DỰ ĐOÁN KẾT QUẢ BÓNG ĐÁ BẰNG THUẬT TOÁN MÁY HỌC

Huỳnh Minh Trí - 18520176 - CS114.K21.KHTN

Link Github:

https://github.com/hmtrii/CS114.K21.KHTN

I. Giới thiệu bài toán

- Bài toán dự đoán kết quả của một trận đấu bóng đá bằng cách sử dụng các thuật toán máy học.
- Dữ liệu dùng để dự đoán: thông tin của 2 đội bóng, bao gồm:
  - Chỉ số đánh giá cầu thủ của FIFA.
  - Những thống kê về lịch sử đấu 2 đội bóng.
- Bài toán phân loại 3 nhãn: thắng, hòa hoặc thua của đội chủ nhà.

II. Xây dựng ứng dụng máy học

1.Thu thập dữ liệu

2.Tiền xử lý dữ liệu

3.Các thuật toán máy học

4.Training

5.Đánh giá

6.Tuning các tham số

7.Dự đoán trên dữ liệu mới

1. Thu thập dữ liệu

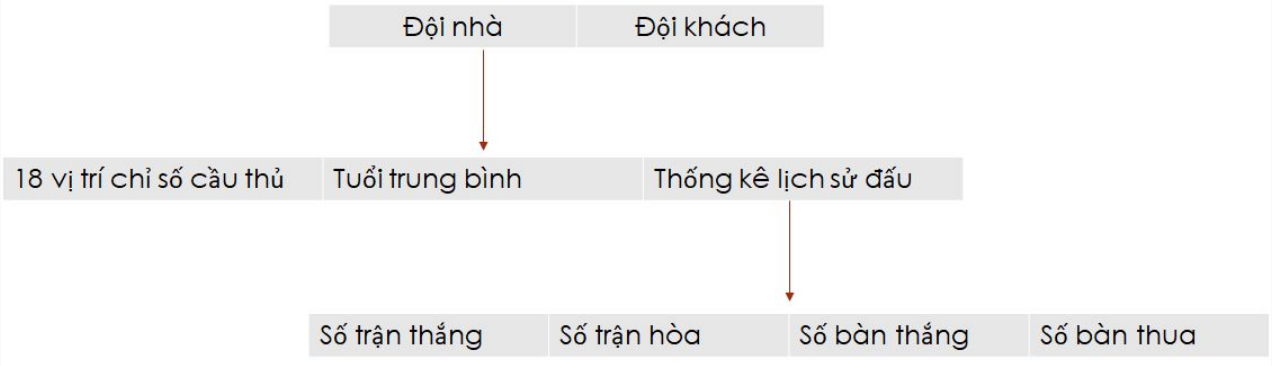
- Sử dụng dữ liệu trong 10 mùa giải (2010-2020) của 4 giải vô địch quốc gia hàng đầu Châu Âu bao gồm: Premier League (Anh), Laliga (Tây Ban Nha), SerieA (Ý) và Bundesliga (Đức).
- Bộ dữ liệu được thu thập gồm 2 nhóm chính:
  - Dữ liệu của các cầu thủ
  - Dữ liệu của các trận đấu
- Sử dụng công cụ **Web Scraper** và **Scrapy** để lấy dữ liệu về từ 2 trang web: https://www.fifaindex.com/ và https://www.skysports.com/

1. Thu thập dữ liệu

- Dữ liệu của các cầu thủ có thành phần chủ yếu là các chỉ số đánh giá từ trò chơi ***FIFA***. Bên cạnh đó còn bao gồm dữ liệu về tên cầu thủ, tên đội đang thi đấu, vị trí thi đấu, số áo thi đấu, tuổi và mùa giải thi đấu tương ứng.
- Dữ liệu trận đấu bao gồm tên 2 đội bóng, thời điểm thi đấu, đội hình ra sân và số bàn thắng mà 2 đội ghi được.
- Kích thước bộ dữ liệu: 15.045 trận đấu và 26.227 cầu thủ.

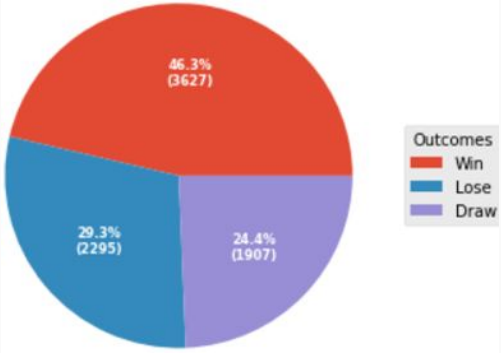
2. Tiền xử lý dữ liệu

- Chuẩn hóa dữ liệu:
  - Xử lý khác biệt về tên đội bóng và số đội bóng ở 2 nhóm dữ liệu.
  - Quản lý các đối tượng: cầu thủ, đội bóng, trận đấu.
- Xây dựng đặc trưng:



2. Tiền xử lý dữ liệu

- Phân chia dữ liệu
  - Train/Validation: 80/20, dùng các trận đấu của mùa giải 2010-2019, bao gồm 13.679 trận đấu.
  - Test: các trận đấu mùa 2020, bao gồm 1.366 trận đấu.
- Sự phân bố của các nhãn trong tập train/validation:



2. Tiền xử lý dữ liệu

- Các đặc trưng về chỉ số đánh giá cầu thủ có giá trị dao động từ 50 đến 90 trong khi các đặc trưng khác như tỉ lệ trận thắng trong 5 trận gần nhất có giá trị từ 0 -1 => Tiến hành scale bộ dữ liệu về khoảng 0 – 1.
- Sử dụng cả 2 bộ dữ liệu chưa được scaled và đã được scaled. Để tìm xem dữ liệu nào sẽ phù hợp cho thuật toán nào và giúp tìm ra cách xử lý dữ liệu phù hợp nhất để đạt kết quả cao nhất cho bài toán.

3. Các thuật toán máy học

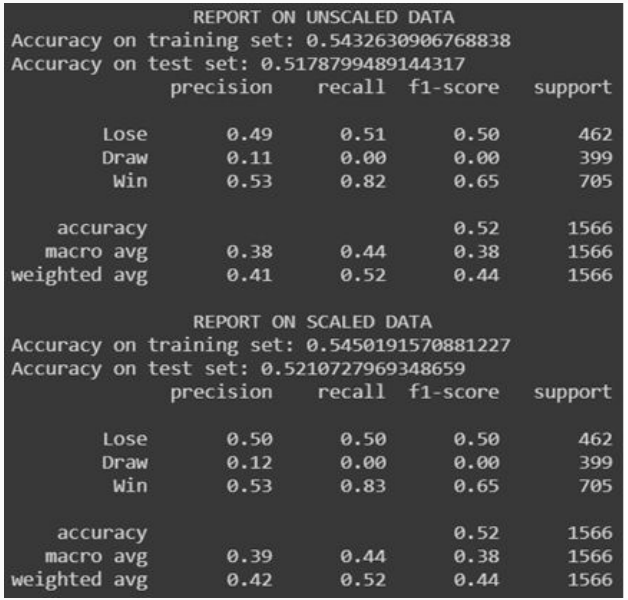
- K-Nearest-Neighbors (k-NN)
- Logistic Regression
- Support Vector Machines (SVMs)
- Neural Network (MLPs)

4. Training

Sử dụng các model với những tham số mặc định từ Scikit-learn.

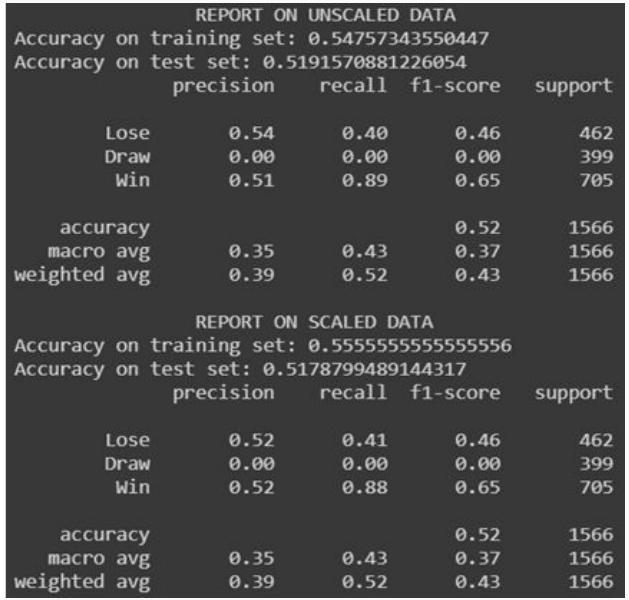
5. Đánh giá

Logistic Regression và SVMs có hiệu quả cao nhất với f1-score = 0.52



```
REPORT ON UNSCALED DATA
Accuracy on training set: 0.5432630906768838
Accuracy on test set: 0.5178799489144317
              precision    recall  f1-score   support

        Lose       0.49      0.51      0.50       462
        Draw       0.11      0.00      0.00       399
         Win       0.53      0.82      0.65       705

    accuracy                           0.52      1566
   macro avg       0.38      0.44      0.38      1566
weighted avg       0.41      0.52      0.44      1566

REPORT ON SCALED DATA
Accuracy on training set: 0.5450191570881227
Accuracy on test set: 0.5210727969348659
              precision    recall  f1-score   support

        Lose       0.50      0.50      0.50       462
        Draw       0.12      0.00      0.00       399
         Win       0.53      0.83      0.65       705

    accuracy                           0.52      1566
   macro avg       0.39      0.44      0.38      1566
weighted avg       0.42      0.52      0.44      1566
```

Logistic Regression



```
REPORT ON UNSCALED DATA
Accuracy on training set: 0.54757343550447
Accuracy on test set: 0.5191570881226054
              precision    recall  f1-score   support

        Lose       0.54      0.40      0.46       462
        Draw       0.00      0.00      0.00       399
         Win       0.51      0.89      0.65       705

    accuracy                           0.52      1566
   macro avg       0.35      0.43      0.37      1566
weighted avg       0.39      0.52      0.43      1566

REPORT ON SCALED DATA
Accuracy on training set: 0.5555555555555556
Accuracy on test set: 0.5178799489144317
              precision    recall  f1-score   support

        Lose       0.52      0.41      0.46       462
        Draw       0.00      0.00      0.00       399
         Win       0.52      0.88      0.65       705

    accuracy                           0.52      1566
   macro avg       0.35      0.43      0.37      1566
weighted avg       0.39      0.52      0.43      1566
```

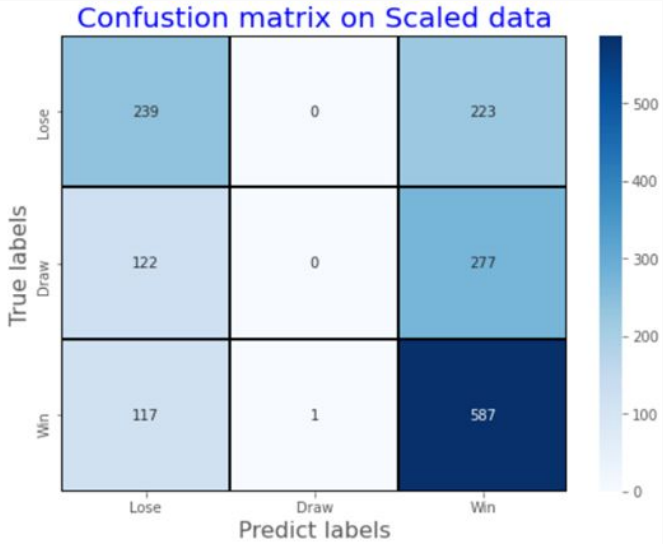
SVMs

6. Tuning các tham số

- Sau khi tuning, Neural Network cho kết quả tốt nhất trên bộ dữ liệu đã được scale.
- f1-score = 0.53

REPORT ON SCALED DATA
Accuracy on training set: 0.5429438058748404
Accuracy on test set: 0.5274584929757343

	precision	recall	f1-score	support
Lose	0.50	0.52	0.51	462
Draw	0.00	0.00	0.00	399
Win	0.54	0.83	0.66	705
accuracy			0.53	1566
macro avg	0.35	0.45	0.39	1566
weighted avg	0.39	0.53	0.44	1566

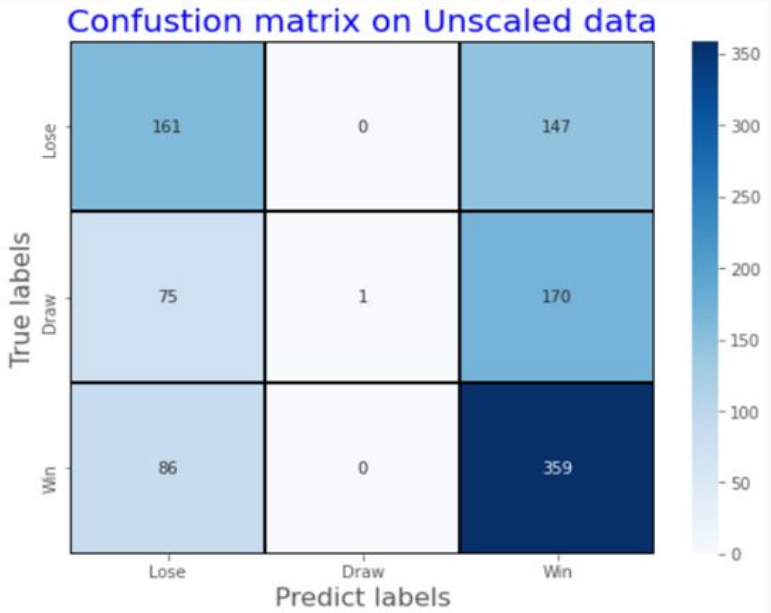


7. Dự đoán trên dữ liệu mới

Áp dụng model Neural network xây dựng được với các tham số tốt nhất trên dữ liệu các mùa giải 2019-2020:

Accuracy: 0.5215215215215215

	precision	recall	f1-score	support
Lose	0.50	0.52	0.51	308
Draw	1.00	0.00	0.01	246
Win	0.53	0.81	0.64	445
accuracy			0.52	999
macro avg	0.68	0.44	0.39	999
weighted avg	0.64	0.52	0.44	999



III. Kết luận

- Bài toán đạt được f1-score = 0.52 trên bộ dữ liệu mới.
- Sau quá trình tuning các tham số, kết quả dự đoán vẫn không tăng quá đáng kể => cần cải thiện bộ dữ liệu.
  - Đưa vào thêm nhiều đặc trưng: phong độ thi đấu, tình hình sức khỏe của từng cầu thủ, yếu tố đánh giá về chiến thuật, huấn luận viên, thời tiết, sân vận động,... Cần có sự phân biệt giữa các giải đấu khác nhau.
  - Khi tune tham số, độ đa dạng về giá trị của các tham số còn ít => thêm số lượng các layers và số node của mỗi layer trong Neural Network.

Tài liệu tham khảo

- Andreas C. Müller & Sarah Guido. Introduction to Machine Learning with Python. 2017.
- Alexandre Bucquet & Vishnu Sarukkai. The Bank is Open: AI in Sports Gambling. 2018.
- Bradley. Predicting Football Matches using EA Player Ratings and Tensorflow. 2018.
- Scikit-learn. https://scikit-learn.org/stable/#


CẢM ƠN!