



**TRƯỜNG ĐẠI HỌC BÁCH KHOA**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CUỐI KỲ**  
**HỌC PHẦN: KHOA HỌC DỮ LIỆU**

**TÊN ĐỀ TÀI: DỰ ĐOÁN ĐỘI THẮNG TRONG**  
**TRẬN ĐẤU BÓNG ĐÁ**

Giảng viên hướng dẫn : TS. Ninh Khánh Duy

Nhóm	4
Họ Và Tên Sinh Viên	Lớp Học Phần
Nguyễn Dương Gia Bảo	20nh91
Vũ Tiến Hùng	
Võ Chí Tài	

ĐÀ NẴNG, 06/2023

## TÓM TẮT

### 1. Vấn đề cần giải quyết

- Bài toán đặt ra: Dự đoán đội chiến thắng trong trận đấu bóng đá
- Input: Dữ liệu các trận đấu bóng đá
  - date: Ngày diễn ra trận đấu
  - time: Thời gian bắt đầu của trận đấu
  - comp: Giải đấu mà trận đấu diễn ra
  - round: Vòng đấu của giải đấu mà trận đấu diễn ra
  - day: Thứ trong tuần khi trận đấu diễn ra
  - venue: Sân nhà - Sân khách
  - result: Kết quả của trận đấu
  - gf: Số bàn thắng ghi được của đội nhà trong trận đấu
  - ga: Số bàn thua của đội nhà trong trận đấu
  - opponent: Đội đối phương
  - poss: Tỷ lệ kiểm soát bóng của đội nhà trong trận đấu
  - referee: Trọng tài bắt trận đấu
  - season: Mùa giải mà trận đấu diễn ra
  - team: Đội bóng tham gia trận đấu (đội nhà)
  - ...
- Output: Kết quả thắng thua của một trận đấu

### 2. Phương pháp giải quyết

Sử dụng phương pháp học có giám sát sử dụng 2 mô hình huấn luyện dữ liệu là mô hình GradientBoostingClassifier và RandomForestClassifier để tìm kết quả bài toán và so sánh kết quả của 2 mô hình

### 3. Kết quả đạt được

- Crawl được dữ liệu từ nhiều nguồn khác nhau
- Hoàn thành bài toán, so sánh kết quả của 2 mô hình
- Có chương trình Demo hoàn chỉnh

## BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai)
Vũ Tiến Hùng	<ul style="list-style-type: none"> <li>- Tìm nguồn dữ liệu</li> <li>- Crawl dữ liệu</li> <li>- Làm sạch dữ liệu</li> <li>-Viết báo cáo và slide phần Crawl dữ liệu</li> </ul>	<ul style="list-style-type: none"> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> </ul>
Võ Chí Tài	<ul style="list-style-type: none"> <li>-Xử lý dữ liệu</li> <li>-Đặc trưng hóa dữ liệu</li> <li>- Viết báo cáo và slide phần xử lý dữ liệu</li> </ul>	<ul style="list-style-type: none"> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> </ul>
Nguyễn Dương Gia Bảo	<ul style="list-style-type: none"> <li>-Tìm bộ siêu tham số của 2 mô hình GradientBoostingClassifier RandomForestClassifier</li> <li>- Thực thi huấn luyện mô hình</li> <li>- Kết quả bài toán</li> <li>-Viết báo cáo và slide phần thực thi mô hình</li> </ul>	<ul style="list-style-type: none"> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> <li>- Đã hoàn thành</li> </ul>

## MỤC LỤC

I. Giới thiệu.....	4
1. Giới thiệu vấn đề .....	4
2. Giải pháp tổng quan.....	4
II. Thu thập và mô tả dữ liệu .....	4
1. Thu thập dữ liệu.....	4
2. Mô tả dữ liệu.....	5
III. Trích xuất đặc trưng .....	8
IV. Mô hình hóa dữ liệu .....	11
1. Tổng kết.....	18
2. Kết quả đạt được.....	18
3. Nhận xét đánh giá.....	20
4. Hướng phát triển.....	20
VI. Tài liệu tham khảo .....	20

## **I. Giới thiệu**

### **1. Giới thiệu vấn đề**

Đề tài dự đoán đội chiến thắng trong một trận đấu bóng đá là một bài toán phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và học máy.

Vấn đề này đặt ra mục tiêu nhằm tăng cường khả năng dự đoán kết quả của các trận đấu bóng đá, giúp người chơi, người đặt cược, hay các đội bóng có thể đưa ra những quyết định tốt hơn về chiến thuật, đội hình, hay các đội chiến lược khác để giành chiến thắng.

### **2. Giải pháp tổng quan**

Bài toán này có thể được giải quyết bằng các phương pháp học máy như: học có giám sát, học không giám sát và học bán giám sát.

Phương pháp học có giám sát dựa trên việc sử dụng dữ liệu huấn luyện để học quy luật thống kê và các mô hình dự đoán, ví dụ như mô hình logistic regression, naive Bayes, decision trees, random forests và các mô hình deep learning như neural networks, convolutional neural networks.

Ở bài toán này, chúng ta sử dụng phương pháp học có giám sát sử dụng 2 mô hình huấn luyện dữ liệu là mô hình GradientBoostingClassifier và RandomForestClassifier để so sánh với nhau.

Việc giải quyết bài toán dự đoán đội chiến thắng trong một trận đấu bóng đá sẽ giúp chúng ta có những quyết định tốt hơn trong các hoạt động đặt cược hay cân nhắc chiến thuật và đội hình cho những trận đấu bóng đá, và từ đó điều đó giúp các đội bóng có thể đạt được thành tích tốt hơn trong các giải đấu.

## **II. Thu thập và mô tả dữ liệu**

### **1. Thu thập dữ liệu**

-Nguồn dữ liệu : <https://fbref.com/en>

-Công cụ thu thập: BeautifulSoup

BeautifulSoup: Là một thư viện Python phổ biến được sử dụng để lấy dữ liệu từ các trang web. Nó cung cấp các công cụ để phân tích cú pháp HTML và XML và thu thập dữ liệu theo các tiêu chí nhất định.

-Cách thức sử dụng công cụ:

- Công cụ BeautifulSoup được đề cập ở trên là thư viện Python BeautifulSoup, để sử dụng thư viện này, bạn cần thực hiện các bước sau đây:

- Cài đặt BeautifulSoup: Bạn có thể cài đặt BeautifulSoup thông qua pip trong command line: ``pip install beautifulsoup4``.
- Import thư viện: Sau khi cài đặt, bạn cần import thư viện BeautifulSoup bằng cách thêm dòng sau vào file Python của mình: **`from bs4 import BeautifulSoup`**
- Lấy nội dung HTML: Tải về nội dung HTML của trang web mà bạn muốn crawl. Bạn có thể sử dụng các thư viện như urllib, requests hoặc Scrapy để lấy nội dung này.
- Sử dụng BeautifulSoup để phân tích cú pháp HTML: Sau khi có nội dung HTML, bạn có thể sử dụng BeautifulSoup để phân tích cú pháp và lấy dữ liệu từ trang web.

-Đầu vào và đầu ra của quá trình thu thập:

- Đầu vào: một liên kết của nguồn dữ liệu
- Đầu ra : 2 file csv : SmallDS\_raw.csv với 1000 mẫu, BigDS\_raw.csv với 10000 mẫu

-Dưới đây là ví dụ để lấy các thẻ `<a>` (liên kết) từ một trang web:

Đoạn mã này sẽ in ra các liên kết URL tìm thấy trên trang web "http://example.com". Bạn có thể thay đổi các thẻ HTML hoặc xử lý dữ liệu theo các yêu cầu cụ thể của bạn.

```
from bs4 import BeautifulSoup
import requests

page = requests.get("http://example.com")
soup = BeautifulSoup(page.content, 'html.parser')
links = soup.find_all('a')

# In các liên kết URL
for link in links:
    print(link.get('href'))
```

Hình 1 Ví dụ về công cụ BeautifulSoup

## 2. Mô tả dữ liệu

-Có 2 tập dữ liệu SmallDS\_raw.csv, BigDS\_raw.csv

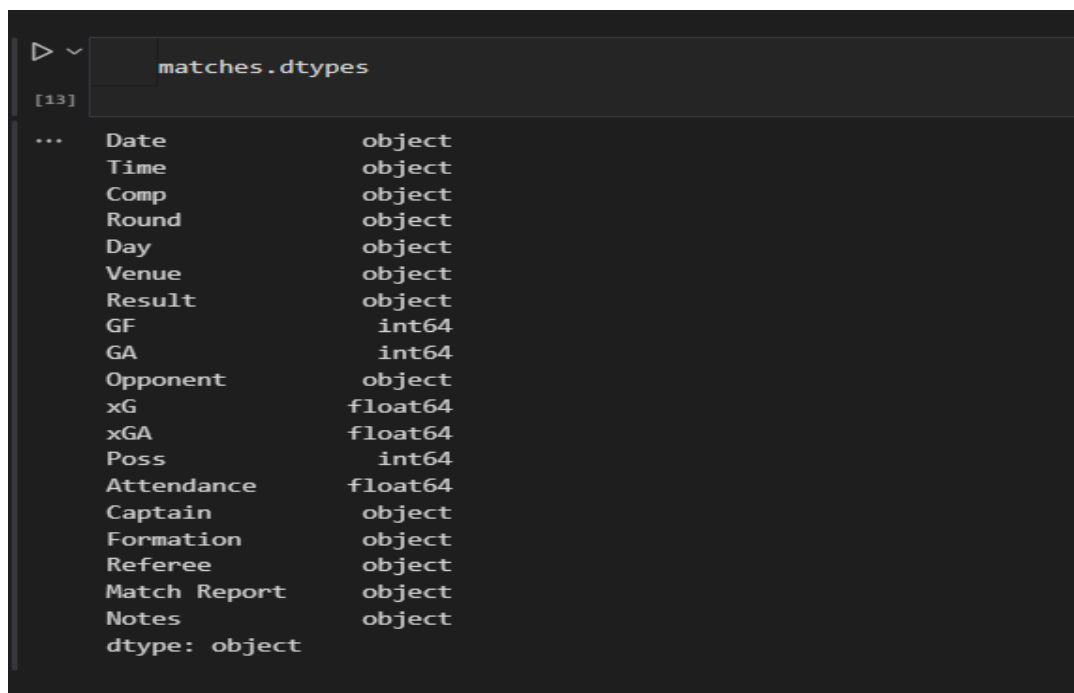
- Số lượng mẫu :
  - SmallDS\_raw.csv : 1000 mẫu
  - BigDS\_raw.csv : 10000 mẫu

- Số đặc trưng của một mẫu (SmallDS\_raw.csv, BigDS\_raw.csv ): 19 đặc trưng

- Date: kiểu dữ liệu thời gian (date) để chỉ ngày tháng.
- Time: kiểu dữ liệu thời gian (time) để chỉ thời điểm trong ngày.
- Comp: kiểu dữ liệu chuỗi (string) để chỉ giải đấu hoặc giải đấu liên quan.
- Round: kiểu dữ liệu chuỗi (string) để chỉ vòng đấu của trận đấu.
- Day: kiểu dữ liệu chuỗi (string) để chỉ ngày trong tuần của trận đấu.
- Venue: kiểu dữ liệu chuỗi (string) để chỉ sân bóng đá diễn ra trận đấu.
- Result: kiểu dữ liệu chuỗi (string) để chỉ kết quả của trận đấu.
- GF: kiểu dữ liệu số nguyên (integer) để chỉ số bàn thắng ghi được của đội chủ nhà (goals for).
- GA: kiểu dữ liệu số nguyên (integer) để chỉ số bàn thắng ghi được của đội khách (goals against).
- Opponent: kiểu dữ liệu chuỗi (string) để chỉ đội bóng đối thủ.
- xG: kiểu dữ liệu số thực (float) để chỉ xG ghi được trong trận đấu.
- xGA: kiểu dữ liệu số thực (float) để chỉ xG bị thủng lưới trong trận đấu.
- Poss: kiểu dữ liệu số thực (float) để chỉ tỷ lệ kiểm soát bóng (possession).
- Attendance: kiểu dữ liệu số nguyên (integer) để chỉ số lượng khán giả có mặt trong trận đấu.
- Captain: kiểu dữ liệu chuỗi (string) để chỉ cầu thủ đội chủ nhà làm đội trưởng.
- Formation: kiểu dữ liệu chuỗi (string) để chỉ hệ thống chiến thuật của đội chủ nhà.
- Referee: kiểu dữ liệu chuỗi (string) để chỉ trọng tài của trận đấu.
- Match Report: kiểu dữ liệu chuỗi (string) để lưu trữ địa chỉ của bài báo cáo/trang web về trận đấu.
- Notes: kiểu dữ liệu chuỗi (string) để lưu trữ các ghi chú và thông tin liên quan khác về trận đấu.

Hình 2: Số đặc trưng của mỗi mẫu

- Kiểu dữ liệu của mỗi đặc trưng: (SmallDS\_raw.csv, BigDS\_raw.csv )



The screenshot shows a Jupyter Notebook interface with a code cell containing the command `matches.dtypes`. The output displays the data types for 19 features of the 'matches' DataFrame. The features are listed on the left, and their corresponding data types are listed on the right. The data types are: object for Date, Time, Comp, Round, Day, Venue, Result, Opponent, Captain, Formation, Referee, Match Report, and Notes; int64 for GF, GA, and Poss; and float64 for xG, xGA, and Attendance.

Feature	Data Type
Date	object
Time	object
Comp	object
Round	object
Day	object
Venue	object
Result	object
GF	int64
GA	int64
Opponent	object
xG	float64
xGA	float64
Poss	int64
Attendance	float64
Captain	object
Formation	object
Referee	object
Match Report	object
Notes	object
dtype:	object

Hình 3: Kiểu dữ liệu của mỗi đặc trưng

- Số mẫu dữ liệu trống của mỗi đặc trưng
  - SmallDS\_raw.csv :

```
matches_small.isna().sum()

[7506]

...  date      0
     time      0
     comp      0
     round     0
     day       0
     venue     0
     result    0
     gf        0
     ga        0
     opponent  0
     xg        0
     xga       0
     poss      0
     sh        0
     sot       0
     dist      0
     fk        0
     pk        0
     pkatt     0
     season    0
     team      0
     dtype: int64
```

Hình 4 : Số mẫu dữ liệu trống của mỗi đặc trưng SmallDS\_raw.csv

- BigDS\_raw.csv :

```
matches_big.isna().sum()

[7516]

...  date      0
     time      0
     comp      0
     round     0
     day       0
     venue     0
     result    0
     gf        0
     ga        0
     opponent  0
     xg       1746
     xga      1746
     poss      2
     sh        2
     sot       2
     dist     1750
     fk       2506
     pk        2
     pkatt     2
     season    0
     team      0
     dtype: int64
```

Hình 5: Số mẫu dữ liệu trống BigDS\_raw.csv



### III. Trích xuất đặc trưng

- Lựa chọn đặc trưng:

- Loại bỏ những đặc trưng không cần thiết : `match report`, `node`, `attendance`, `referee`, `formation`, `caption`
- Thêm các đặc trưng mới từ file raking.csv : `rank\_before`, `rank\_before\_dif`, `point\_by\_rank\_before`
- Tiến hành trượt các đặc trưng: "xg", "xga", "gf", "ga", "goals\_dif", "sh", "sot", "dist", "fk", "pk", "pkatt", "point\_win"

Lựa chọn 19 đặc trưng sau: `venue`, `opp`, `hour`, `day`, `rank\_before`, `rank\_before\_dif`, `point\_by\_rank\_before`, `xg\_rolling`, `xga\_rolling`, `gf\_rolling`, `ga\_rolling`, `sh\_rolling`, `goals\_dif\_rolling`, `sot\_rolling`, `dist\_rolling`, `fk\_rolling`, `pk\_rolling`, `pkatt\_rolling` và `point\_win\_rolling`.

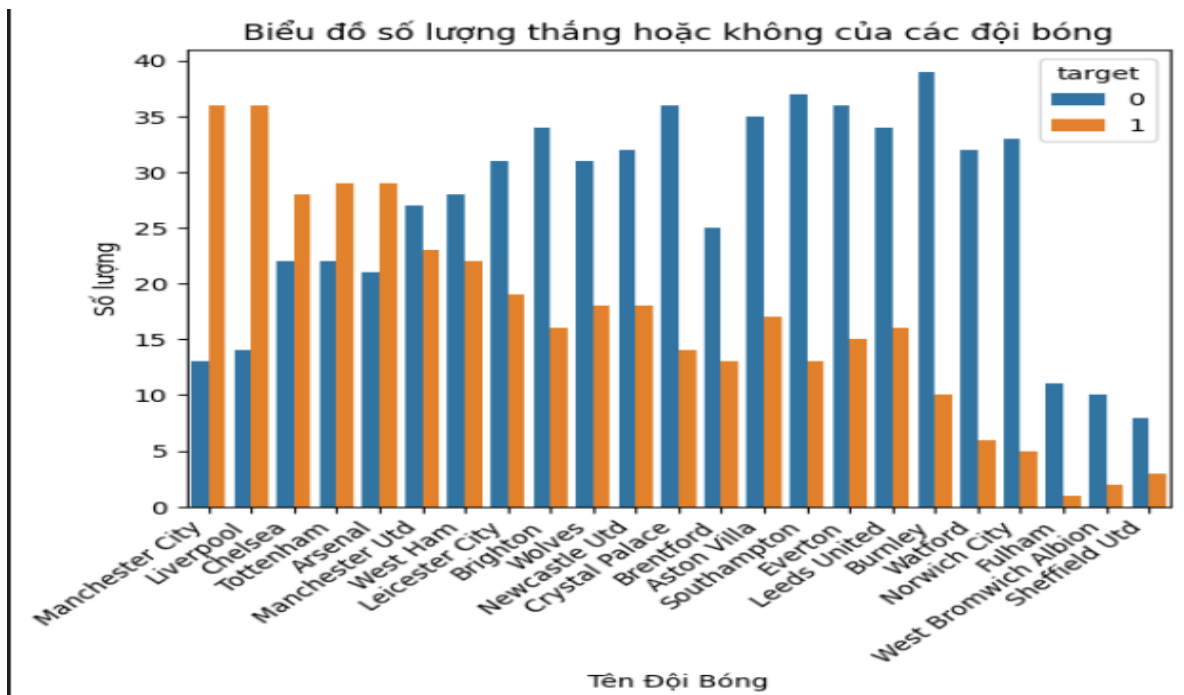
+ Làm sạch dữ liệu:

- Chuyển đổi về cùng một kiểu dữ liệu : ở cột `ga` và `gf`
- Chuyển đổi ở cột `team` và `opponent` về cùng một tên gọi
- Sau khi làm sạch dữ liệu ta có 2 file SmallDS.csv và BigDS.csv

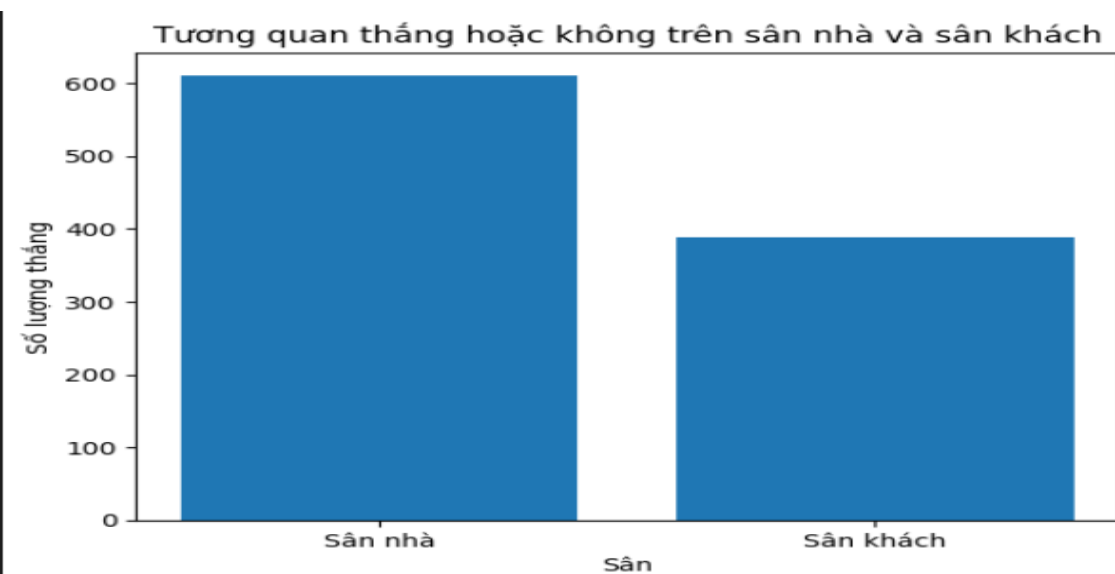
+ Xử lý dữ liệu :

- Xử lý dữ liệu trống : Điền các giá trị trung bình của các ô trống trong các cột
- Chuyển đổi dữ liệu:
  - Chuyển đổi dữ liệu về số cột: `venue`, `opponent`, `round`, `result`, `date`, `time`
  - Thêm cột `target` là 1 nếu cột result là 'W' và là 0 nếu 'D' hoặc 'L'
  - Thêm cột goals\_dif (số bàn thắng đội nhà trừ số bàn thắng đội đối đầu)

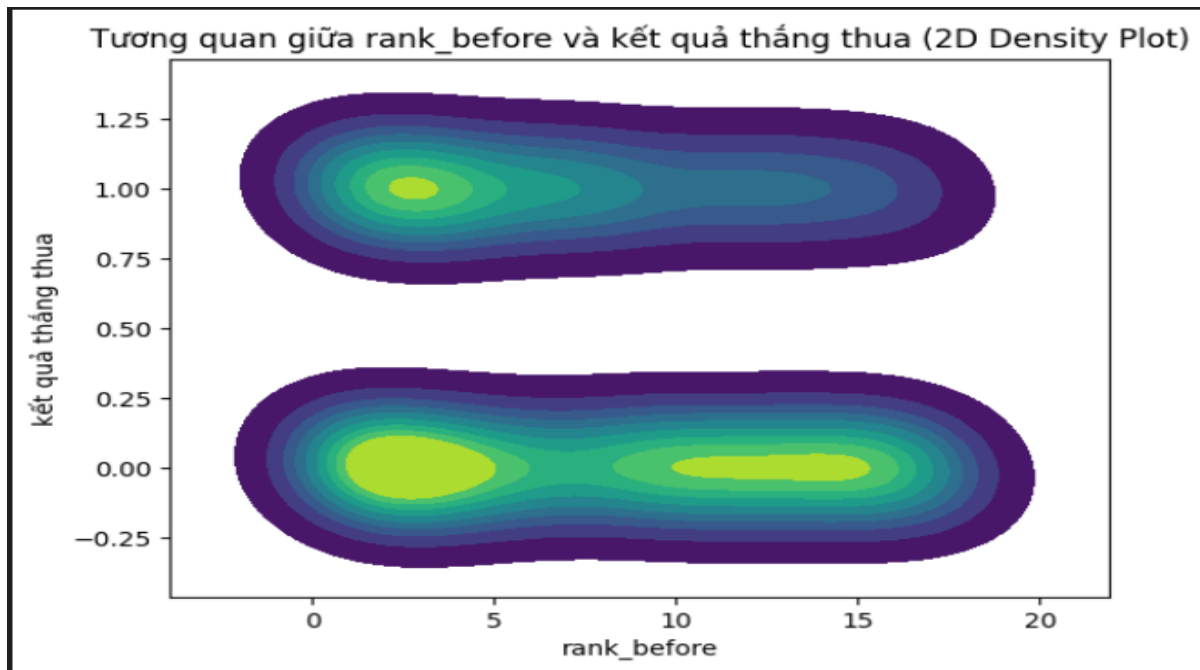
- Trực quan hóa dữ liệu:



Hình 7: Biểu đồ số lượng thắng hoặc không của các đội bóng



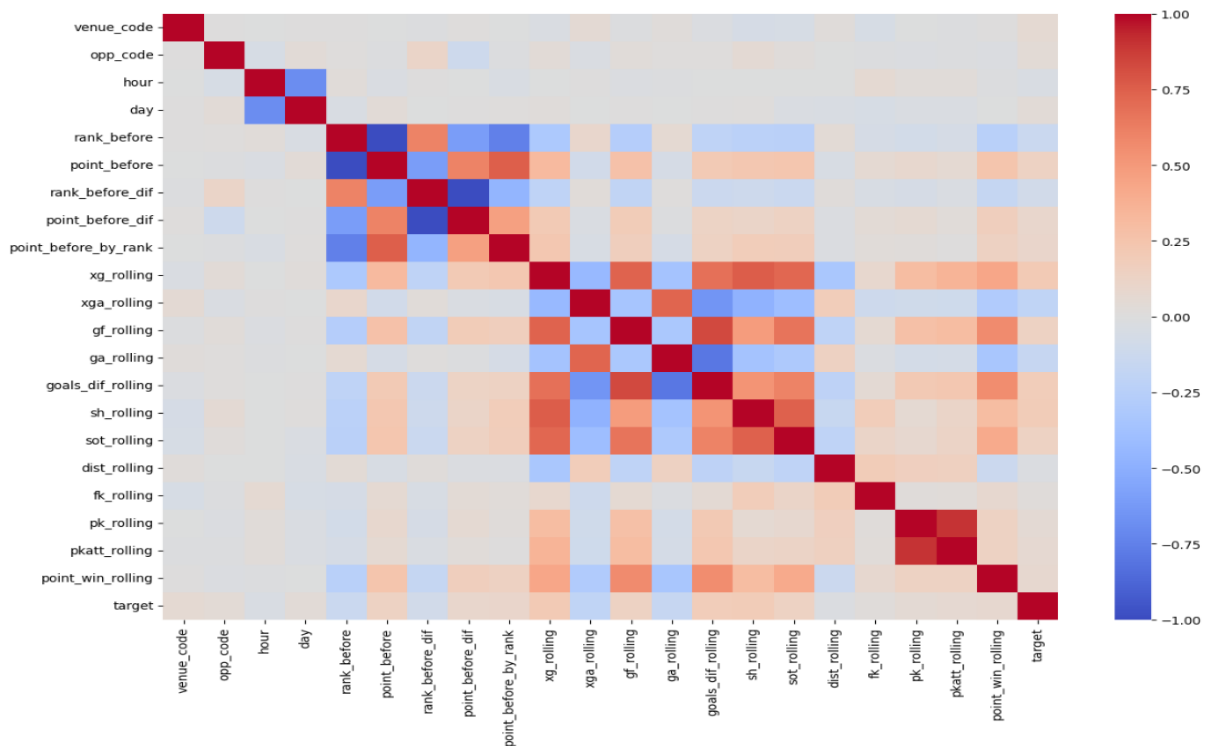
Hình 8: Biểu đồ tương quan thắng hoặc không trên sân nhà và sân khách



Hình 9: Biểu đồ tương quan giữa rank\_before và kết quả thua

- Quan sát các đặc trưng:

- Biểu đồ thể hiện mối quan hệ và sự tương quan
- Màu sắc càng đỏ đậm thể hiện mối quan hệ và sự tương qua càng lớn và màu xanh càng đậm thì ngược lại



Hình 10: Quan sát các đặc trưng

#### IV. Mô hình hóa dữ liệu

- Sử dụng 2 mô hình GradientBoostingClassifier và RandomForestClassifier

+GradientBoostingClassifier:

GradientBoostingClassifier là mô hình dự báo kết hợp của nhiều cây quyết định, với mỗi cây được xây dựng dựa trên các trọng điểm khác nhau cho những điểm dữ liệu mà mô hình trước đã không dự báo đúng. Tổng quan về thuật toán xây dựng một mô hình dự báo trên một mẫu dữ liệu, tính toán sai số của mô hình trên nó, sau đó xây dựng một mô hình mới nhằm dự báo sai số, và lặp lại quá trình này cho đến khi đạt được sai số tối thiểu.

Đối với thuật toán GradientBoostingClassifier, bộ tham số chính là:

- learning rate: kích thước bước của quá trình cập nhật trọng số sau mỗi lượt xử lý.
- number of estimators: số lượng cây quyết định được sử dụng trong quá trình huấn luyện.
- max depth: độ sâu tối đa cho cây quyết định.
- min sample leaf: số lượng điểm dữ liệu nhỏ nhất mà phải tồn tại trên các lá của cây quyết định mới được tạo.

+RandomForestClassifier:

RandomForestClassifier là một mô hình học máy dựa trên rừng cây quyết định, trong đó mỗi cây quyết định được xây dựng trên một tập con của các điểm dữ liệu được lấy mẫu ngẫu nhiên từ tập dữ liệu huấn luyện. Các dự đoán được tính bằng cách lấy trung bình các dự đoán của tất cả các cây.

Đối với thuật toán RandomForestClassifier, các bộ tham số quan trọng đó:

- n\_estimators: số lượng cây quyết định trong rừng.
- max\_depth: độ sâu tối đa của cây quyết định.
- min\_samples\_split: số lượng điểm dữ liệu nhỏ nhất để tiếp tục phân tách cây.
- min\_samples\_leaf: số lượng điểm dữ liệu nhỏ nhất có mặt trên một lá cây.
- max\_features: số lượng đặc trưng được sử dụng trong quá trình xây dựng cây quyết định.

-Chia dữ liệu thành 2 tập train và test

Tập train : tập dữ liệu từ ngày 01/01/2022 trở về trước

Tập test : tập dữ liệu từ ngày 01/01/2022 trở về sau

Tỉ lệ dữ liệu: train/test : 7/3

- Bộ siêu tham số của 2 mô hình huấn luyện:

SmallDS.csv

+ Mô hình GradientBoostingClassifier:

```
model = GradientBoostingClassifier(random_state = 5)
params = {
    'learning_rate': [0.01, 0.05, 0.1],
    'min_samples_leaf': [2, 3, 5],
    'min_samples_split': [10, 12, 13],
    'max_depth': [2,3]
}
model_cv = GridSearchCV(model, params, cv = 5, n_jobs = -1, verbose = False)
```

Hình 11: GridSearchCV với GradientBoostingClassifier SmallDS.csv

Kết quả của bộ siêu tham số GradientBoostingClassifier

```
▼ GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.05, max_depth=2, min_samples_leaf=3,
min_samples_split=15, random_state=5)
```

Hình 12: Bộ siêu tham số GradientBoostingClassifier SmallDS.csv

+ Mô hình RandomForestClassifier:

```
model = RandomForestClassifier(random_state = 5)
params = {
    'max_depth':[2,3,5],
    'min_samples_split': [5, 10, 15 ],
    'min_samples_leaf': [2, 5, 15],
    'n_estimators' : [100,150]
}
model_cv = GridSearchCV(model, params, cv = 5, n_jobs = -1, verbose = False)
```

Hình 13 : GridSearchCV với RandomForestClassifier SmallDS.csv

Kết quả của bộ siêu tham số RandomForestClassifier

```
▼ RandomForestClassifier
RandomForestClassifier(max_depth=5, min_samples_leaf=15, min_samples_split=5,
n_estimators=200, random_state=5)
```

Hình 14: Bộ siêu tham số RandomForestClassifier SmallDS.csv

BigDS.csv

+ Mô hình GradientBoostingClassifier:

```

model = GradientBoostingClassifier(random_state = 5)
params = {
    'learning_rate': [0.1, 0.25, 0.2],
    'n_estimators': [100, 125],
    'max_depth': [2, 3],
    'min_samples_leaf': [2, 3, 5],
    'min_samples_split': [10, 12, 13],
}
model_cv = GridSearchCV(model, params, cv = 2, n_jobs = -1, verbose = False)

```

Hình 15: GridSearchCV với GradientBoostingClassifier BigDS.csv

Kết quả của bộ siêu tham số GradientBoostingClassifier

```

▼ GradientBoostingClassifier
GradientBoostingClassifier(min_samples_leaf=3, min_samples_split=13,
                           n_estimators=125, random_state=5)

```

Hình 16: Bộ siêu tham số GradientBoostingClassifier BigDS.csv

+ Mô hình RandomForestClassifier:

```

model = RandomForestClassifier(random_state = 5)
params = {
    'max_depth': [3, 5],
    'min_samples_split': [5, 10, 15],
    'min_samples_leaf': [2, 5, 15],
    'n_estimators': [100, 150, 200]
}
model_cv = GridSearchCV(model, params, cv = 5, n_jobs = -1, verbose = False)

```

Hình 17: GridSearchCV với RandomForestClassifier BigDS.csv

Kết quả của bộ siêu tham số RandomForestClassifier

```

▼ RandomForestClassifier
RandomForestClassifier(max_depth=5, min_samples_leaf=15, min_samples_split=5,
                       n_estimators=200, random_state=5)

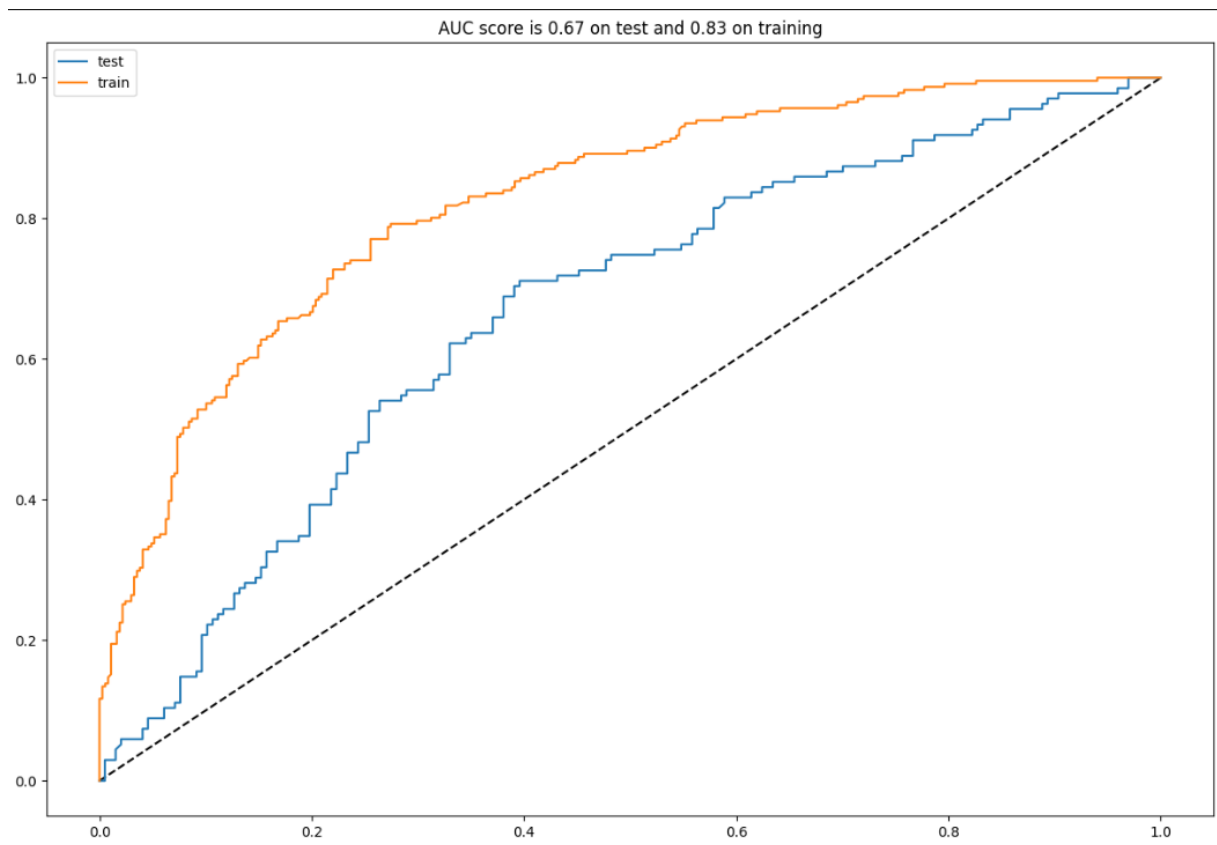
```

Hình 18: Bộ siêu tham số RandomForestClassifier BigDS.csv

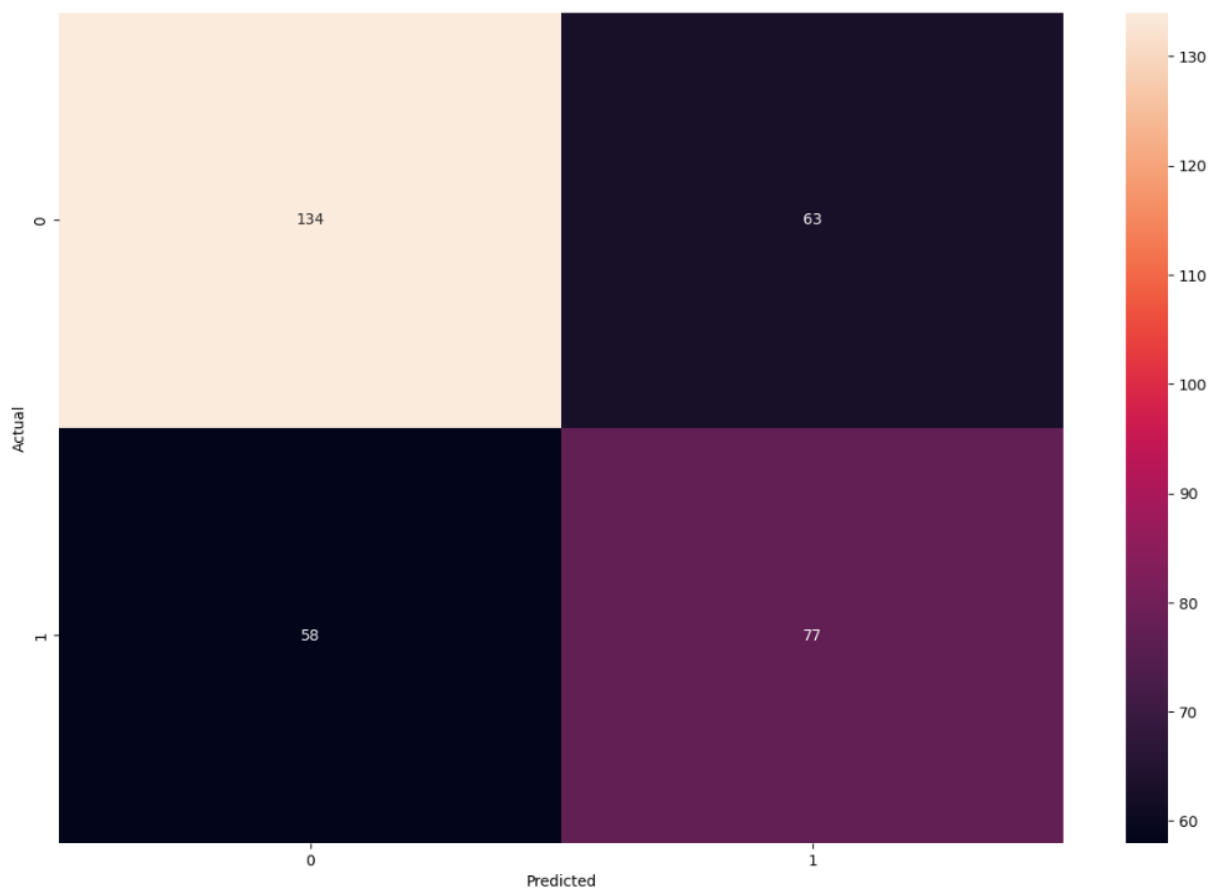
- Đồ thị thể hiện hiệu suất

+ Mô hình GradientBoostingClassifier :

SmallDS.csv

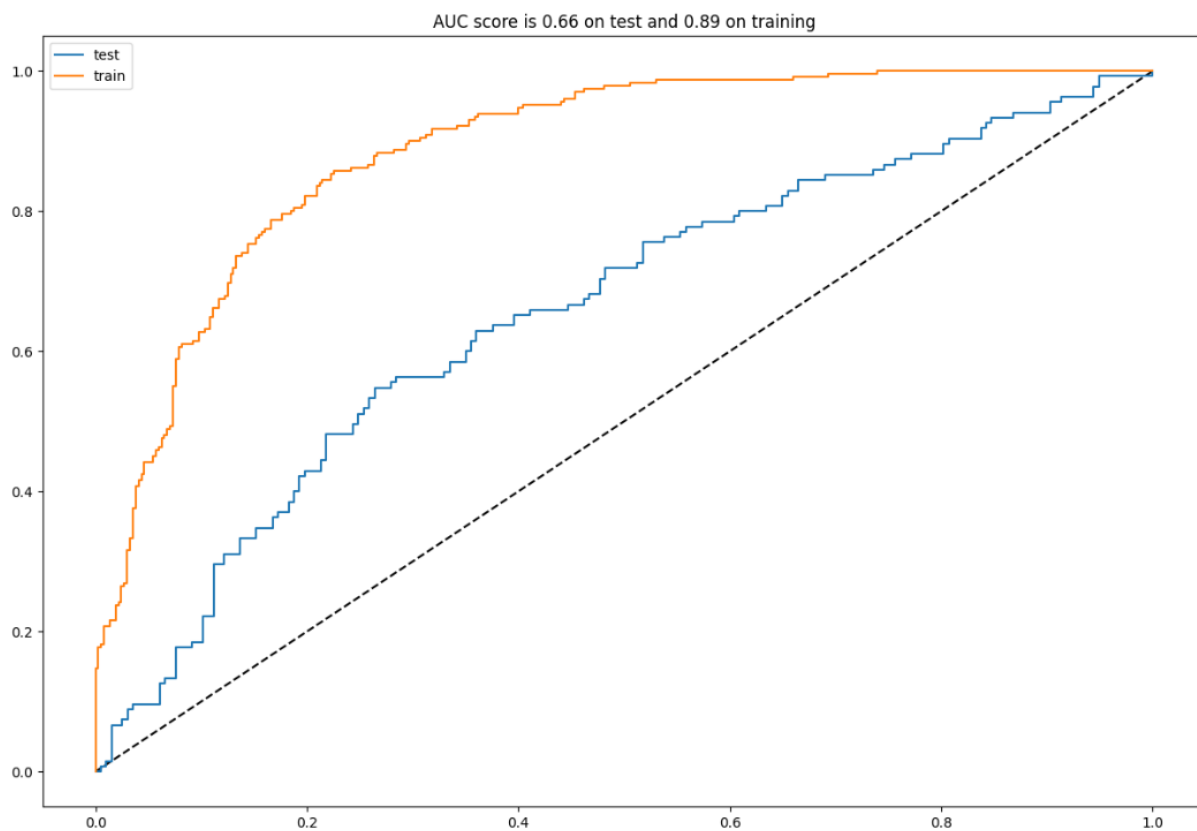


Hình 19: Đồ thị AUC trên tập test và train GradientBoostingClassifier SmallDS.csv

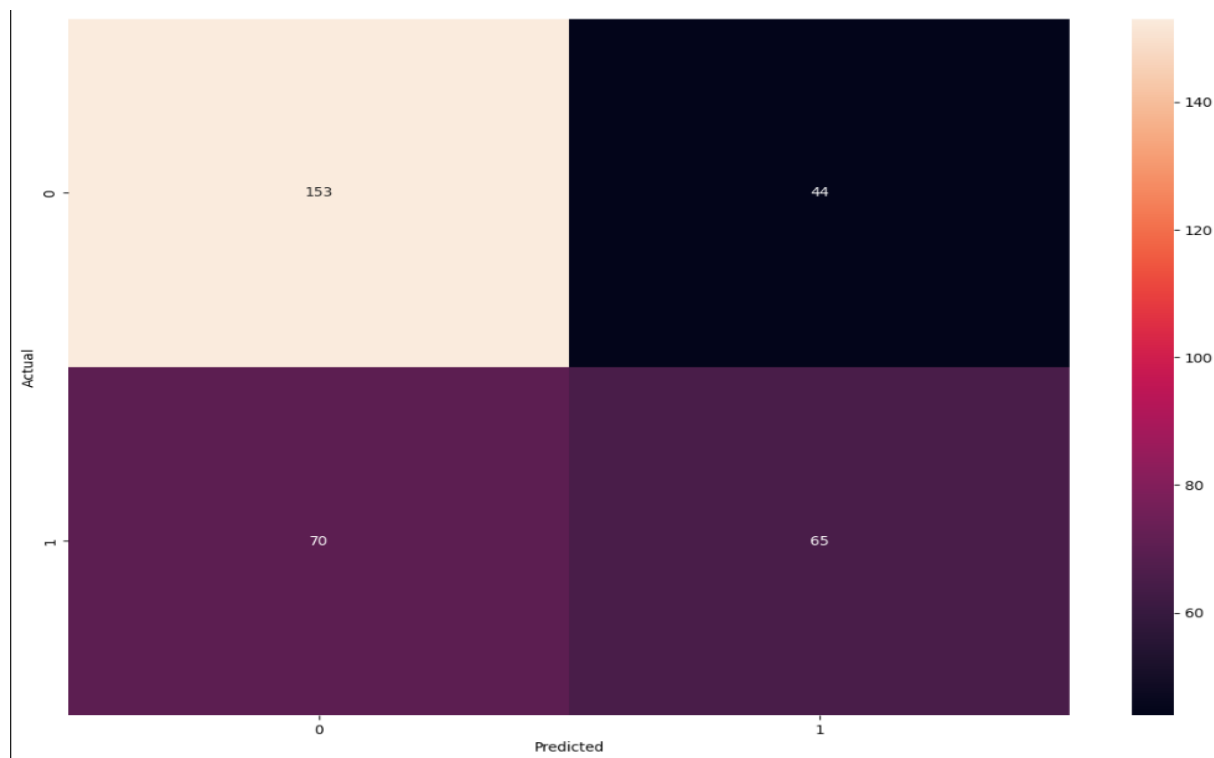


Hình 20: Đồ thị thể hiện hiệu suất GradientBoostingClassifier SmallDS.csv

## + Mô hình RandomForestClassifier

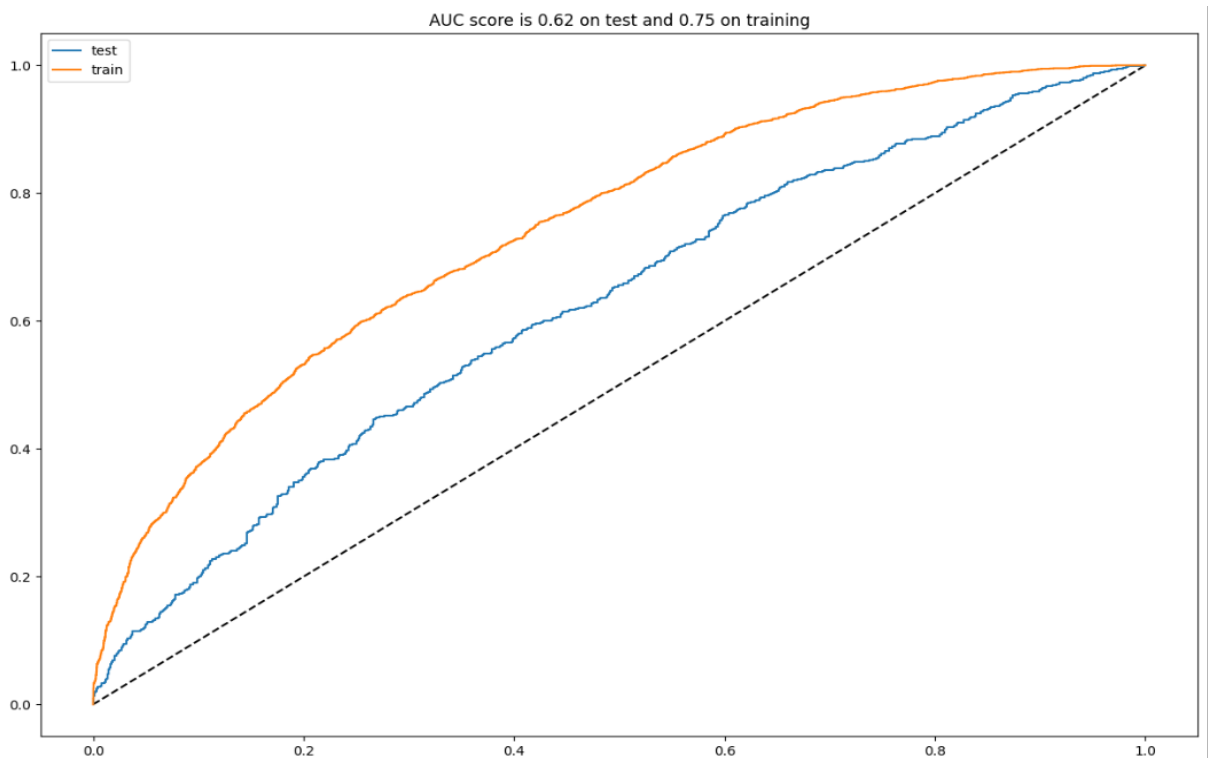


Hình 21: Đồ thị AUC trên tập test và train RandomForestClassifier SmallDS.csv

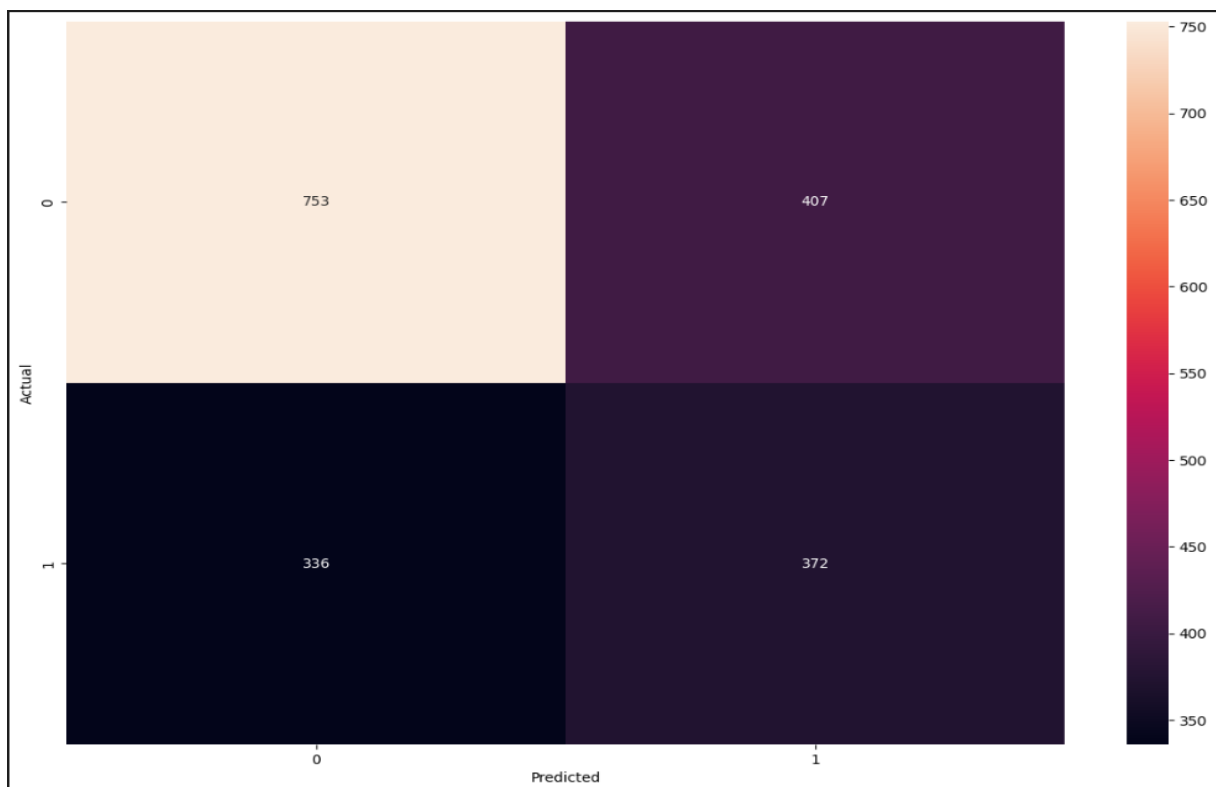


Hình 22: Đồ thị thể hiện hiệu suất RandomForestClassifier SmallDS.csv

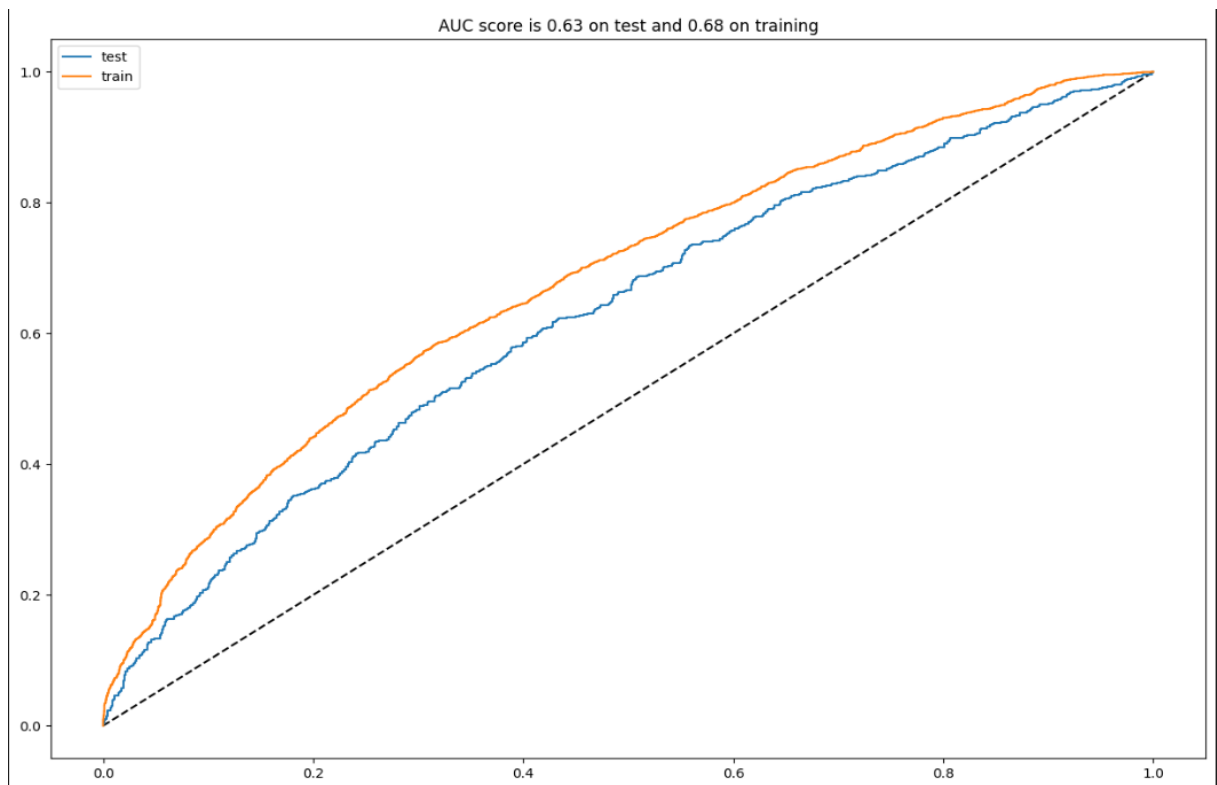




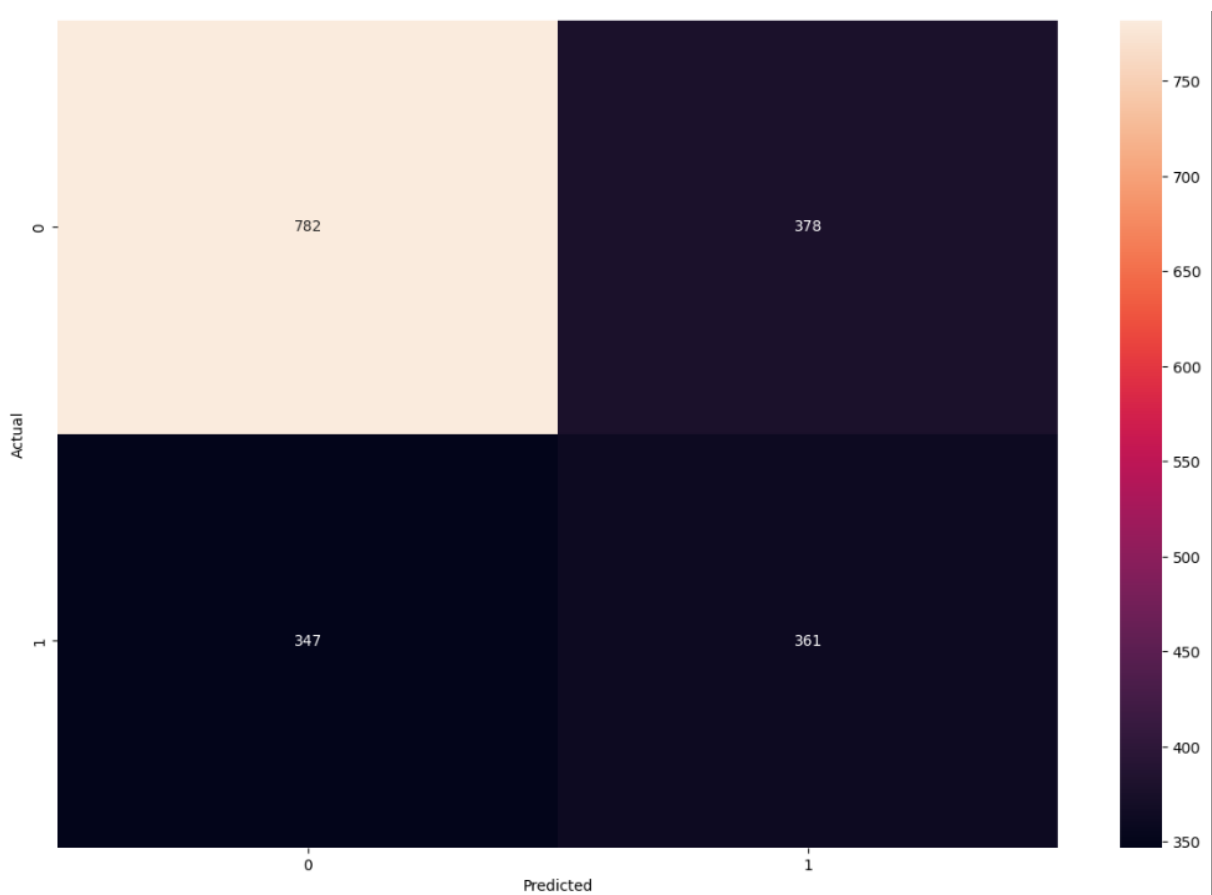
Hình 23: Đồ thị AUC trên tập test và train GradientBoostingClassifier BigDS.csv



Hình 24: Đồ thị thể hiện hiệu suất GradientBoostingClassifier BigDS.csv



Hình 25: Đồ thị AUC trên tập test và train RandomForestClassifier BigDS.csv



Hình 26: Đồ thị thể hiện hiệu suất RandomForestClassifier BigDS.csv

## V. Kết luận

### 1. Tổng kết

Sau khi thực hiện phân tích và xây dựng mô hình Machine Learning để dự đoán kết quả trận đấu bóng đá dựa trên các feature liên quan, chúng ta đã tiến hành các công việc chính sau:

- Tiền xử lí dữ liệu: Tách các feature dạng chữ ra thành feature dạng số, loại bỏ các feature không cần thiết, chọn ra các feature quan trọng từ mô hình SVM.
- Thực hiện huấn luyện và đánh giá sử dụng các mô hình Machine Learning như SVM, Random Forest và Gradient Boosting Classifier.
- Tinh chỉnh mô hình sử dụng kỹ thuật tìm kiếm siêu tham số với `GridSearchCV`.
- Đánh giá độ hiệu quả của mô hình thông qua các độ đo đánh giá như accuracy, area under the curve

### 2. Kết quả đạt được

Kết quả dự đoán so với thực tế với cột `actual` là cột thực tế, cột `predicted` là cột dự đoán SmallDS.csv

+ Mô hình GradientBoostingClassifier

	actual_x	predicted_x	date	season_x	round_x	comp_x	team_x	opponent_x	venue_x	result_x	actual_y	predicted_y	season_y	round_y	comp_y	team_y	opponent_y	venue_y	result_y
0	1	1	2022-02-10	2021	Matchweek 24	Premier League	Arsenal	Wolves	Away	W	0	1	2021	Matchweek 24	Premier League	Wolves	Arsenal	Home	L
1	1	1	2022-02-19	2021	Matchweek 26	Premier League	Arsenal	Brentford	Home	W	0	0	2021	Matchweek 26	Premier League	Brentford	Arsenal	Away	L
2	1	1	2022-02-24	2021	Matchweek 20	Premier League	Arsenal	Wolves	Home	W	0	0	2021	Matchweek 20	Premier League	Wolves	Arsenal	Away	L
3	1	1	2022-03-06	2021	Matchweek 28	Premier League	Arsenal	Watford	Away	W	0	1	2021	Matchweek 28	Premier League	Watford	Arsenal	Home	L
4	1	1	2022-03-13	2021	Matchweek 29	Premier League	Arsenal	Leicester City	Home	W	0	1	2021	Matchweek 29	Premier League	Leicester City	Arsenal	Away	L
5	0	1	2022-03-16	2021	Matchweek 27	Premier League	Arsenal	Liverpool	Home	L	1	0	2021	Matchweek 27	Premier League	Liverpool	Arsenal	Away	W
6	1	1	2022-03-19	2021	Matchweek 30	Premier League	Arsenal	Aston Villa	Away	W	0	0	2021	Matchweek 30	Premier League	Aston Villa	Arsenal	Home	L
7	0	1	2022-04-04	2021	Matchweek 31	Premier League	Arsenal	Crystal Palace	Away	L	1	0	2021	Matchweek 31	Premier League	Crystal Palace	Arsenal	Home	W
8	0	1	2022-04-09	2021	Matchweek 32	Premier League	Arsenal	Brighton	Home	L	1	0	2021	Matchweek 32	Premier League	Brighton	Arsenal	Away	W
9	0	1	2022-04-16	2021	Matchweek 33	Premier League	Arsenal	Southampton	Away	L	1	0	2021	Matchweek 33	Premier League	Southampton	Arsenal	Home	W
10	1	1	2022-04-20	2021	Matchweek 25	Premier League	Arsenal	Chelsea	Away	W	0	1	2021	Matchweek 25	Premier League	Chelsea	Arsenal	Home	L
11	1	1	2022-04-23	2021	Matchweek 34	Premier League	Arsenal	Manchester Utd	Home	W	0	1	2021	Matchweek 34	Premier League	Manchester Utd	Arsenal	Away	L
12	1	1	2022-05-01	2021	Matchweek 35	Premier League	Arsenal	West Ham	Away	W	0	0	2021	Matchweek 35	Premier League	West Ham	Arsenal	Home	L
13	1	1	2022-05-08	2021	Matchweek 36	Premier League	Arsenal	Leeds United	Home	W	0	0	2021	Matchweek 36	Premier League	Leeds United	Arsenal	Away	L
14	0	1	2022-05-12	2021	Matchweek 22	Premier League	Arsenal	Tottenham	Away	L	1	1	2021	Matchweek 22	Premier League	Tottenham	Arsenal	Home	W
15	0	0	2022-05-16	2021	Matchweek 37	Premier League	Arsenal	Newcastle Utd	Away	L	1	0	2021	Matchweek 37	Premier League	Newcastle Utd	Arsenal	Home	W
16	1	1	2022-05-22	2021	Matchweek 38	Premier League	Arsenal	Everton	Home	W	0	0	2021	Matchweek 38	Premier League	Everton	Arsenal	Away	L
17	0	0	2022-02-09	2021	Matchweek 24	Premier League	Aston Villa	Leeds United	Home	D	0	1	2021	Matchweek 24	Premier League	Leeds United	Aston Villa	Away	D
18	0	0	2022-02-13	2021	Matchweek 25	Premier League	Aston Villa	Newcastle Utd	Away	L	1	1	2021	Matchweek 25	Premier League	Newcastle Utd	Aston Villa	Home	W
19	0	0	2022-02-19	2021	Matchweek 26	Premier League	Aston Villa	Watford	Home	L	1	1	2021	Matchweek 26	Premier League	Watford	Aston Villa	Away	W
20	1	0	2022-02-26	2021	Matchweek 27	Premier League	Aston Villa	Brighton	Away	W	0	0	2021	Matchweek 27	Premier League	Brighton	Aston Villa	Home	L

Hình 27: Kết quả dự đoán GradientBoostingClassifier SmallDS.csv

Kết quả dự đoán là kết quả của từng trận : cột `predicted\_x` là dự đoán chiến thắng của team và cột `predicted\_y` là dự đoán chiến thắng của team bên phía đối đầu :

- Kết quả dự đoán chiến thắng chính xác 66.26%
- Kết quả dự đoán thua hoặc hòa 78,3%

## + Mô hình RandomForestClassifier

	actual_x	predicted_x	date	season_x	round_x	comp_x	team_x	opponent_x	venue_x	result_x	actual_y	predicted_y	season_y	round_y	comp_y	team_y	opponent_y	venue_y	result_y
0	1	1	2022-02-10	2021	Matchweek 24	Premier League	Arsenal	Wolves	Away	W	0	0	2021	Matchweek 24	Premier League	Wolves	Arsenal	Home	L
1	1	1	2022-02-19	2021	Matchweek 26	Premier League	Arsenal	Brentford	Home	W	0	0	2021	Matchweek 26	Premier League	Brentford	Arsenal	Away	L
2	1	1	2022-02-24	2021	Matchweek 20	Premier League	Arsenal	Wolves	Home	W	0	0	2021	Matchweek 20	Premier League	Wolves	Arsenal	Away	L
3	1	1	2022-03-06	2021	Matchweek 28	Premier League	Arsenal	Watford	Away	W	0	0	2021	Matchweek 28	Premier League	Watford	Arsenal	Home	L
4	1	1	2022-03-13	2021	Matchweek 29	Premier League	Arsenal	Leicester City	Home	W	0	1	2021	Matchweek 29	Premier League	Leicester City	Arsenal	Away	L
5	0	1	2022-03-16	2021	Matchweek 27	Premier League	Arsenal	Liverpool	Home	L	1	1	2021	Matchweek 27	Premier League	Liverpool	Arsenal	Away	W
6	1	1	2022-03-19	2021	Matchweek 30	Premier League	Arsenal	Aston Villa	Away	W	0	0	2021	Matchweek 30	Premier League	Aston Villa	Arsenal	Home	L
7	0	1	2022-04-04	2021	Matchweek 31	Premier League	Arsenal	Crystal Palace	Away	L	1	0	2021	Matchweek 31	Premier League	Crystal Palace	Arsenal	Home	W
8	0	1	2022-04-09	2021	Matchweek 32	Premier League	Arsenal	Brighton	Home	L	1	0	2021	Matchweek 32	Premier League	Brighton	Arsenal	Away	W
9	0	1	2022-04-16	2021	Matchweek 33	Premier League	Arsenal	Southampton	Away	L	1	0	2021	Matchweek 33	Premier League	Southampton	Arsenal	Home	W
10	1	1	2022-04-20	2021	Matchweek 25	Premier League	Arsenal	Chelsea	Away	W	0	1	2021	Matchweek 25	Premier League	Chelsea	Arsenal	Home	L

Hình 28: Kết quả dự đoán RandomForestClassifier SmallDS.csv

Kết quả dự đoán là kết quả của từng trận : cột `predicted\_x` là dự đoán chiến thắng của team và cột `predicted\_y` là dự đoán chiến thắng của team bên phía đối đầu :

- Kết quả dự đoán chiến thắng chính xác 62.76%
- Kết quả dự đoán thua hoặc hòa 77.66%

## BigDS.csv

## + Mô hình GradientBoostingClassifier

	actual_x	predicted_x	date	season_x	round_x	comp_x	team_x	opponent_x	venue_x	result_x	actual_y	predicted_y	season_y	round_y	comp_y	team_y	opponent_y	venue_y	result_y
0	0	0	2021-08-13	2021	Matchweek 1	Premier League	Arsenal	Brentford	Away	L	1	1	2021	Matchweek 1	Premier League	Brentford	Arsenal	Home	W
1	0	1	2021-08-22	2021	Matchweek 2	Premier League	Arsenal	Chelsea	Home	L	1	1	2021	Matchweek 2	Premier League	Chelsea	Arsenal	Away	W
2	0	0	2021-08-28	2021	Matchweek 3	Premier League	Arsenal	Manchester City	Away	L	1	1	2021	Matchweek 3	Premier League	Manchester City	Arsenal	Home	W
3	1	1	2021-09-11	2021	Matchweek 4	Premier League	Arsenal	Norwich City	Home	W	0	0	2021	Matchweek 4	Premier League	Norwich City	Arsenal	Away	L
4	1	0	2021-09-18	2021	Matchweek 5	Premier League	Arsenal	Burnley	Away	W	0	1	2021	Matchweek 5	Premier League	Burnley	Arsenal	Home	L
5	1	1	2021-09-26	2021	Matchweek 6	Premier League	Arsenal	Tottenham	Home	W	0	0	2021	Matchweek 6	Premier League	Tottenham	Arsenal	Away	L
6	0	0	2021-10-02	2021	Matchweek 7	Premier League	Arsenal	Brighton	Away	D	0	0	2021	Matchweek 7	Premier League	Brighton	Arsenal	Home	D
7	0	0	2021-10-18	2021	Matchweek 8	Premier League	Arsenal	Crystal Palace	Home	D	0	0	2021	Matchweek 8	Premier League	Crystal Palace	Arsenal	Away	D
8	1	0	2021-10-22	2021	Matchweek 9	Premier League	Arsenal	Aston Villa	Home	W	0	0	2021	Matchweek 9	Premier League	Aston Villa	Arsenal	Away	L
9	1	0	2021-10-30	2021	Matchweek 10	Premier League	Arsenal	Leicester City	Away	W	0	1	2021	Matchweek 10	Premier League	Leicester City	Arsenal	Home	L
10	1	1	2021-11-07	2021	Matchweek 11	Premier League	Arsenal	Watford	Home	W	0	0	2021	Matchweek 11	Premier League	Watford	Arsenal	Away	L

Hình 29: Kết quả dự đoán GradientBoostingClassifier BigDS.csv

Kết quả dự đoán là kết quả của từng trận : cột `predicted\_x` là dự đoán chiến thắng của team và cột `predicted\_y` là dự đoán chiến thắng của team bên phía đối đầu :

- Kết quả dự đoán chiến thắng chính xác 51.89%
- Kết quả dự đoán thua hoặc hòa 76,65%

## + Mô hình RandomForestClassifier

	actual_x	predicted_x	date	season_x	round_x	comp_x	team_x	opponent_x	venue_x	result_x	actual_y	predicted_y	season_y	round_y	comp_y	team_y	opponent_y	venue_y	result_y
0	0	0	2021-08-13	2021	Matchweek 1	Premier League	Arsenal	Brentford	Away	L	1	1	2021	Matchweek 1	Premier League	Brentford	Arsenal	Home	W
1	0	1	2021-08-22	2021	Matchweek 2	Premier League	Arsenal	Chelsea	Home	L	1	1	2021	Matchweek 2	Premier League	Chelsea	Arsenal	Away	W
2	0	0	2021-08-28	2021	Matchweek 3	Premier League	Arsenal	Manchester City	Away	L	1	1	2021	Matchweek 3	Premier League	Manchester City	Arsenal	Home	W
3	1	0	2021-09-11	2021	Matchweek 4	Premier League	Arsenal	Norwich City	Home	W	0	0	2021	Matchweek 4	Premier League	Norwich City	Arsenal	Away	L
4	1	0	2021-09-18	2021	Matchweek 5	Premier League	Arsenal	Burnley	Away	W	0	0	2021	Matchweek 5	Premier League	Burnley	Arsenal	Home	L
5	1	1	2021-09-26	2021	Matchweek 6	Premier League	Arsenal	Tottenham	Home	W	0	0	2021	Matchweek 6	Premier League	Tottenham	Arsenal	Away	L
6	0	0	2021-10-02	2021	Matchweek 7	Premier League	Arsenal	Brighton	Away	D	0	0	2021	Matchweek 7	Premier League	Brighton	Arsenal	Home	D
7	0	0	2021-10-18	2021	Matchweek 8	Premier League	Arsenal	Crystal Palace	Home	D	0	0	2021	Matchweek 8	Premier League	Crystal Palace	Arsenal	Away	D
8	1	1	2021-10-22	2021	Matchweek 9	Premier League	Arsenal	Aston Villa	Home	W	0	0	2021	Matchweek 9	Premier League	Aston Villa	Arsenal	Away	L
9	1	0	2021-10-30	2021	Matchweek 10	Premier League	Arsenal	Leicester City	Away	W	0	1	2021	Matchweek 10	Premier League	Leicester City	Arsenal	Home	L
10	1	1	2021-11-07	2021	Matchweek 11	Premier League	Arsenal	Watford	Home	W	0	0	2021	Matchweek 11	Premier League	Watford	Arsenal	Away	L

Hình 30: Kết quả dự đoán RandomForestClassifier BigDS.csv

Kết quả dự đoán là kết quả của từng trận : cột `predicted\_x` là dự đoán chiến thắng của team và cột `predicted\_y` là dự đoán chiến thắng của team bên phía đối đầu :

- Kết quả dự đoán chiến thắng chính xác 52.24%
- Kết quả dự đoán thua hoặc hòa 76.12%

### 3. Nhận xét đánh giá

- Tỷ lệ chia dữ liệu giữa trận thắng và không thắng mất cân bằng dẫn đến kết quả dự đoán bị lệch sang dự đoán có thiên hướng lệch sang một bên.
- Đạt kết quả cao hơn trên tập dữ liệu SmallDS.csv vì tập BigDS.csv sử dụng kết quả của nhiều mùa giải và nhiều giải đấu nên ảnh hưởng đến kết quả dự đoán .
- Mô hình GradientBoostingClassifier đạt kết quả dự đoán thấp hơn mô hình RandomForestClassifier tuy nhiên cả TPR( True Positive Rate) train GradientBoostingClassifier đạt kết quả cao hơn RandomForestClassifier

### 4. Hướng phát triển

Tuy nhiên, để cải thiện kết quả hiện tại của mô hình dự đoán đội thắng trận đấu bóng đá, chúng ta có thể thực hiện một số hướng phát triển như sau:

- Tăng cường dữ liệu: Từ các trận đấu trước đó có thể sinh ra thêm dữ liệu ở các điều kiện thực tế khác nhau để có được dữ liệu đầy đủ hơn để huấn luyện và phát triển mô hình.
- Tinh chỉnh các tham số mô hình: Có thể tinh chỉnh các tham số của mô hình để tăng cường khả năng dự đoán. Ví dụ, tăng số lượng cây trong thuật toán Random Forest, điều chỉnh các siêu tham số của Gradient Boosting, hay tăng độ sâu của cây quyết định.
- Tổng hợp lại, để cải thiện kết quả dự đoán đội thắng trong một trận đấu bóng đá, chúng ta có thể thực hiện một hoặc một số hướng phát triển như trên để phát triển mô hình tốt hơn và chính xác hơn.

## VI. Tài liệu tham khảo

[https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html?fbclid=IwAR2wwjDQdazbo6M3S0gJ4m\\_nO2OZhMCCDNwUope\\_oHO9xTYyTdSxjEnYUdc](https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html?fbclid=IwAR2wwjDQdazbo6M3S0gJ4m_nO2OZhMCCDNwUope_oHO9xTYyTdSxjEnYUdc)

<https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?fbclid=IwAR1AToug4SF678Dmm75675bzuF49CrYjCqh2ZMJnG8fRL8WQZYGnig1JnLY>