HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN I

BÁO CÁO BÀI TẬP LỚN

MÔN: LẬP TRÌNH VỚI PYTHON

GIẢNG VIÊN	**: KIM NGỌC BÁCH**
SINH VIÊN	**: THÁI DUY TIẾN**
MÃ SINH VIÊN	**: B22DCCN727**
NHÓM	**: 11**

Hà Nội – 11/2024

Lời nói đầu

Trong thời đại công nghệ số hiện nay, việc thu thập và xử lý dữ liệu đã trở thành một phần không thể thiếu trong nhiều lĩnh vực. Với sự phát triển mạnh mẽ của Internet, các doanh nghiệp ngày càng chú trọng đến việc phân tích dữ liệu tiếp cận khách hàng một cách hiệu quả và tiện lợi hơn.

Bài tập lớn môn Python này được thực hiện với mục tiêu thu thập dữ liệu từ internnet và xử lý dữ liệu sử dụng ngôn ngữ lập trình Python. Dự án này không chỉ giúp sinh viên nắm vững các kiến thức cơ bản về lập trình với Python mà còn cung cấp những kỹ năng thực tế trong việc xử lý dữ liệu .

Trong quá trình thực hiện dự án, chúng tôi sẽ áp dụng ngôn ngữ Python và các thư viện hỗ trợ của Python như Pandas, **Matplotlib** và **Seaborn.**

Chúng tôi hy vọng rằng, thông qua dự án này, các sinh viên sẽ có cơ hội áp dụng những kiến thức đã học vào thực tế, từ đó nâng cao kỹ năng lập trình và chuẩn bị tốt hơn cho công việc trong tương lai.

PHÁT BIỂU BÀI TOÁN:

PHÂN TÍCH VÀ ĐỊNH GIÁ CẦU THỦ DỰA TRÊN DỮ LIỆU MÙA GIẢI PREMIER LEAGUE 2023-2024

Mục tiêu: Xây dựng một chương trình Python để thu thập, phân tích dữ liệu thống kê cầu thủ từ mùa giải Premier League 2023-2024 và sử dụng các thuật toán máy học để phân loại và định giá cầu thủ.

Các bước thực hiện chính:

1. **Thu thập dữ liệu thống kê cầu thủ**:
    - Thu thập dữ liệu thống kê từ trang web **fbref.com** cho các cầu thủ đã chơi trên 90 phút trong mùa giải Premier League 2023-2024.
    - Các chỉ số cần thu thập gồm: quốc tịch, đội bóng, vị trí, tuổi, thời gian thi đấu, hiệu suất ghi bàn, kiến tạo, số thẻ phạt, chỉ số kỳ vọng (xG, xAG...), chỉ số chuyền bóng, phòng ngự, kiểm soát bóng và nhiều chỉ số chi tiết khác.
2. **Lưu trữ và sắp xếp dữ liệu**:
    - Ghi dữ liệu thu thập được ra file results.csv.
    - Sắp xếp danh sách cầu thủ theo tên và tuổi, điền giá trị "N/a" nếu chỉ số không có sẵn hoặc không áp dụng.
3. **Phân tích dữ liệu thống kê cầu thủ**:
    - Tìm top 3 cầu thủ có điểm cao nhất và thấp nhất ở từng chỉ số.
    - Tính các chỉ số thống kê như trung vị, trung bình và độ lệch chuẩn cho từng chỉ số của cầu thủ trong toàn giải và từng đội.
    - Tìm đội bóng có chỉ số cao nhất ở mỗi chỉ số và nhận xét đội có phong độ tốt nhất.
4. **Phân cụm cầu thủ sử dụng K-means**:
    - Sử dụng thuật toán **K-means** để phân các cầu thủ thành các nhóm có chỉ số tương đồng.
    - Áp dụng thuật toán giảm số chiều **PCA** để giảm số chiều dữ liệu xuống 2D, hỗ trợ trực quan hóa phân cụm trên mặt phẳng 2D.
5. **Vẽ biểu đồ và so sánh cầu thủ**:
    - Vẽ biểu đồ histogram phân bố của mỗi chỉ số trên toàn giải và từng đội.
    - Viết chương trình vẽ biểu đồ radar để so sánh các cầu thủ dựa trên danh sách các chỉ số cụ thể được chỉ định.
6. **Thu thập dữ liệu và định giá cầu thủ**:
    - Thu thập giá trị chuyển nhượng từ trang **footballtransfers.com**.
    - Đề xuất mô hình định giá cầu thủ dựa trên các chỉ số quan trọng như tuổi, số bàn thắng, số kiến tạo, số phút thi đấu, v.v., với phương pháp hồi quy tuyến tính hoặc cây quyết định để dự đoán giá trị chuyển nhượng.

I. Giải Quyết Bài Toán:

A. Bài Tập 1:

1. Giới thiệu

Chương trình Python được xây dựng nhằm thu thập dữ liệu thống kê của các cầu thủ từ giải Ngoại hạng Anh mùa 2023-2024, từ đó phân tích hiệu suất của họ dựa trên các chỉ số quan trọng. Dữ liệu được thu thập từ trang fbref.com, một nguồn đáng tin cậy và phổ biến về thống kê bóng đá, với mục tiêu lọc ra các cầu thủ có thời gian thi đấu trên 90 phút.

2. Cấu trúc chương trình

Chương trình gồm hai phần chính:

- **Thu thập dữ liệu từ trang fbref.com**: Dữ liệu được lấy từ nhiều trang khác nhau, mỗi trang cung cấp các chỉ số thống kê khác nhau như số phút thi đấu, bàn thắng, chuyền bóng, và các chỉ số phòng ngự.
- **Xử lý và kết hợp dữ liệu**: Các tệp dữ liệu từ các trang khác nhau được làm sạch và kết hợp thành một bảng chính duy nhất. Kết quả cuối cùng được lưu trong tệp result.csv.

3. Các hàm chính

- **Hàm crawler**:
  - **Chức năng**: Thu thập dữ liệu từ một URL nhất định và lưu dưới dạng tệp CSV riêng.
  - **Chi tiết hoạt động**: Gửi yêu cầu HTTP GET đến URL fbref.com, sau đó sử dụng BeautifulSoup để phân tích và trích xuất dữ liệu từ các bình luận HTML (commented HTML), vì dữ liệu bảng thường được ẩn dưới dạng comment. Từ đó, chương trình trích xuất các hàng và cột dữ liệu cần thiết vào từ điển ans1.
  - **Đầu ra**: Tạo các tệp CSV table2.csv đến table10.csv cho từng phần dữ liệu.
- **Hàm clean_data**:
  - **Chức năng**: Làm sạch và kết hợp dữ liệu từ nhiều tệp CSV thành một bảng chính.
  - **Chi tiết hoạt động**: Hàm này đọc các tệp CSV table1.csv đến table10.csv, sau đó loại bỏ các cột trùng lặp không cần thiết và kết hợp các bảng dựa trên cột chung. Hàm cũng lọc ra các cầu thủ có số phút thi đấu trên 90 và sắp xếp danh sách theo tên cầu thủ và độ tuổi.
  - **Đầu ra**: Lưu dữ liệu đã làm sạch vào tệp result.csv để phục vụ phân tích.

4. Quy trình thực hiện chương trình

- **Khởi tạo**: Chương trình bắt đầu bằng việc lấy URL của trang fbref.com chứa dữ liệu thống kê cầu thủ cho mùa giải 2023-2024.
- **Thu thập dữ liệu**: Sử dụng hàm crawler với từng URL và ID của từng phần dữ liệu để lấy dữ liệu cầu thủ từ các trang như thủ môn, dứt điểm, chuyền bóng, phòng ngự, v.v.
- **Làm sạch dữ liệu**: Hàm clean_data thực hiện việc lọc cầu thủ và sắp xếp danh sách theo yêu cầu.
- **Lưu trữ dữ liệu**: Kết quả cuối cùng được lưu vào result.csv, chứa dữ liệu thống kê hoàn chỉnh cho các cầu thủ đạt yêu cầu về thời gian thi đấu.

5. Kết quả

Chương trình đã tạo tệp result.csv, bao gồm các cầu thủ có thời gian thi đấu hơn 90 phút, được sắp xếp theo tên và độ tuổi. Các chỉ số không áp dụng hoặc không có sẵn được đánh dấu là “N/a”.

6. Nhận xét và cải tiến

- **Tính chính xác**: Chương trình sử dụng phương pháp phân tích cú pháp HTML, giúp thu thập chính xác các chỉ số cần thiết. Tuy nhiên, do fbref.com thường xuyên cập nhật cấu trúc HTML, chương trình cần được bảo trì định kỳ.
- **Đề xuất cải tiến**:
    - Thêm kiểm tra HTTP response để phát hiện lỗi khi truy cập trang web.
    - Sử dụng xử lý ngoại lệ (try-except) để tiếp tục chương trình ngay cả khi có lỗi nhỏ.
    - Sử dụng danh sách DataFrames và thực hiện kết hợp một lần để tối ưu hóa tốc độ khi kết hợp các tệp CSV.

7. Kết luận

Chương trình này cung cấp phương pháp tự động hóa việc thu thập và phân tích dữ liệu cầu thủ Ngoại hạng Anh mùa 2023-2024, giúp tiết kiệm thời gian và công sức. Với những cải tiến được đề xuất, chương trình có thể hoạt động linh hoạt và hiệu quả hơn.

B. Bài Tập 2:

1. Tìm top 3 cầu thủ có điểm cao nhất và thấp nhất ở mỗi chỉ số

Mục tiêu: Xác định ba cầu thủ có điểm cao nhất và ba cầu thủ có điểm thấp nhất ở mỗi chỉ số (ví dụ: số bàn thắng, số pha kiến tạo, v.v.).

Cách giải quyết:

- Sử dụng dữ liệu từ tệp result.csv, lọc ra các cột chứa các chỉ số cầu thủ.
- Với mỗi chỉ số, sắp xếp giá trị từ cao đến thấp và chọn ba cầu thủ có điểm số cao nhất, sau đó sắp xếp từ thấp đến cao để chọn ba cầu thủ có điểm số thấp nhất.
- Kết quả bao gồm danh sách top 3 cầu thủ và bottom 3 cầu thủ cho từng chỉ số.

2. Tính trung vị, trung bình và độ lệch chuẩn của mỗi chỉ số cho toàn giải và từng đội

Mục tiêu: Tính toán các giá trị trung vị, trung bình và độ lệch chuẩn cho mỗi chỉ số, áp dụng cho toàn bộ giải đấu và từng đội.

Cách giải quyết:

- Đọc dữ liệu từ result.csv, sau đó xác định các cột chỉ số.
- Tính trung vị, trung bình và độ lệch chuẩn cho từng chỉ số của toàn bộ giải đấu.
- Sau đó, nhóm dữ liệu theo đội bóng để tính trung vị, trung bình và độ lệch chuẩn của mỗi chỉ số cho từng đội.
- Ghi kết quả vào file results2.csv theo định dạng được yêu cầu, trong đó hàng đầu tiên là các thống kê tổng thể, tiếp theo là các thống kê cho từng đội.

3. Vẽ và lưu histogram phân bố của mỗi chỉ số của các cầu thủ trong toàn giải

Mục tiêu: Tạo các biểu đồ histogram cho thấy phân bố của từng chỉ số trong toàn bộ giải đấu, để giúp hiểu rõ hơn về sự phân bố điểm số của các chỉ số này.

Cách giải quyết:

- Đọc dữ liệu từ result.csv, chọn ra các cột chỉ số để vẽ histogram.
- Đối với mỗi chỉ số, vẽ histogram để biểu diễn phân bố điểm số của cầu thủ trong giải.
- Lưu các biểu đồ histogram dưới dạng ảnh riêng biệt trong thư mục histograms, và giới hạn tối đa 20 chỉ số để tránh quá tải.

4. Tìm đội bóng có điểm cao nhất ở mỗi chỉ số và xác định đội có phong độ tốt nhất
Mục tiêu: Xác định đội bóng có điểm số cao nhất ở mỗi chỉ số và từ đó đánh giá đội nào có phong độ tốt nhất trong giải đấu.
Cách giải quyết:

- Tính trung bình của từng chỉ số cho mỗi đội bóng.
- Xác định đội có điểm trung bình cao nhất ở từng chỉ số.
- Đếm số lần mỗi đội đứng đầu ở các chỉ số để đánh giá phong độ tổng thể, đội có nhiều chỉ số đứng đầu nhất được coi là đội có phong độ tốt nhất giải đấu.

Tổng kết

Bốn bài toán trên đều xoay quanh việc phân tích các chỉ số cầu thủ và đội bóng trong giải đấu. Các phương pháp sử dụng bao gồm phân tích thống kê (trung bình, trung vị, độ lệch chuẩn), trực quan hóa dữ liệu (histogram), và phân tích thứ hạng để đánh giá phong độ. Kết quả giúp chúng ta có cái nhìn sâu sắc về hiệu suất của từng cầu thủ cũng như sức mạnh của các đội bóng trong giải đấu.

C. BÀI TẬP 3:

1. Phân cụm cầu thủ dựa trên các chỉ số
Mục tiêu: Tìm các nhóm cầu thủ có chỉ số giống nhau, từ đó phân tích và nhận dạng các nhóm cầu thủ dựa trên phong cách hoặc hiệu suất.

- **Chuẩn bị dữ liệu**:
  - Đọc dữ liệu từ tệp CSV (result.csv), chỉ chọn các cột chứa dữ liệu số (chẳng hạn như tuổi, số bàn thắng, số đường chuyền).
  - Điền các giá trị khuyết thiếu bằng trung bình của từng cột để tránh gây sai lệch trong quá trình phân tích.
- **Chuẩn hóa dữ liệu**:
  - Dữ liệu được chuẩn hóa về cùng một thang đo bằng StandardScaler, giúp thuật toán phân cụm không bị ảnh hưởng bởi sự chênh lệch lớn trong các chỉ số.
- **Xác định số cụm tối ưu**:
  - Dùng phương pháp "elbow" (khuỷu tay) để xác định số cụm tối ưu. Phương pháp này tính toán độ biến thiên (inertia) của từng số lượng cụm từ 1 đến 10 và vẽ đồ thị biểu diễn. Điểm "gãy" trên đồ thị sẽ là số cụm phù hợp nhất.
  - Ví dụ, nếu điểm "gãy" ở k = 4, chúng ta sẽ chọn 4 cụm.
- **Kết quả**:
  - Thêm nhãn cụm vào DataFrame ban đầu và tính trung bình các chỉ số cho từng cụm, giúp nhận biết đặc điểm nổi bật của từng nhóm.
  - Dữ liệu phân cụm được lưu vào file clustered_players.csv.

Trực quan hóa: Vẽ biểu đồ phân tán các cụm bằng 2 chỉ số đầu tiên (hoặc bất kỳ hai chỉ số nào). Mỗi điểm trên biểu đồ đại diện cho một cầu thủ, màu sắc biểu thị cụm của cầu thủ đó.
Nhận xét:

- Phân cụm giúp nhận diện các nhóm cầu thủ có đặc điểm nổi bật như phong cách chơi hoặc vai trò trong đội.
- Số cụm có thể thay đổi tùy thuộc vào dữ liệu và mục tiêu phân tích.

2. Phân cụm với PCA để giảm số chiều
Mục tiêu: Sử dụng K-means và kỹ thuật giảm chiều PCA để trực quan hóa dữ liệu trên mặt phẳng 2D, giúp dễ dàng phân tích đặc điểm của các cụm.

- **Lựa chọn và chuẩn bị dữ liệu**:

    - Đọc dữ liệu từ tệp results2.csv và chọn các cột quan trọng như age, nationality, team, position.
- **Chuẩn hóa dữ liệu**:
    - Dữ liệu được chuẩn hóa về cùng một thang đo, loại bỏ sự ảnh hưởng của chênh lệch giữa các chỉ số.
- **Phân cụm K-means**:
    - Chọn số cụm là 4 (theo giả định từ đoạn mã trước) và gán nhãn cụm cho từng cầu thủ.
- **Giảm số chiều bằng PCA**:
    - Dữ liệu được giảm xuống 2 chiều bằng PCA, giúp trực quan hóa phân cụm trên mặt phẳng 2D.
    - Các thành phần chính (PC1 và PC2) đại diện cho các yếu tố tổng hợp từ các chỉ số gốc.

Trực quan hóa:

- Vẽ biểu đồ phân tán các cầu thủ dựa trên hai thành phần chính (PC1 và PC2), với mỗi cụm được biểu diễn bằng một màu khác nhau.
- Phân tích đặc điểm chung của từng cụm, hỗ trợ việc xác định các kiểu cầu thủ tương tự về các chỉ số.

Nhận xét:

- PCA giúp đơn giản hóa dữ liệu và cung cấp cái nhìn trực quan về cấu trúc của các cụm.
- Phương pháp này có thể không giữ lại toàn bộ thông tin của dữ liệu gốc, nhưng vẫn hữu ích trong việc quan sát và phân tích các cụm cầu thủ.

3. Vẽ biểu đồ radar để so sánh cầu thủ

Mục tiêu: Xây dựng biểu đồ radar để so sánh các chỉ số giữa hai cầu thủ cụ thể dựa trên danh sách các chỉ số do người dùng chỉ định.

- **Nhận đầu vào từ dòng lệnh**:
    - Người dùng cung cấp tên của hai cầu thủ (--p1, --p2) và danh sách các chỉ số (--Attribute) muốn so sánh.
- **Lấy dữ liệu cầu thủ**:
    - Đọc dữ liệu từ results2.csv và trích xuất các chỉ số cần so sánh cho hai cầu thủ.
    - Nếu không tìm thấy cầu thủ hoặc chỉ số không hợp lệ, chương trình thông báo lỗi.
- **Thiết lập biểu đồ radar**:
    - Tạo các góc cho biểu đồ radar để thể hiện các chỉ số.
    - Thêm các đường nối giữa các điểm và tô màu cho vùng biểu đồ của mỗi cầu thủ để dễ so sánh.
- **Trực quan hóa**:
    - Vẽ biểu đồ radar để hiển thị rõ ràng sự khác biệt về chỉ số giữa hai cầu thủ.

Nhận xét:

- Biểu đồ radar là công cụ hiệu quả để so sánh đa chỉ số, giúp nhận diện những điểm mạnh, điểm yếu của từng cầu thủ.
- Đoạn mã này cung cấp một cách tiếp cận linh hoạt, có thể áp dụng cho nhiều chỉ số khác nhau tùy ý người dùng.

III. Kết luận

Báo cáo đã phân tích và phân cụm các cầu thủ dựa trên chỉ số thi đấu tại giải Ngoại hạng Anh mùa 2023-2024. Các kỹ thuật K-means và PCA đã được sử dụng để phân nhóm và trực quan hóa dữ liệu hiệu quả, giúp nhận biết các nhóm cầu thủ có đặc điểm tương đồng. Cuối cùng, biểu đồ radar hỗ trợ trong việc so sánh chi tiết giữa các cầu thủ, cung cấp cái nhìn sâu sắc về các chỉ số cụ thể và phong cách chơi.

Phân cụm và biểu đồ radar giúp nâng cao khả năng phân tích và hỗ trợ các nhà quản lý bóng đá trong việc đánh giá cầu thủ và xây dựng chiến thuật.

D. BÀI TẬP 4

Mục tiêu: Để định giá cầu thủ trong bóng đá, có nhiều phương pháp khác nhau có thể được áp dụng, tùy thuộc vào dữ liệu có sẵn, mục tiêu nghiên cứu, và các yếu tố mà bạn muốn xem xét. Dưới đây là một số phương pháp phổ biến mà bạn có thể xem xét:

1. Phân tích thống kê

- **Mô hình hồi quy** : Sử dụng hồi quy tuyến tính hoặc hồi quy logistic để dự đoán giá trị của cầu thủ dựa trên các yếu tố như tuổi, vị trí, thành tích thi đấu, số trận đấu, bàn thắng ghi được, và các chỉ số khác.
- **Phân tích đa biến** : Phân tích tác động của nhiều biến đồng thời lên giá trị cầu thủ, cho phép xác định mối quan hệ phức tạp giữa các yếu tố.

2. Phân tích dữ liệu cầu thủ

- **Chỉ số hiệu suất** : Sử dụng các chỉ số như bàn thắng, kiến tạo, tỷ lệ chuyền bóng chính xác, và thời gian ra sân để đánh giá hiệu suất của cầu thủ và ảnh hưởng của nó đến giá trị.
- **K-means**: Sử dụng phân nhóm K-means để phân loại cầu thủ thành các nhóm có tính chất tương tự, từ đó có thể định giá cầu thủ trong từng nhóm.

3. Mô hình định giá tương lai

- **Dự đoán giá trị chuyển nhượng tương lai** : Sử dụng dữ liệu lịch sử chuyển nhượng để dự đoán giá trị chuyển nhượng trong tương lai, dựa trên các yếu tố như xu hướng thị trường và phát triển cá nhân của cầu thủ.

4. Mô hình mạng nơ-ron

- **Deep learning** : Sử dụng các mô hình mạng nơ-ron để học từ dữ liệu lớn, bao gồm các chỉ số và thông tin về cầu thủ, giúp cải thiện độ chính xác trong việc định giá.

5. Mô hình dựa trên thị trường

- **So sánh giá trị thị trường** : Phân tích giá trị của cầu thủ tương tự trên thị trường để ước tính giá trị của cầu thủ đang xem xét. Phương pháp này thường dựa trên dữ liệu giá chuyển nhượng gần đây cho các cầu thủ có đặc điểm tương tự.

6. Phân tích định tính

- **Đánh giá từ chuyên gia** : Lấy ý kiến từ các huấn luyện viên, nhà quản lý, và các chuyên gia trong ngành để đưa ra đánh giá về cầu thủ.
- **Tình hình thị trường** : Theo dõi cảm xúc của người hâm mộ, truyền thông và các yếu tố khác có thể ảnh hưởng đến giá trị của cầu thủ.

Kết luận

Phương pháp định giá cầu thủ tốt nhất thường là sự kết hợp của nhiều phương pháp trên, kết hợp cả phân tích định lượng và định tính để đưa ra một cái nhìn toàn diện và chính xác về giá trị của cầu thủ. Bạn có thể bắt đầu bằng việc thu thập dữ liệu và áp dụng một số phương pháp trên để xem phương pháp nào hoạt động tốt nhất trong trường hợp cụ thể của bạn.