name’: Tên đầy đủ của đội bóng.

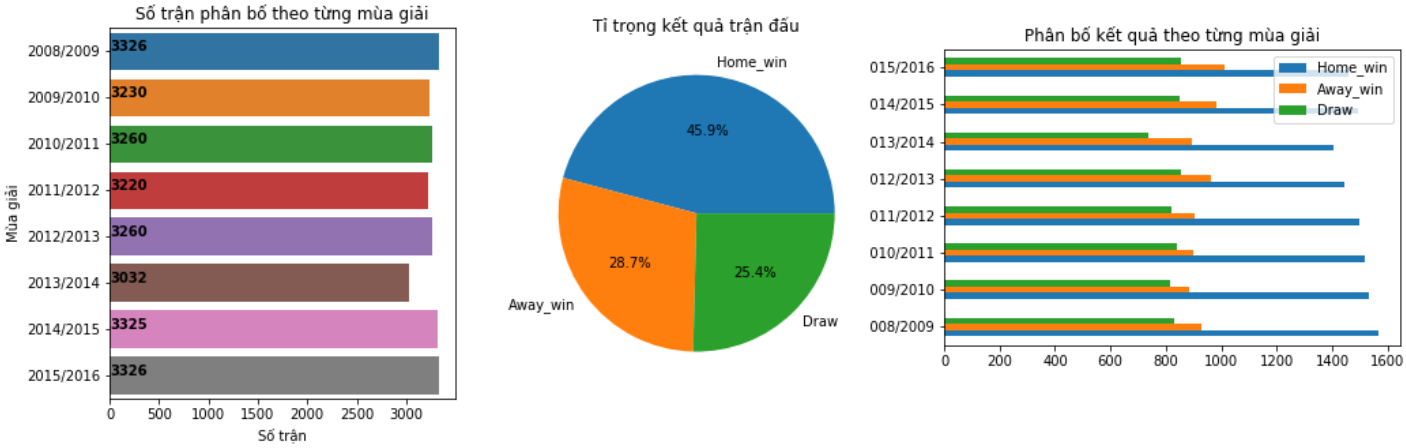
Đối chiếu dữ liệu dựa trên các thông tin lấy được:

- Dựa trên ‘home_team_api_id’ và ‘away_team_api_id’ cùng ‘team_api_id_id’ để xác tên các đội bóng tham gia vào các trận đấu cụ thể.



Hình 2.01-1 Top các đội bóng với các thông số khác nhau.

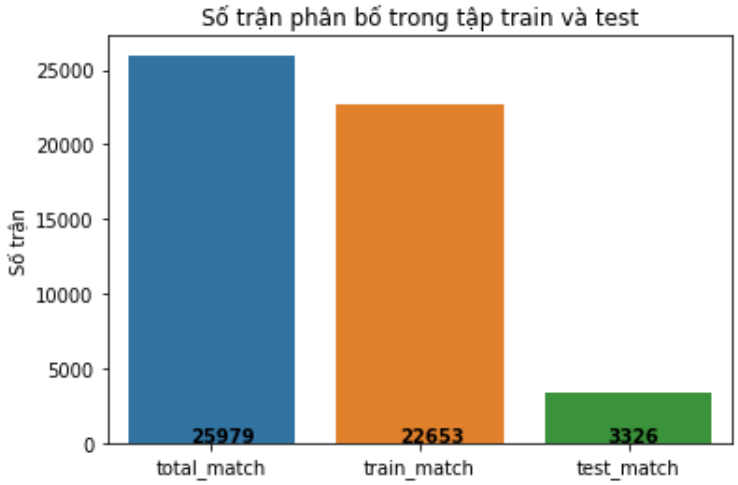
- Dựa trên ‘home_team_goal’ và ‘away_team_api’ để xác định kết quả thực sự của các trận đấu.



Hình 2.01-2 Phân tích tổng hợp tập dữ liệu

Tách ra hai tập dữ liệu dựa trên ‘season’:

- Tập train: từ ‘season’ 2008/2009 tới 2014/2015 nhằm mục đích huấn luyện mô hình.
- Tập test: ‘season’ 2015/2016 (gần nhất) nhằm mục đích kiểm tra mô hình huấn luyện được.

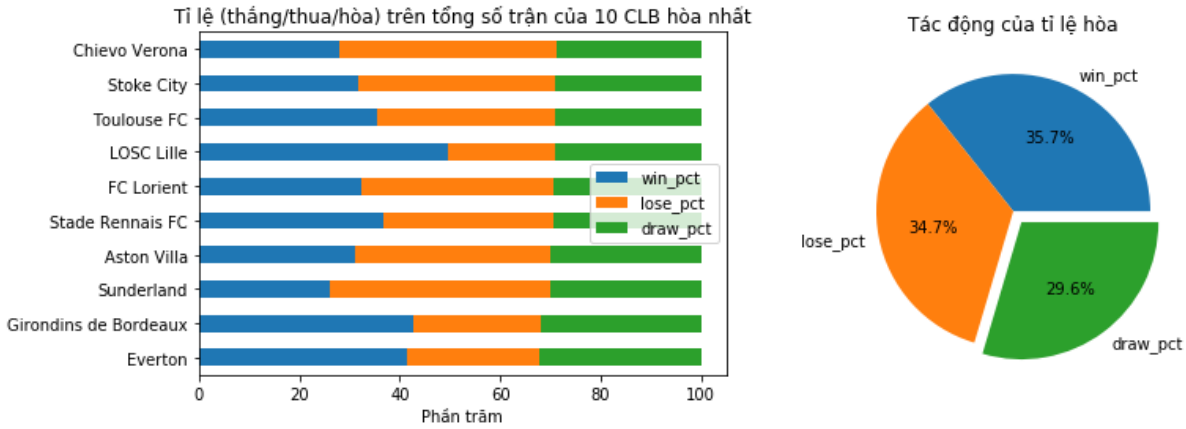


Hình 2.01-3 Phân bố tập dữ liệu 'train' và 'test'

2.02 Phân tích tập huấn luyện (train)

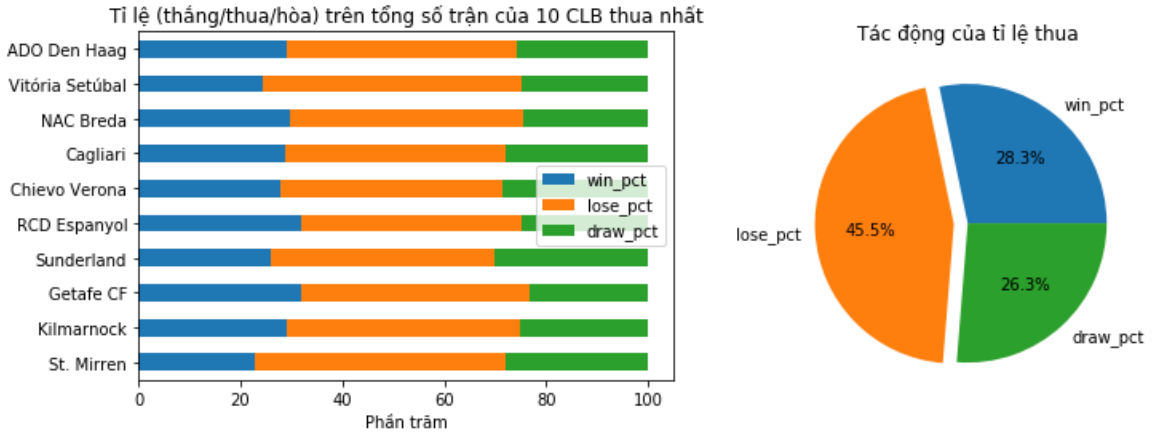
Hiệu năng thi đấu của từng đội (so sánh tỉ lệ thắng thua và hòa trên tổng số trận họ thi đấu):

- Với các đội có số trận hòa nhiều nhất thì sự ảnh hưởng của tỉ lệ hòa chiếm tỉ trọng thấp nhất. Điều này đồng nghĩa tỉ số hòa không có nhiều ý nghĩa để khai thác.



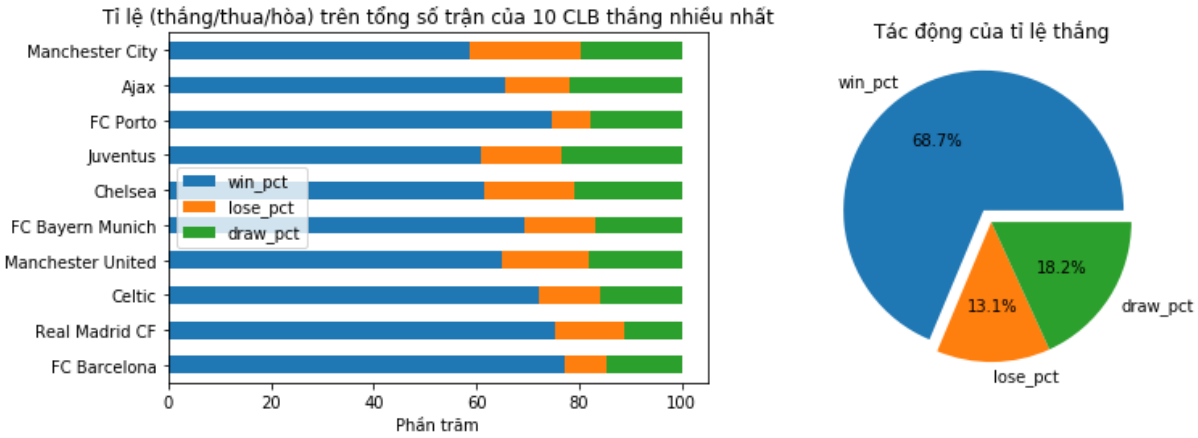
Hình 2.02-1 Độ ảnh hưởng của tỉ lệ hòa.

- Với các đội có số trận thua nhiều nhất thì sự ảnh hưởng của tỉ lệ thua có tỉ trọng cao nhất và chiếm gần một nửa (45.5%). Tỷ trọng này vẫn cần xem xét để chọn là yếu tố huấn luyện.



Hình 2.02-2 Độ ảnh hưởng của tỉ lệ thua

- Với các đội có số trận thắng nhiều nhất thì tỉ lệ thắng chiếm quá nữa (68.7%) và lớn một cách áp đảo các chỉ số khác.

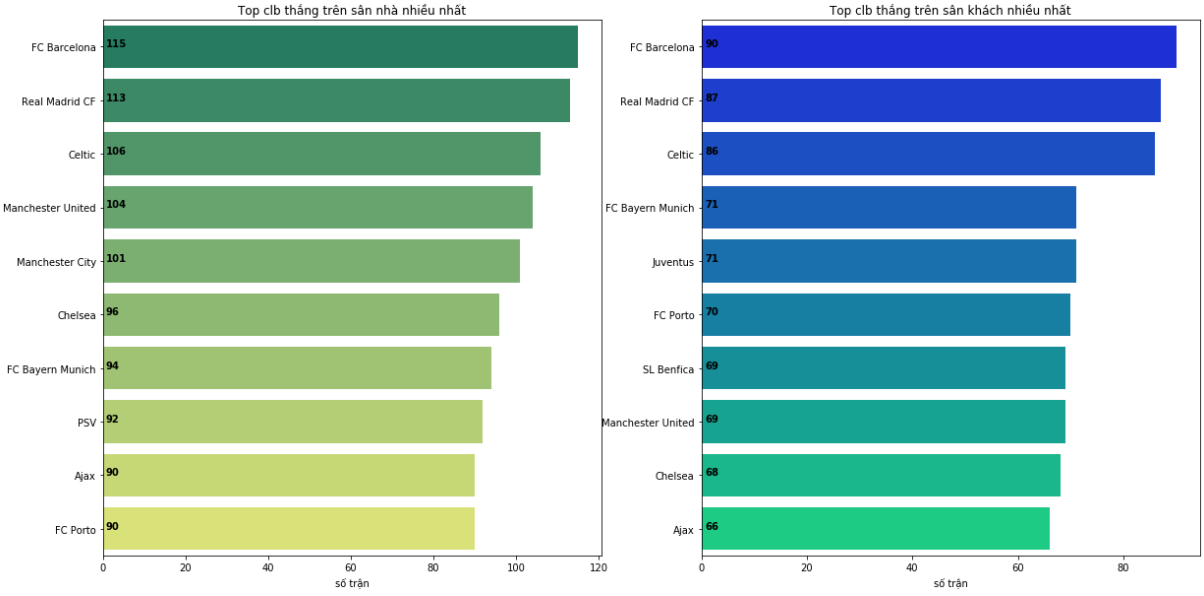


Hình 2.02-3 Độ ảnh hưởng của tỉ lệ thắng

2.03 Tính tỷ lệ thắng trên sân nhà và sân khách của từng đội

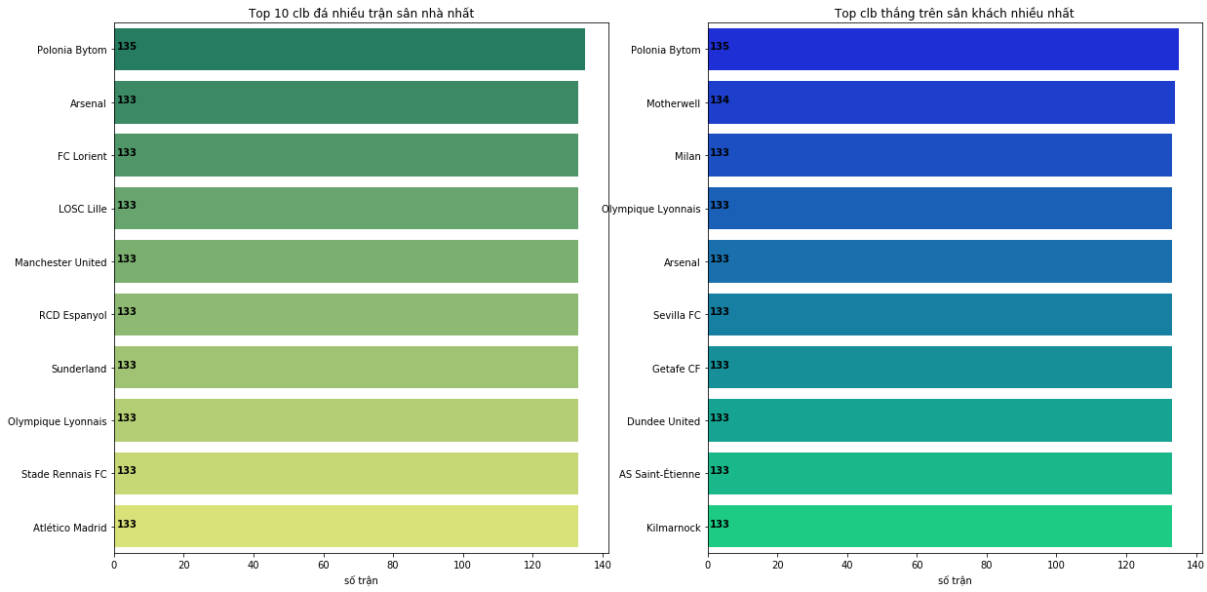
Đếm các thống số cần thiết trên từng đội từng đội:

- Số trận thắng trên sân nhà (1).
- Số trận thắng trên sân khách (2).



Hình 2.03-1 Top các đội dựa trên số trận thắng sân nhà và sân khách

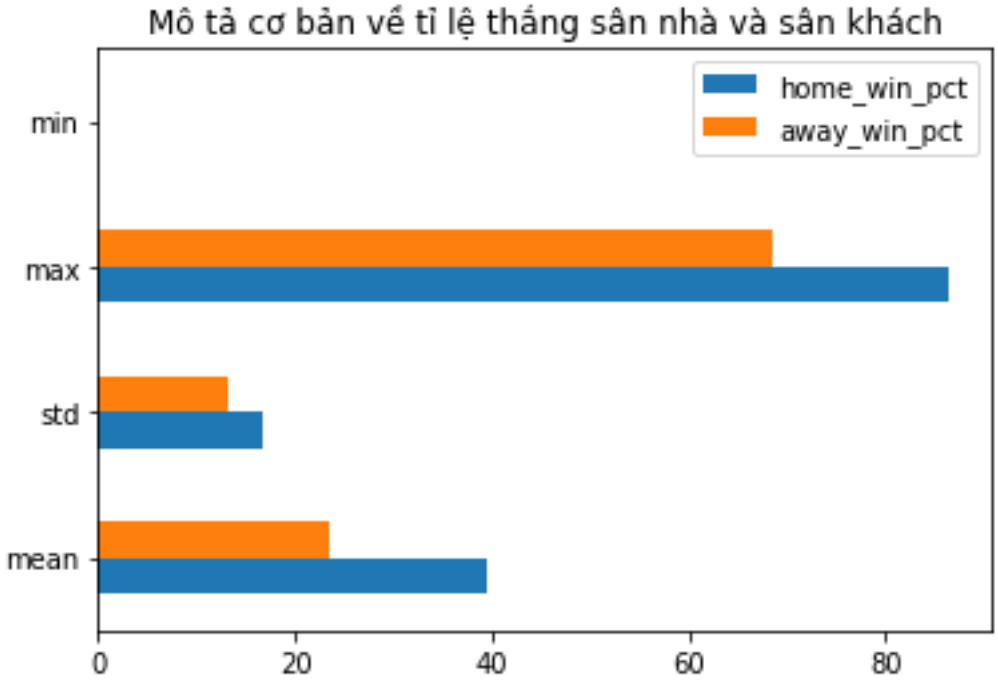
- Số trận thi đấu trên trận nhà (3).
- Số trận thi đấu trên trận khách (4).



Hình 2.03-2 Top các đội thi đấu nhiều nhất

Tính tỷ lệ thắng trên sân nhà và sân khách của từng đội:

- A = (1) / (3) * 100
- B = (2) / (4) * 100



Hình 2.03-3 Mô tả cơ bản về tỉ lệ thắng đội nhà và đội khách

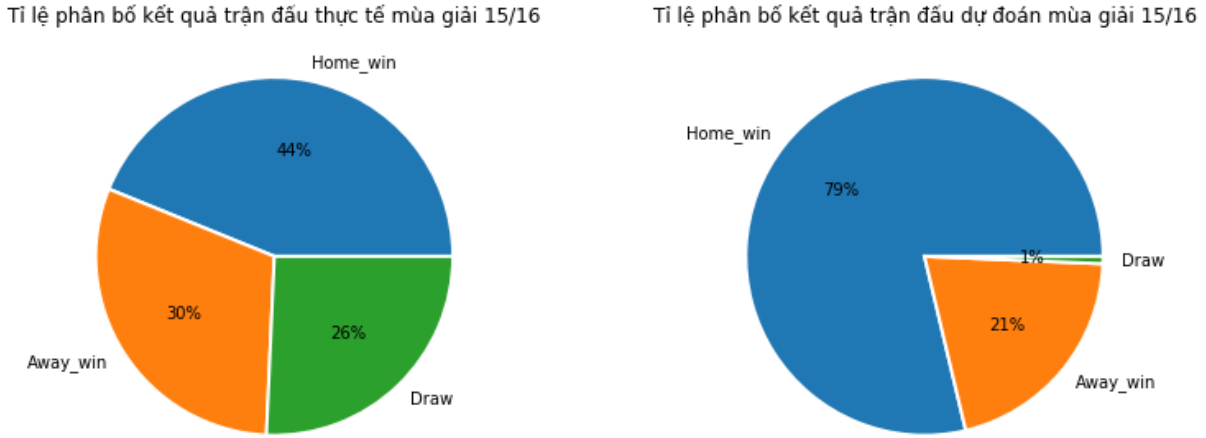
Nhận xét:

- Giá trị tối thiểu đều xuất hiện 0, đồng nghĩa là có các đội bóng không thắng bất ký một trận bào, hoặc các đội bóng mới lên hạng lần đầu nên không có thông số gì và sẽ được gán giá trị 0.
- Giá trị tối đa ở sân nhà là 86.46, và sân khách là 68.42.
- Giá trị trung bình ở nhà là 39.35, và lớn hơn ở sân khách là 23.5. Điều này cho thấy rằng tỉ lệ thắng của đội nhà sẽ cao hơn.
- Độ lệch chuẩn của tỉ lệ thắng là 16.75 với đội nhà và 13.2 với đội khách.

CHƯƠNG III: Tổng Kết

3.01 Dự đoán

Sau khi lấy được tỉ lệ thắng sân nhà và sân khách của từng đội từ tập huấn luyện. Ta có số liệu đó và dựa trên mô hình đề ra từ giả thuyết để tiến hành dự đoán kết các trận đấu.



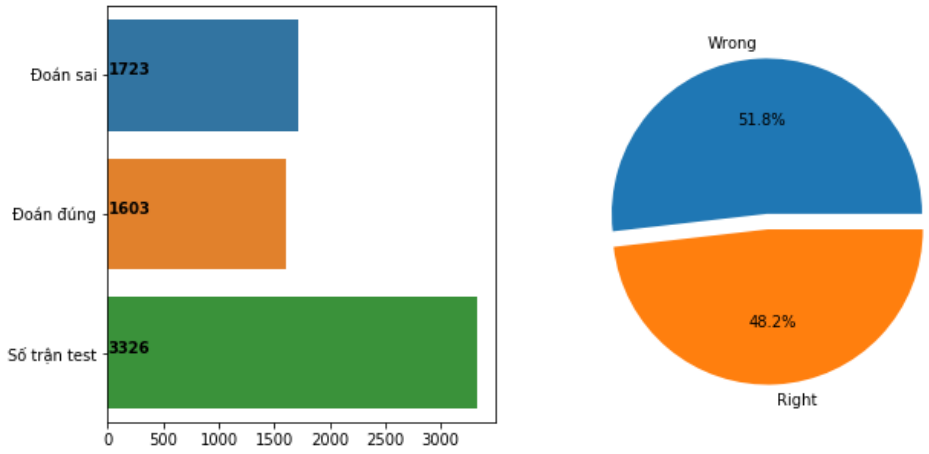
Hình 3.01-1 Phân bố kết quả thực tế và kết quả dự đoán.

Như đồ thị ở trên cho thấy mức độ sai lệch là khá lớn, đặc biệt là trên hai tập 'đội nhà thắng' và kết quả 'hòa'.

3.02 Đánh giá tổng quát

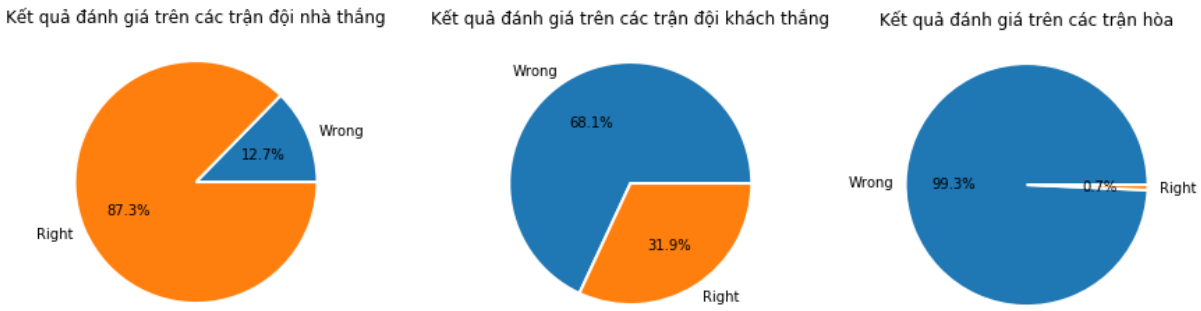
Phương pháp đánh giá:

- So sánh trực tiếp giữa các kết quả thực tế và kết quả dự đoán.
- Tính tỷ lệ phần trăm kết quả dự đoán trùng với thực tế để làm thang đo độ chính xác của mô hình.



Hình 3.03-1 Kết quả đánh giá tổng quát

3.03 Đánh giá trên các vùng khác



Hình 3.03-2 Kết quả đánh giá dựa trên các vùng tách biệt

3.04 Tổng kết

Dù tỉ lệ dự đoán chuẩn xác các trận đấu chỉ là 48.2%, nhưng lại đúng tới 87.3% các trận có kết quả là đội nhà thắng. Từ đó thấy rằng đây là một mô hình phù hợp cho việc dự đoán kết quả đội nhà thắng hay không.

Trong trường hợp đội nhà không thắng, điều này sẽ dẫn tới hai cơ hội khác có thể xảy ra, hòa hoặc thua (đội khách thắng). Đây cũng có thể xem là hương phát triển trong tương lai vì tỷ lệ dự đoán chính xác từ mô hình trên cho thấy là rất thấp.