DỰ ĐOÁN ĐỘI THẮNG TRONG TRẬN ĐẤU BÓNG ĐÁ

NHÓM 4

Danh sách thành viên

Vũ Tiến Hùng

Tìm nguồn dữ liệu .

Crawl dữ liệu.

Làm sạch dữ liệu.

Thêm các đặc trưng.

Lựa chọn đặc trưng.

Võ Chí Tài

Xử lý dữ liệu.

Đặc trưng hóa dữ liệu.

Viết báo cáo.

Nguyễn Dương Gia Bảo

Tìm hiểu về 2 mô hình.

Tìm bộ siêu tham số phù hợp 2 mô hình.

Thực thi huấn luyện mô hình.

Đánh giá kết quả.

Mục tiêu và giải pháp

- MỤC TIÊU

Tăng cường khả năng dự đoán kết quả của các trận đấu bóng đá, giúp 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.

- GIẢI PHÁP

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 RamdomForestClassifier trên hai tập dữ liệu lớn và nhỏ để tìm ra kết quả, đánh giá và so sánh với nhau.

Thu thập dữ liệu

Dữ liệu được thu thập:

- Nguồn dữ liệu : https://fbref.com/en
- Công cụ thu thập : Thư viện Python - BeautifulSoup

Dữ liệu thu thập gồm : Xếp hạng mùa trước đội bóng ở một giải bóng đá và các thông số liên quan tới trận đấu và các cú sút của các đội bóng trong các giải đấu ở Anh trong 6 năm gần nhất

Shooting 2022-2023 Manchester City: All Competitions Share & Export ▾ Glossary

For Manchester City | Against Manchester City

For Manchester City										Standard										Expected				
Date	Time	Comp	Round	Day	Venue	Result	GF	GA	Opponent	Gls	Sh	SoT	SoT%	G/Sh	G/SoT	Dist	FK	PK	PKatt	xG	npxG	npxG/Sh	G-xG	np:G-xG
2022-07-30	17:00 (23:00)	Community Shield	FA Community Shield	Sat	Neutral	L	1	3	Liverpool	1	14	8	57.1	0.07	0.13			0	0					
2022-08-07	16:30 (22:30)	Premier League	Matchweek 1	Sun	Away	W	2	0	West Ham	2	13	1	7.7	0.08	1.00	18.7	1	1	1	2.2	1.4	0.11	-0.2	-0.4
2022-08-13	15:00 (21:00)	Premier League	Matchweek 2	Sat	Home	W	4	0	Bournemouth	3	19	7	36.8	0.16	0.43	17.5	0	0	0	1.7	1.7	0.09	+1.3	+1.3
2022-08-21	16:30 (22:30)	Premier League	Matchweek 3	Sun	Away	D	3	3	Newcastle Utd	3	21	10	47.6	0.14	0.30	16.2	1	0	0	2.1	2.1	0.10	+0.9	+0.9
2022-08-27	15:00 (21:00)	Premier League	Matchweek 4	Sat	Home	W	4	2	Crystal Palace	4	18	5	27.8	0.22	0.80	14.1	0	0	0	2.2	2.2	0.13	+1.8	+1.8
2022-08-31	19:30 (01:30)	Premier League	Matchweek 5	Wed	Home	W	6	0	Nott'ham Forest	6	17	9	52.9	0.35	0.67	14.8	0	0	0	3.3	3.3	0.20	+2.7	+2.7
2022-09-03	17:30 (23:30)	Premier League	Matchweek 6	Sat	Away	D	1	1	Aston Villa	1	13	4	30.8	0.08	0.25	16.8	1	0	0	2.1	2.1	0.16	-1.1	-1.1
2022-09-06	21:00 (02:00)	Champions Lg	Group stage	Tue	Away	W	4	0	Sevilla	4	24	9	37.5	0.17	0.44	16.9	1	0	0	3.6	3.6	0.16	+0.4	+0.4
2022-09-14	20:00 (02:00)	Champions Lg	Group stage	Wed	Home	W	2	1	Dortmund	2	12	3	25.0	0.17	0.67	17.3	0	0	0	1.0	1.0	0.08	+1.0	+1.0
2022-09-17	12:30 (18:30)	Premier League	Matchweek 8	Sat	Away	W	3	0	Wolves	3	16	7	43.8	0.19	0.43	19.4	0	0	0	1.1	1.1	0.07	+1.9	+1.9

Trích xuất đặc trưng

- Làm sạch dữ liệu :
  - ✓ Lựa chọn các đặc trưng phù hợp cho việc dự đoán.
  - ✓ Chuyển đổi dữ liệu ở các cột về một mẫu thống nhất.
  - ✓ Kết quả đạt được : 2 file SmallDS.csv và BigDS.csv

- Thêm các đặc trưng mới :
  - ✓ Thêm các đặc trưng : Chênh lệch xếp hạng của các đội bóng , điểm và chênh lệch điểm của đội bóng trong các giải đấu.

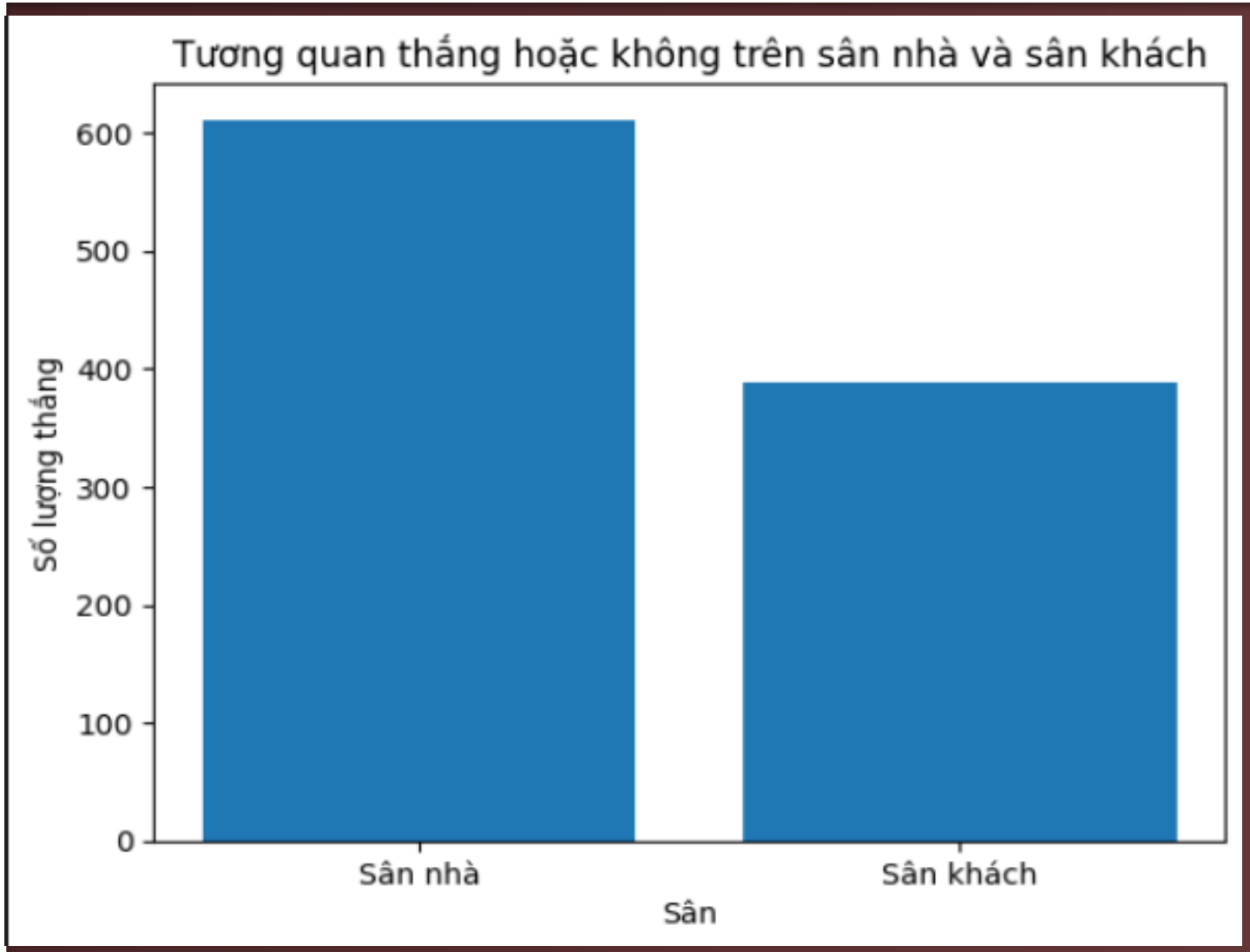
- Xử lý dữ liệu :
  - ✓ Xử lý dữ liệu trống
  - ✓ Chuyển đổi dữ liệu:

Chuyển đổi một số đặc trưng về dạng số.

Nhãn được sử dụng ‘target’ là 1 nếu kết quả trận đấu là thắng còn hòa và thua là 0

- Lựa chọn các đặc trưng :

Trực quan hóa dữ liệu

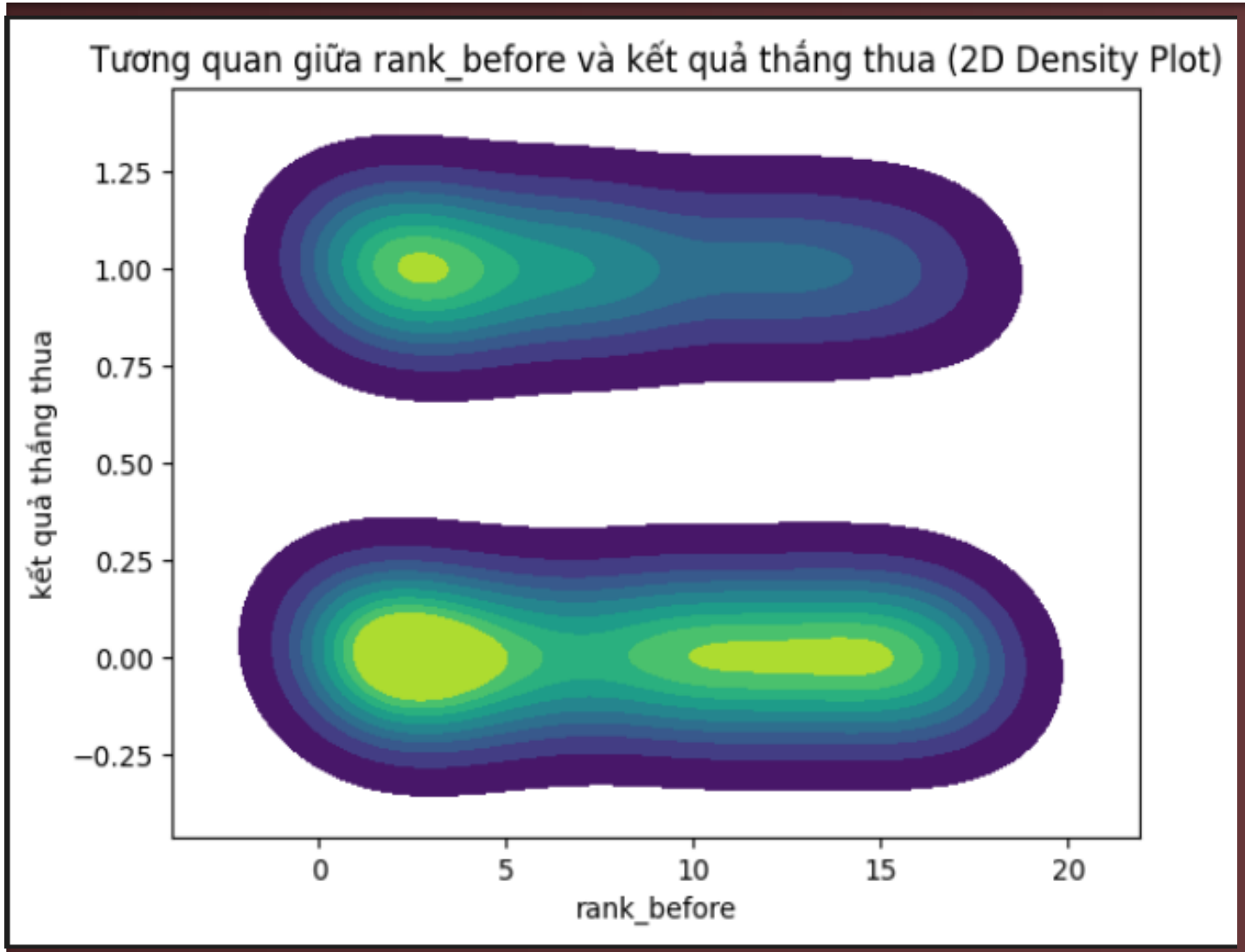


NHẬN XÉT

- Sân nhà thì chiếm tỉ lệ thắng cao hơn sân khách

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

Trực quan hóa dữ liệu

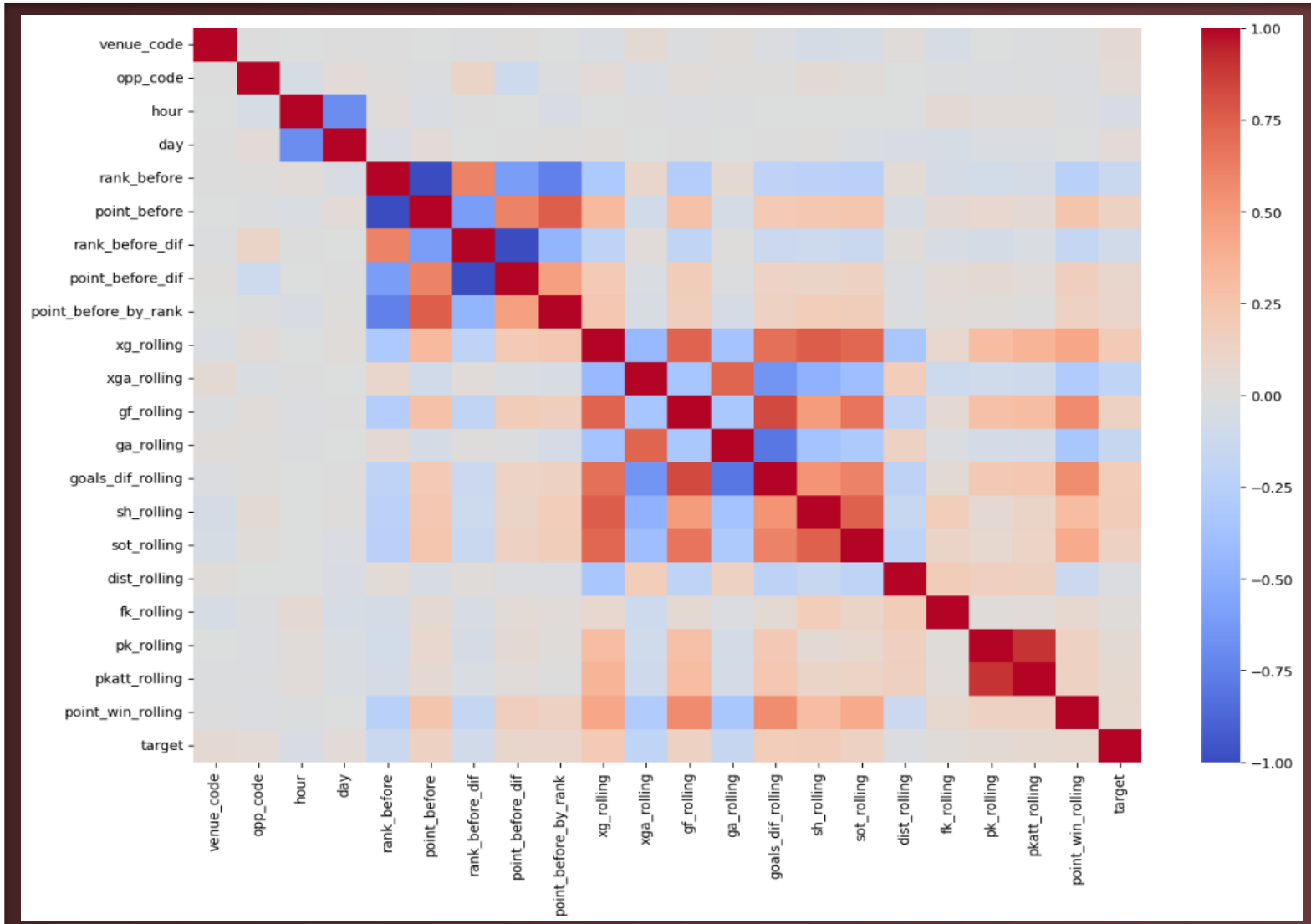


NHẬN XÉT

- Ta thấy dữ liệu khi rank cao (1-5) thì dữ liệu tập trung ở target 1(thắng) và vẫn có một vài dữ liệu rank cao nhưng vẫn đạt kết quả target 0 (Thua hoặc hòa)

Hình 9: Biểu đồ tương quan giữa rank_before và kết quả thua

Trực quan hóa dữ liệu



NHẬN XÉT

- 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

Mô hình hóa dữ liệu

- Các mô hình sử dụng:
  - ✓ Gradient Boosting Classifier
  - ✓ Ramdom Forest Classifier

Mô hình hóa dữ liệu

- Gradient Boosting Classifier

  - ✓ Bộ tham số chính là

    - ❑ Learning rate
    - ❑ min_samples_split
    - ❑ min_samples_leaf
    - ❑ max_depth

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

Bộ siêu tham số GradientBoostringClassifier SmallDS.csv

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

Bộ siêu tham số GradientBoostringClassifier BigDS.csv

Mô hình hóa dữ liệu

- Ramdom Forest Classifier

  - ✓ Bộ tham số chính là
    - ❑ n_estimators
    - ❑ max_depth
    - ❑ min_samples_split
    - ❑ min_samples_leaf

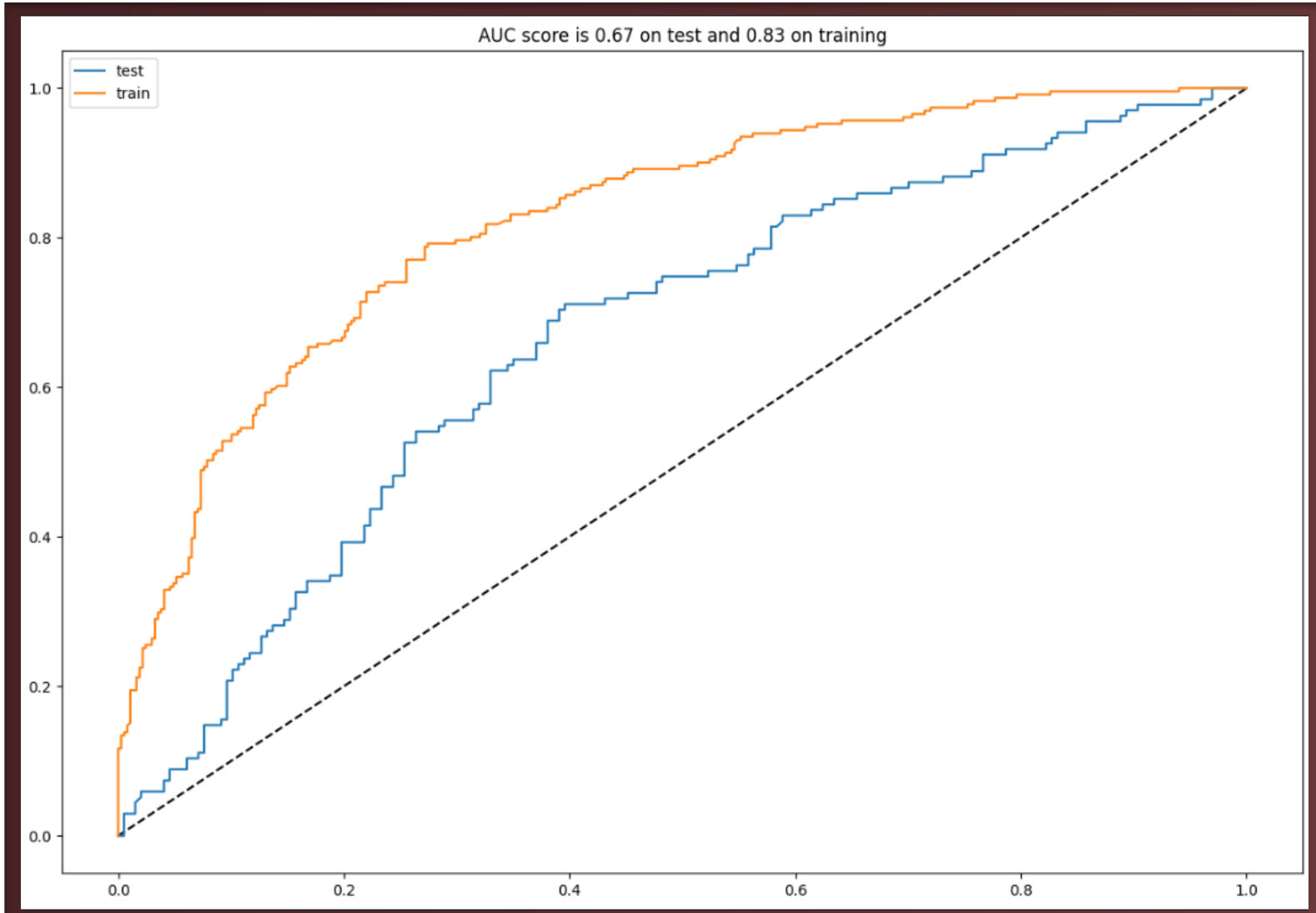
```
RandomForestClassifier
RandomForestClassifier(max_depth=5, min_samples_leaf=15, min_samples_split=5,
                       n_estimators=200, random_state=5)
```

Bộ siêu tham số RandomForestClassifier SmallDS.csv

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

Bộ siêu tham số RandomForestClassifier BigDS.csv

Đồ thị thể hiện hiệu suất

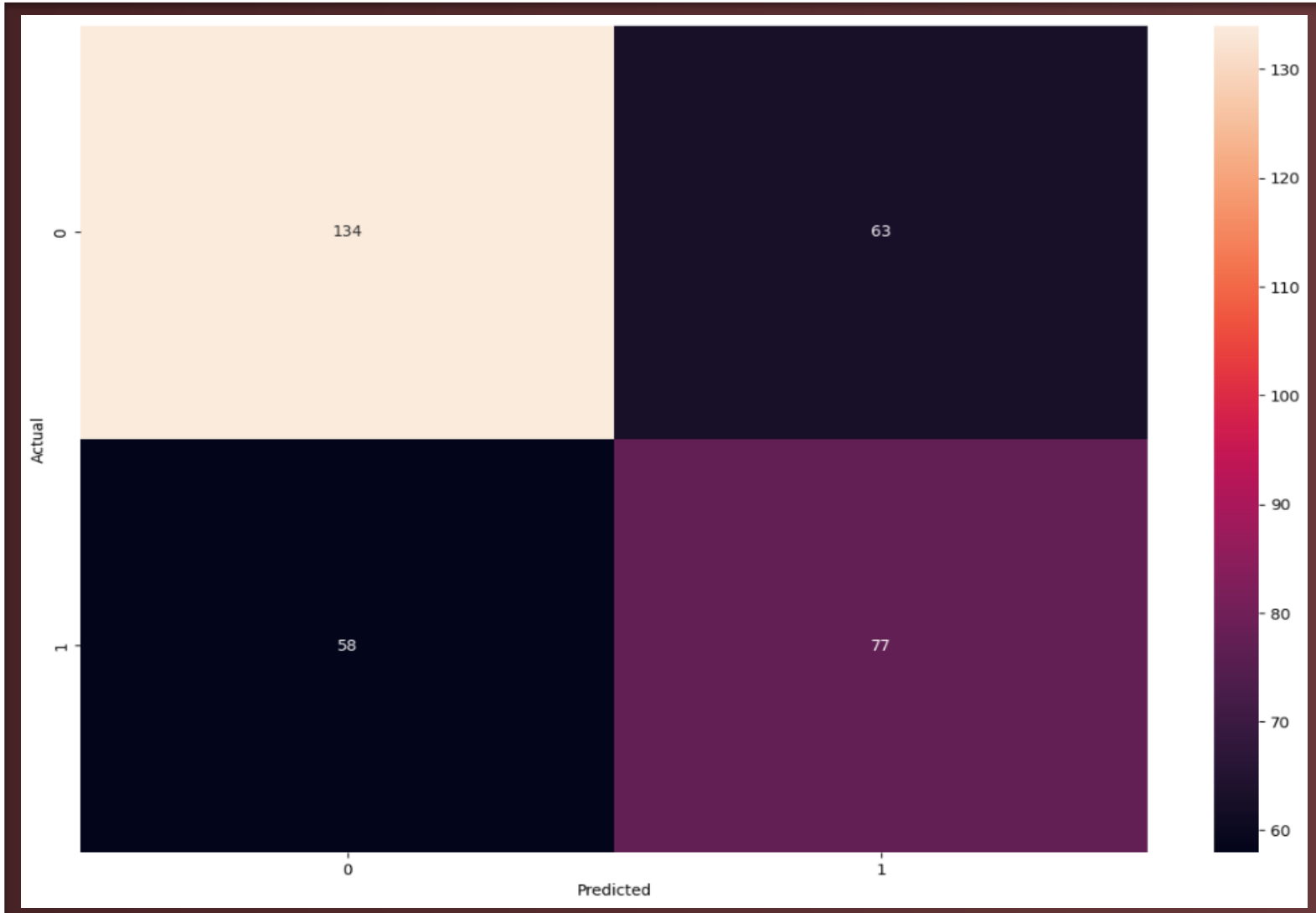


Nhận xét:

- AUC score trên tập test là 0.67 và trên tập train là 0.83
- Trục Ox: FPR = FP/ (FP + TN)
- Trục Oy: TPR = TP/ (TP + FN)
- Trong đó
- FP : kết quả dự đoán trận không thắng sai
- TN: kết quả dự đoán trận thắng đúng
- TP: kết quả dự đoán trận không thắng đúng
- FN: kết quả dự đoán trận thắng sai

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

Đồ thị thể hiện hiệu suất

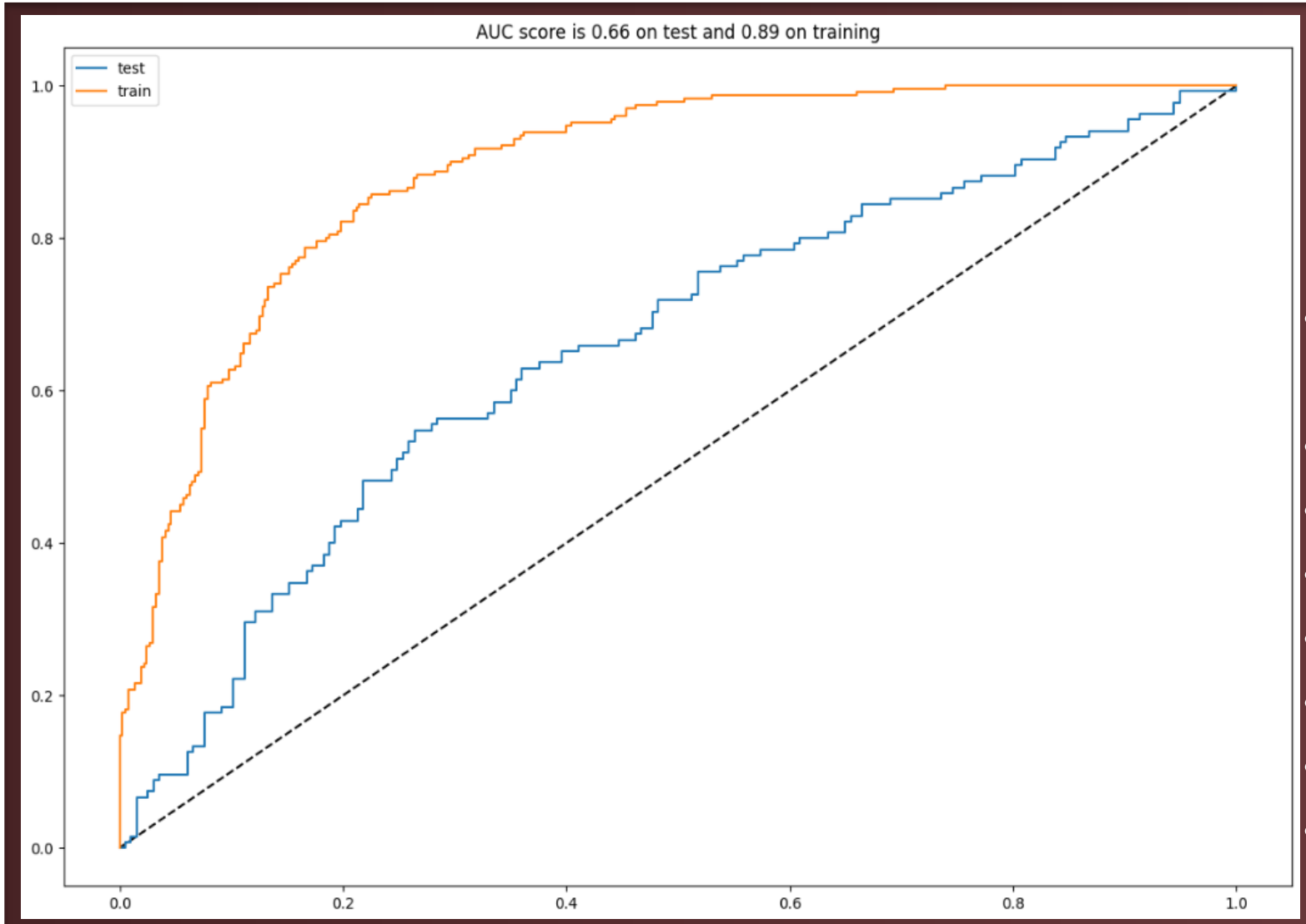


Nhận xét:

- Trục Ox là kết quả dự đoán
- Trục Oy là kết quả thực
- Kết quả dự đoán trận không thắng đúng là 134
- Kết quả dự đoán trận thắng đúng là 77
- Kết quả dự đoán trận thắng sai là 58
- Kết quả dự đoán trận không thắng sai là 63

Đồ thị thể hiện hiệu suất GradientBoostingClassifier SmallDS.csv

Đồ thị thể hiện hiệu suất

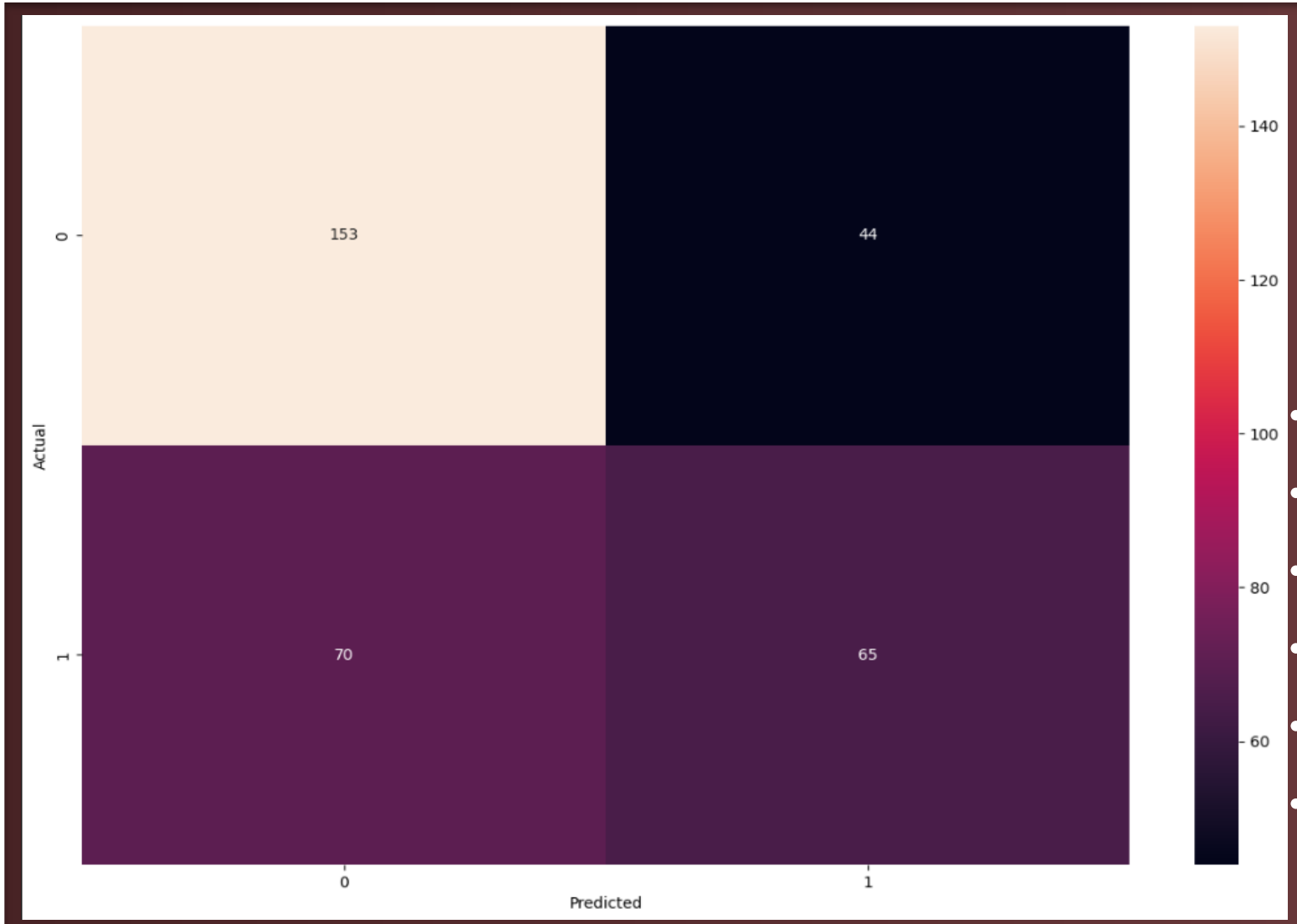


Nhận xét:

- AUC score trên tập test là 0.66 và trên tập train là 0.89
- Trục Ox: FPR = FP/ (FP + TN)
- Trục Oy: TPR = TP/ (TP + FN)
- Trong đó
- FP : kết quả dự đoán trận không thắng sai
- TN: kết quả dự đoán trận thắng đúng
- TP: kết quả dự đoán trận không thắng đúng
- FN: kết quả dự đoán trận thắng sai

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

Đồ thị thể hiện hiệu suất

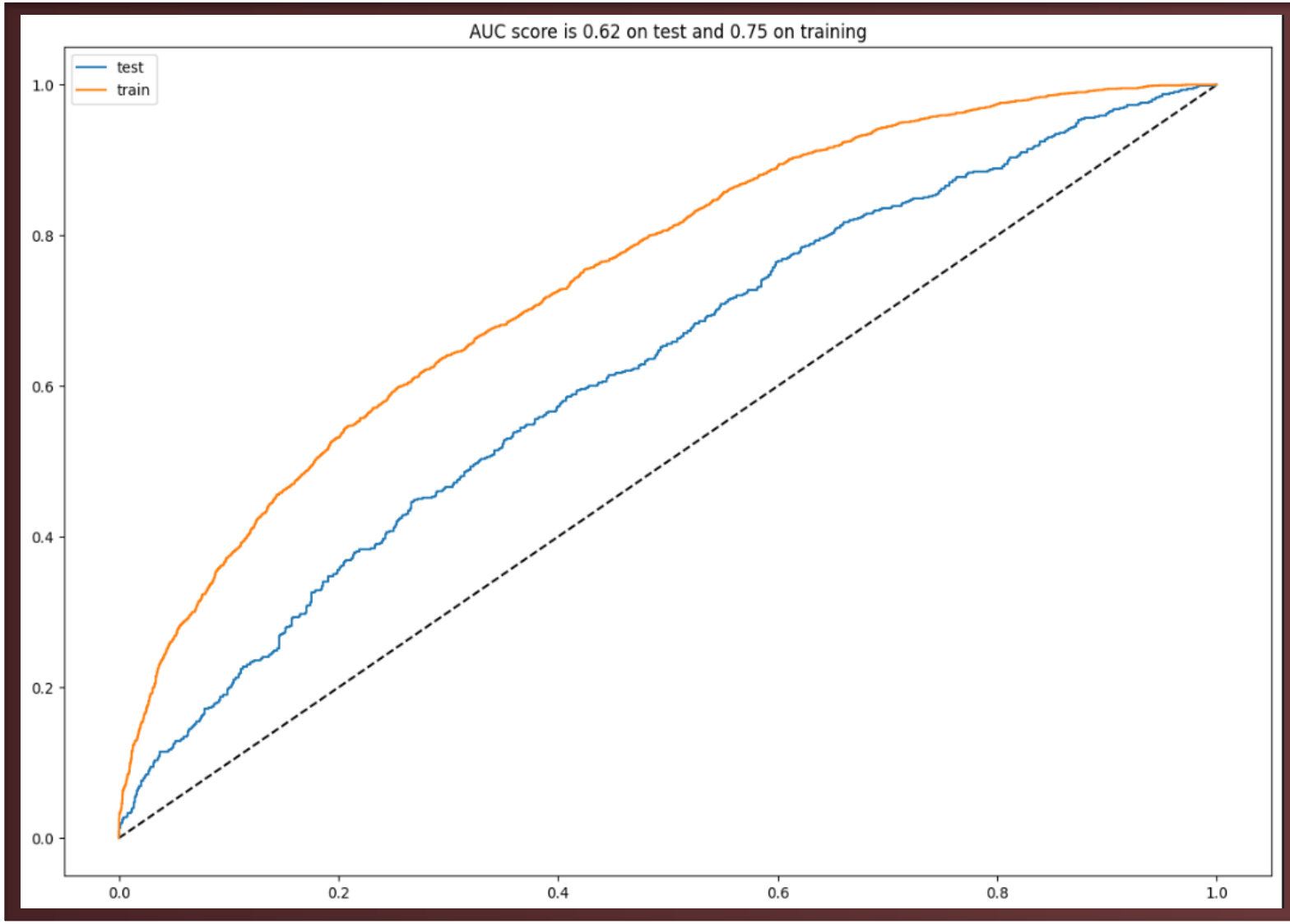


Nhận xét:

- Trục Ox là kết quả dự đoán
- Trục Oy là kết quả thực
- Kết quả dự đoán trận không thắng đúng là 153
- Kết quả dự đoán trận thắng đúng là 65
- Kết quả dự đoán trận thắng sai là 44
- Kết quả dự đoán trận không thắng sai là 70

Đồ thị thể hiện hiệu suất RandomForestClassifier SmallDS.csv

Đồ thị thể hiện hiệu suất

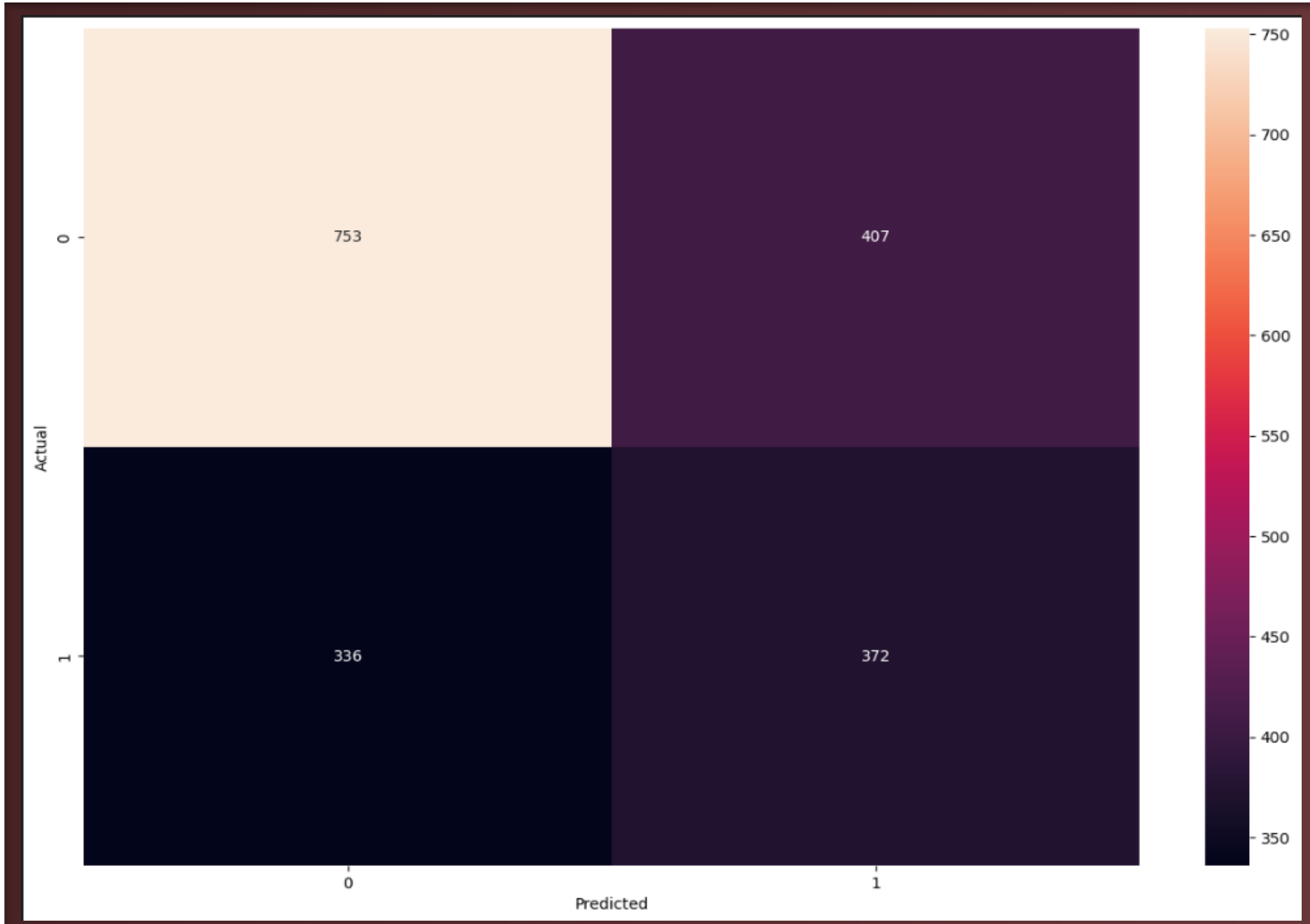


Nhận xét:

- AUC score trên tập test là 0.62 và trên tập train là 0.75
- Trục Ox: FPR = FP/ (FP + TN)
- Trục Oy: TPR = TP/ (TP + FN)
- Trong đó
- FP : kết quả dự đoán trận không thắng sai
- TN: kết quả dự đoán trận thắng đúng
- TP: kết quả dự đoán trận không thắng đúng
- FN: kết quả dự đoán trận thắng sai

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

Đồ thị thể hiện hiệu suất

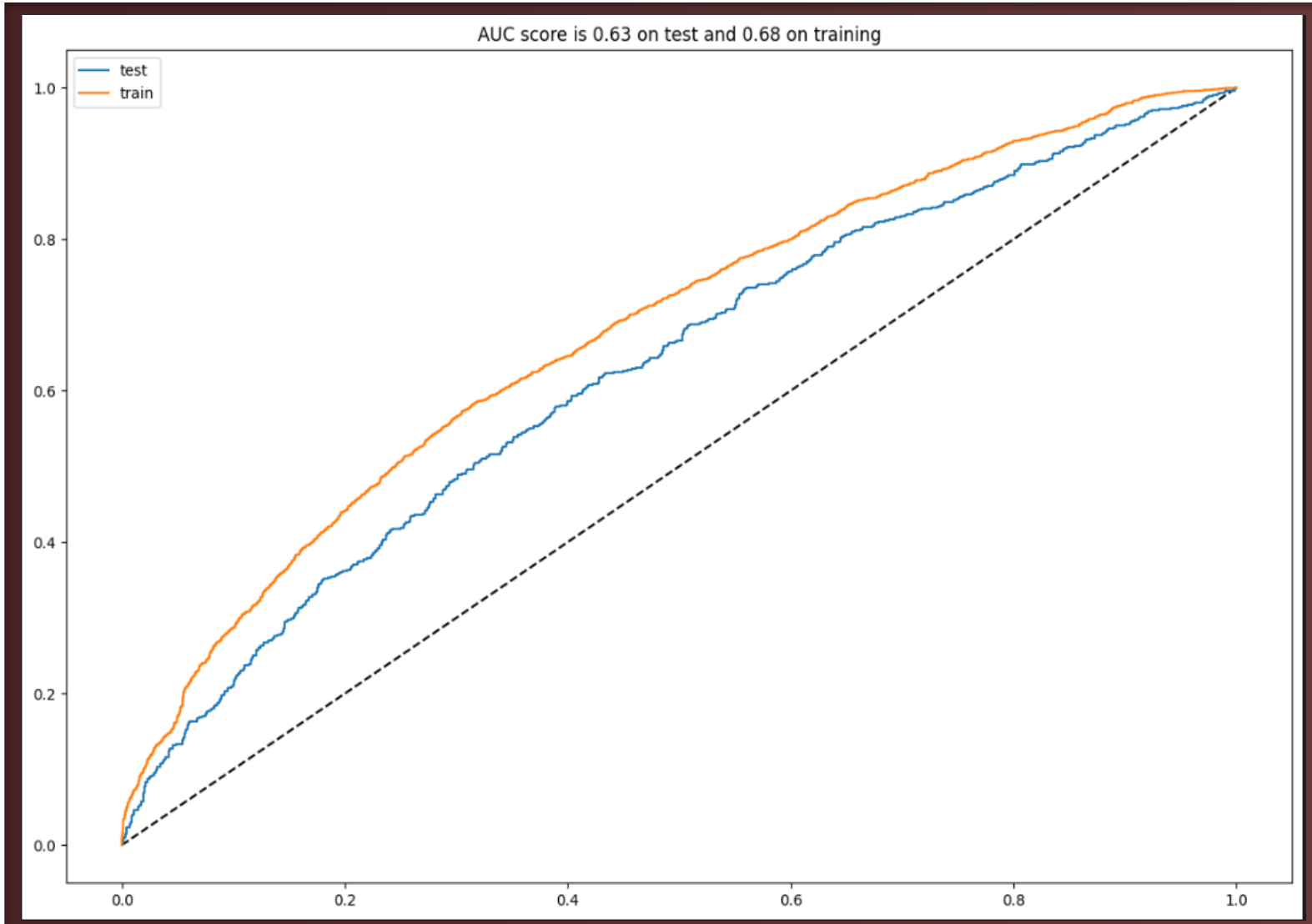


Nhận xét:

- Trục Ox là kết quả dự đoán
- Trục Oy là kết quả thực
- Kết quả dự đoán trận không thắng đúng là 753
- Kết quả dự đoán trận thắng đúng là 372
- Kết quả dự đoán trận thắng sai là 407
- Kết quả dự đoán trận không thắng sai là 336

Đồ thị thể hiện hiệu suất GradientBoostingClassifier BigDS.csv

Đồ thị thể hiện hiệu suất

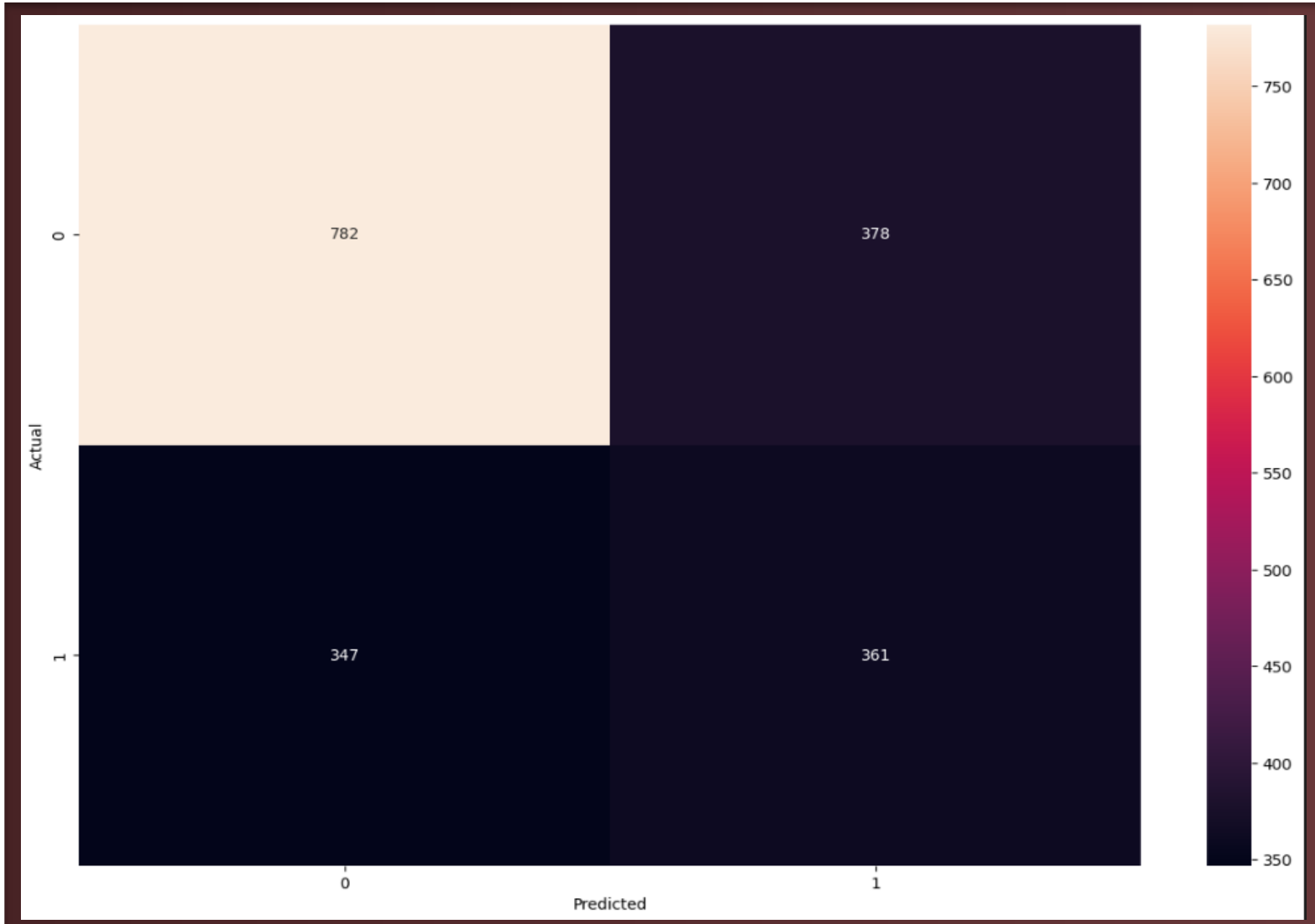


Nhận xét:

- AUC score trên tập test là 0.63 và trên tập train là 0.68
- Trục Ox: FPR = FP/ (FP + TN)
- Trục Oy: TPR = TP/ (TP + FN)
- Trong đó
- FP : kết quả dự đoán trận không thắng sai
- TN: kết quả dự đoán trận thắng đúng
- TP: kết quả dự đoán trận không thắng đúng
- FN: kết quả dự đoán trận thắng sai

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

Đồ thị thể hiện hiệu suất



Nhận xét:

- Trục Ox là kết quả dự đoán
- Trục Oy là kết quả thực
- Kết quả dự đoán trận không thắng đúng là 782
- Kết quả dự đoán trận thắng đúng là 361
- Kết quả dự đoán trận thắng sai là 378
- Kết quả dự đoán trận không thắng sai là 347

Đồ thị thể hiện hiệu suất RandomForestClassifier BigDS.csv

Kết quả đạt được

Mô hình Gradient Boosting Classifier

SmallDS.csv

- Kết quả dự đoán chính xác chiến thắng 66.26%
- Kết quả dự đoán chính xác không thắng 78,3%
- Accuracy : 63,5%

BigDS.csv

- Kết quả dự đoán chính xác chiến thắng 51.89%
- Kết quả dự đoán chính xác không thắng 76,65%
- Accuracy : 60,2%

Mô hình Random Forest Classifier

SmallDS.csv

- Kết quả dự đoán chính xác chiến thắng 62.76%
- Kết quả dự đoán chính xác không thắng 77,66%
- Accuracy : 65,6%

BigDS.csv

- Kết quả dự đoán chiến thắng chính xác 52.24%
- Kết quả dự đoán thua hoặc hòa 76,12%
- Accuracy : 61,18%

Kết luận

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 trong quá khứ và nhiều giải đấu nên ảnh hưởng đến kết quả dự đoán .
- Mô hình GradientBoostingClassifier đạt kết quả dự đoán thấp hơn mô hình Random ForestClassifier tuy nhiên cả TPR (True Positive Rate) trên GradientBoostingClassifier đạt kết quả cao hơn RandomForestClassifier

Thank you!