ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Phân tích và trực quan dữ liệu nhằm đề xuất xây dựng đội hình cầu thủ trong game FIFA

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Hoàng Minh	20521609
2	Nguyễn Minh Tiến	20522010
3	Tạ Nhật Minh	20521614
4	Nguyễn Thiện Thuật	20521998

TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

Hiện nay, các trò chơi điện tử có xu hướng phát triển mạnh và dần dần được nhiều người công nhận là một bộ môn thể thao. Trong đó có cả những game bóng đá và FIFA là một trò chơi điển hình. Nó là một tựa game giả lập các đội bóng giúp cho người chơi có thể xây dựng đội bóng riêng của mình từ các cầu thủ nổi tiếng.

Đề tài được thực hiện với mong muốn đề xuất giải pháp là xây dựng được mô hình máy học gợi ý cầu thủ từ kết quả của việc phân tích bộ dữ liệu.

Nhóm đã tiến hành phân tích bộ dữ liệu chỉ số các cầu thủ trong trò chơi nhằm tìm ra các chỉ số đặt trưng cho từng nhóm cầu thủ, từ đó sử dụng các mô hình máy học không giám sát (unsupervise learning) để xây dựng mô hình gợi ý cầu thủ phù hợp với mong muốn của người chơi.

Các công cụ xử lý dữ liệu mà nhóm sử dụng như Pandas, Numpy và thư viện học máy Sklearn. Ngoài ra còn có công cụ trực quan như Matplotlib, Seaborn kết hợp suy luận từ vốn hiểu biết và các kết quả tìm được từ quá trình phân tích dữ liệu để từ đó có thể sử dụng các thuật toán máy học như K-means, Hierarchical để xây dựng mô hình gợi ý.

Thông qua thực hiện đề tài, nhóm đã ứng dụng được các công cụ và tư duy xử lý dữ liệu được học từ môn học để ứng dụng vào đề tài nhằm tìm ra các chỉ số quan trọng của cầu thủ trong bộ dữ liệu và ứng dụng được nó để xây dựng được mô hình gợi ý.

2. NỘI DUNG

2.1. Dataset

Tên bộ dữ liệu: FIFA22_OFFICIAL_DATASET

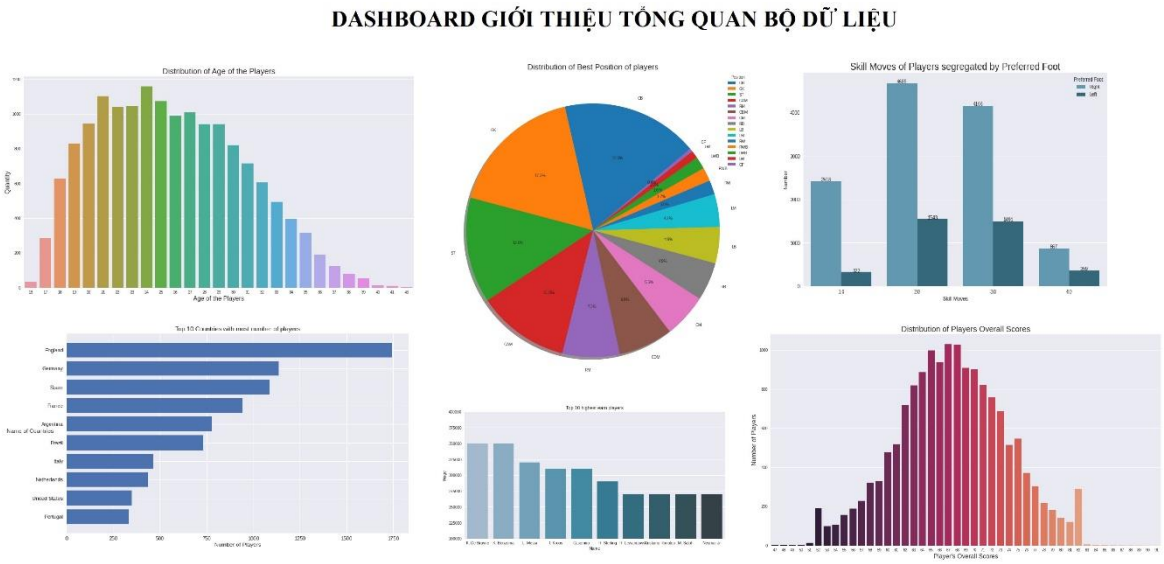
Nguồn dữ liệu: Kaggle

Cấu trúc dataset: Bộ data gồm có 16710 dòng dữ liệu và 65 cột là chi tiết về các thông tin và số liệu thống kê của các cầu thủ

Số thứ tự	Thuộc tính tương ứng trên Dataset	Nội dung
1 → 6	ID → Flag	Các thông tin cở bản của cầu thủ: ID, tên, tuổi, hình ảnh, quốc tịch, quốc kì.
7	Overall	Chỉ số trung bình của cầu thủ.
8	Potential	Tiềm năng của cầu thủ.
9 →10	Club → Club Logo	Tên và Logo câu lạc bộ cầu thủ đang thi đấu.

11 → 12	Value → Wage	Giá trị chuyển nhượng và lương của cầu thủ.
13	Special	Điểm đánh giá đóng góp trong lối chơi của đội
14 → 27	Preferred Foot → Weight	Các thông tin đặc trưng cá nhân của cầu thủ: chân thuận, độ nổi tiếng, tần suất sử dụng chân không thuận, kĩ năng di chuyển, tần suất hoạt động trên sân, thể hình, chân dung, vị trí thi đấu, số áo, ngày gia nhập câu lạc bộ, được mượn từ câu lạc bộ khác, thời hạn hợp đồng, chiều cao cân nặng.
28 → 37	Crossing → BallControl	Các số liệu về kĩ thuật cá nhân của cầu thủ: tạt cánh, dứt điểm, khả năng ghi bàng bằng đầu, chuyền bóng tầm ngắn-trung-dài, rê bóng, đọ xoáy của bóng khi sút, khả năng sút phạt, kiểm soát bóng.
38 → 40	Acceleration → Agility	Các số liệu về tốc độ của cầu thủ: khả năng tăng tốc, chạy nước rút, nhanh nhẹn.
41 → 42	Reactions → Balance	Khả năng phản ứng và giữ thăng bằng của cầu thủ.
43 → 48	ShotPower → Aggression	Các chỉ số về sức mạnh của cầu thủ: lực sút, bật nhảy, thể lực, sức mạnh, sút xa, mức độ hiếu chiến.
49 → 56	Interception → SlidingTackle	Các thông số về kĩ năng cá nhân của cầu thủ: đánh chặn, chạy chỗ, tầm nhìn chiến lược, sút penalty, sự điềm tĩnh, tổ chức phòng ngự, cướp bóng, trượt bóng.
57 → 61	GKDiving → GKReflexes	Các thông số cơ bản của một thủ môn: bay người cản phá, khả năng xử lý và giữ bóng bằng tay, phát bóng, khả năng chọn vị trí đứng, phản xạ.
62 → 65	BestPosition → DefensiveAwareness	Vị trí tốt nhất, đánh giá tổng quan, điều kiện chấm dứt hợp đồng, nhận thức phòng ngự của cầu thủ

Bảng thông tin các thuộc tính



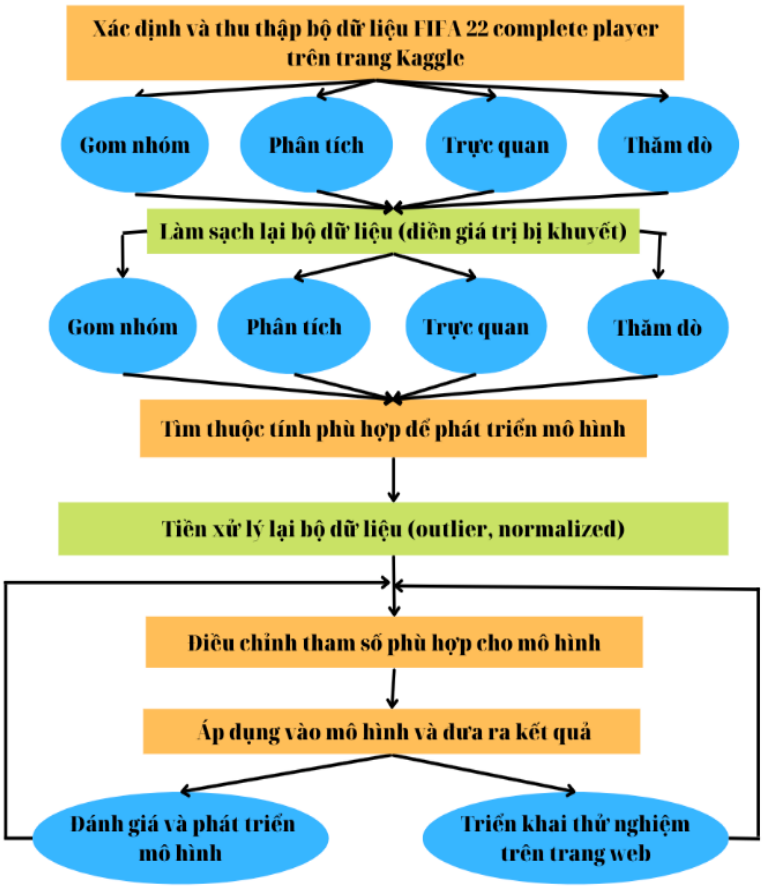
Dashboard giới thiệu tổng quan bộ dữ liệu

Một vài nhận xét về data:

- Độ tuổi thi đấu của các cầu thủ dao động từ 15 đến 45 tuổi. Tuy nhiên phần lớn nằm ở khoảng từ 18 đến 33, và có sự trẻ hóa trong giới cầu thủ khi số lượng cầu cầu thủ trẻ vượt trội hơn số lượng cầu thủ lớn tuổi.
- Giữa các nhóm cầu thủ, các cầu thủ chơi ở vị trí CB, GK, ST và CAM chiếm phần lớn. Các vị trí đá cánh có tỉ trọng thấp vì đòi hỏi khả năng di chuyển, phán đoán tình huống, và kĩ thuật cao.
- Kĩ năng di chuyển của các cầu thủ được cải thiện, đa phần ở mức 2 và 3, kĩ năng giữa các cầu thủ là khá tương đồng với nhau và các cầu thủ thuận chân phải luôn chiếm ưu thế về số lượng so với các cầu thủ còn lại.
- Chỉ số đánh giá chung của các cầu thủ phân bố đều về hai phía và chỉ số trung bình tương đối cao (68 đến 70), cho thấy chất lượng cầu thủ được nâng cao hơn so với trước và không có sự cách biệt quá lớn giữa các cầu thủ như trước.
- Đối với các quốc có lượng cầu thủ nhiều nhất, ta có thể thấy được chính sách đầu tư và phát triển bóng đá, đào tạo các tài năng. Điều này góp phần khơi dậy hứng thú cho những người có định hướng và đi theo con đường phát triển thành cầu thủ chuyên nghiệp.
- Các cầu thủ có mức lương cao có độ nổi tiếng cao được nhiều người biết đến và hâm mộ và có đánh giá chất lượng chuyên môn cao.

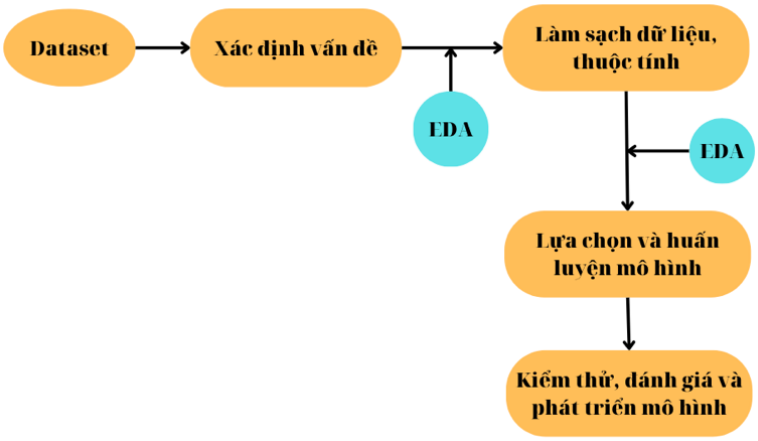
2.2. Quy trình thực hiện đề tài

Trong quá trình nghiên cứu và chuẩn bị cho bộ dữ liệu, chúng em đã quyết định thu thập dữ liệu thô từ trang Kaggle. Trang web này có chứa một lượng lớn đa dạng bộ dữ liệu phù hợp cho nhu cầu phân tích và tìm hiểu qua nhiều năm. Chúng em đã chọn FIFA 22 complete player làm bộ dữ liệu cho đề tài này. Sau khi xác định được bộ dữ liệu, nhận định được bộ dữ liệu còn bị khuyết ở khá nhiều thuộc tính, việc tiếp theo được xác định là làm sạch lại bộ dữ liệu này. Sau đó chúng em tiến hành gom nhóm, phân tích, trực quan, thăm dò từ đó đưa ra được các thuộc tính có các chỉ số phù hợp với cầu thủ để có thể phát triển mô hình, đồng thời xử lý các outlier, normalize để chuẩn bị cho bước tiếp theo. Về phần mô hình, nhóm chúng em đã xây dựng mô hình và đồng thời điều chỉnh bộ tham số để có thể phù hợp với bộ dữ liệu và cho ra kết quả tốt nhất. Cuối cùng, chúng em sẽ đưa ra đánh giá dựa trên kết quả dự đoán của mô hình và triển khai các thử nghiệm trên trang web.



Quy trình thực hiện đề tài

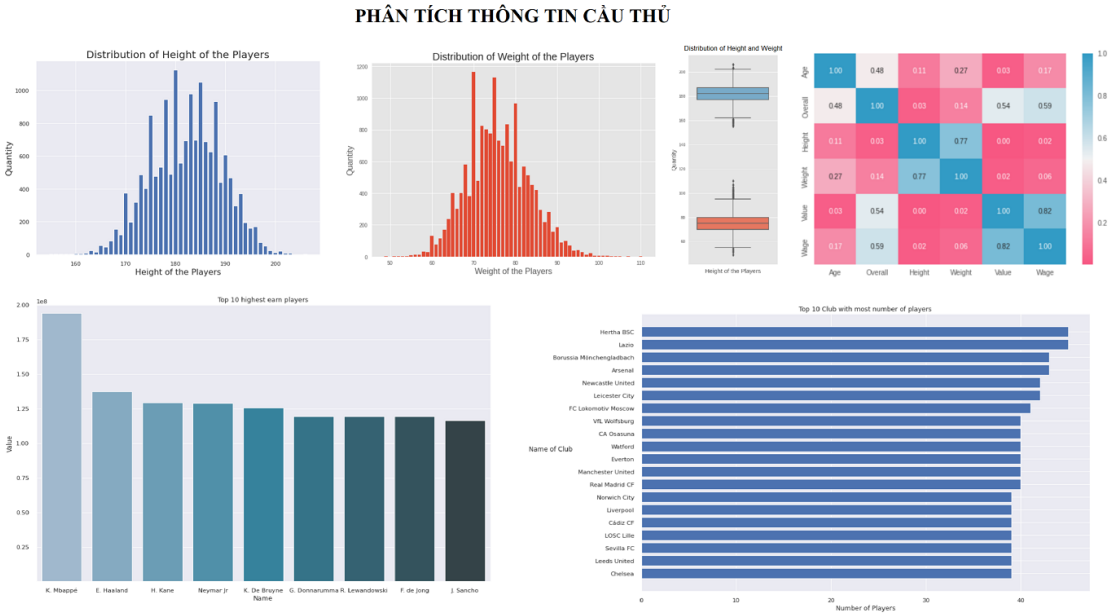
2.3. Phân tích và trực quan dữ liệu



Quy trình Phân tích dữ liệu

Từ bộ dữ liệu đã thu thập được, chúng em tiến hành thực hiện quá trình phân tích và trực quan dữ liệu thông qua một quy trình phân tích cụ thể như hình vẽ. Đầu tiên khi có dataset, tiến hành xác định vấn đề để hướng đến những mô hình huấn luyện. Dựa trên những thuộc tính đã có của bộ dataset. Sau đó tiến hành làm sạch dữ liệu và thuộc tính với nhiều phương pháp khác nhau, sau đó lựa chọn các phương pháp mang về độ hiệu quả lớn nhất cho mô hình dự đoán. Bước tiếp theo, chúng em tiến hành thực hiện nghiên cứu, lựa chọn các mô hình phù hợp cho dự án ở nhiều nguồn tài liệu khác nhau. Và cuối cùng, dựa vào những kiến thức có được, chúng em sẽ đưa mô hình dựa đoán vào những điều kiện khác nhau để kiểm thử, đánh giá và sẽ hướng phát triển mô hình trở nên tốt hơn, hoàn thiện hơn.

2.3.1. Phân tích thông tin cầu thủ



Dashboard phân tích thông tin cầu thủ

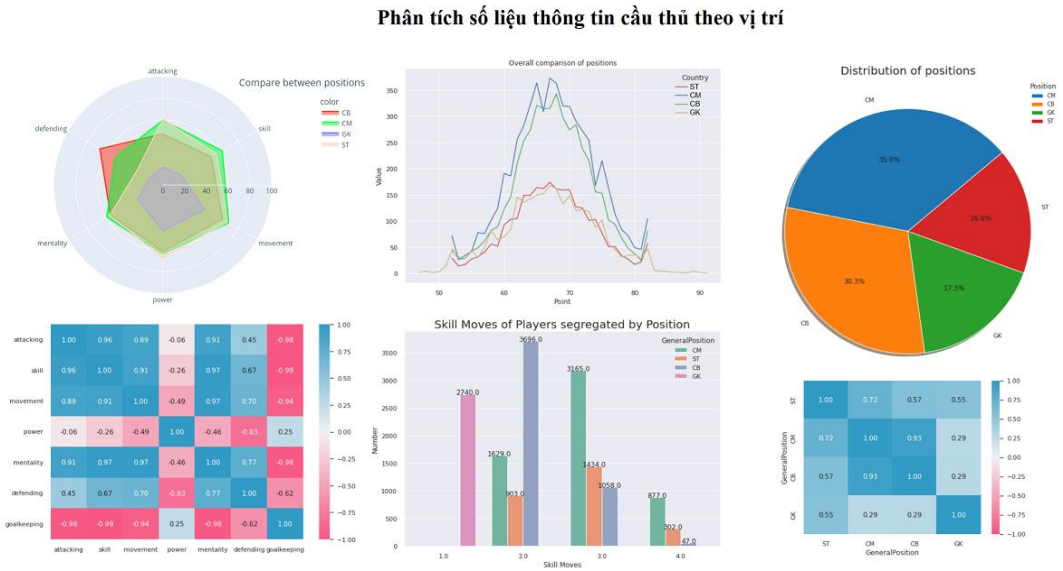
Câu hỏi: Top 10 các cầu thủ có giá trị cao nhất trong bộ dataset. Và giá trị cầu thủ liên quan đến chất lượng của cầu thủ hay không?

- Dựa vào dashboard thông tin cầu thủ ta có thể thấy được top 10 cầu thủ có giá trị cao nhất. Trong 10 cầu thủ này, K. Mbappé có giá trị cao nhất (gần 200 tỷ) và vượt trội hơn so với phần còn lại. Những cầu thủ khác trong top 10 có giá trị khá tương đương nhau, giao động trong khoảng 110 – 130 tỷ.
- Giá trị của cầu thủ có tương quan khá tốt so với chỉ số trung bình của cầu thủ. Từ đó ta có thể đưa ra dự đoán rằng các cầu thủ nằm trong top 10 cầu thủ đắt giá nhất được đánh giá là có chất lượng hàng đầu.

Câu hỏi: Top 20 câu lạc bộ sở hữu nhiều cầu thủ trong đội nhất có chứa các cầu thủ thuộc top 10 cầu thủ có giá trị cao nhất không?

- Thông qua dashoard thông tin cầu thủ ta có thể dễ dàng nhìn ra được không có cầu thủ nào trong 10 cầu thủ có giá trị cao nhất đang thi đấu cho top 20 câu lạc bộ sở hữu nhiều cầu thủ nhất trong đội. Và sau khi khảo sát các câu lạc bộ có nhiều cầu thủ đang thi đấu nhất thì ta cũng dễ dàng thấy được sức mạnh của những câu lạc bộ này chỉ nằm ở mức tầm trung của bóng đá thế giới.
- Từ những quan sát trên ta có thể chuẩn đoán được rằng giá trị của cầu thủ có ảnh hưởng đến số lượng thành viên của các câu lạc bộ. Từ đó có thể ảnh hưởng đến quá trình xây dựng một đội bóng, khi muốn có nhiều cầu thủ có giá trị cao thì sẽ khó khăn hơn trong việc xây dựng đến một đội bóng cùng chung một câu lạc bộ.

2.3.2. Phân tích các chỉ số cầu thủ



Dashboard phân tích các chỉ số cầu thủ

Câu hỏi: Những cầu thủ chơi ở vị trí nào có những điểm tương đồng với nhau?

- Thông qua dashboard ta có thể thấy được ST và GK cùng với CM và CB có những điểm tương đồng với nhau. Từ đó có khả năng thay đổi vị trí của các cầu thủ ở vị trí đó nếu như một vị trí được đề xuất bị thiếu hoặc dư.
- Bên cạnh đó ta còn rút ra được những cầu thủ có vị trí ở trên sân gần nhau không hẳn có những điểm tương đồng với nhau.

Câu hỏi: Hãy cho biết những vị trí được chơi nhiều và mức độ cạnh tranh của vị trí đó.

- Qua dashboard ta có thể thấy được hai vị trí có nhiều cầu thủ nhất là CB và CM, cùng với các chỉ số về chuyên môn của cầu thủ ở 2 vị trí này cao hơn so với 2 vị trí còn lại. Điều này đồng nghĩa với việc khả năng cạnh tranh của 2 vị trí này lớn hơn so với các vị trí khác. Điều này có thể giúp cho việc lựa chọn các cầu thủ dễ dàng hơn trong việc tạo ra một đội hình khi có nhiều sự lựa chọn cho cùng một vị trí.

Câu hỏi: Những thông số nào có khả năng ảnh hưởng với nhau nhất và cho kết luận.

- Qua dashboard ta có thể thấy những thông số ảnh hưởng với nhau nhất là skill với attaking (0.96), skill với movement (0.91) hay attacking với (-0.98). Với những thông số như trên đã phản ảnh được phần nào sự ảnh hưởng của các thông số với nhau. Khi cần xây dựng một một hình theo một thông số có thể ảnh hưởng đến các thông số khác và kết với với chúng để tạo nên đội hình không ảnh hưởng đến ý định ban đầu.
- Ví dụ: một người chơi muốn chơi với đội hình có attacking cao nhưng không muốn chỉ số skill cao thì sẽ khó khăn hơn trong việc chọn được hình ưng ý.

2.3.3. Kết quả các khám phá thu được từ phân tích từ bộ dữ liệu

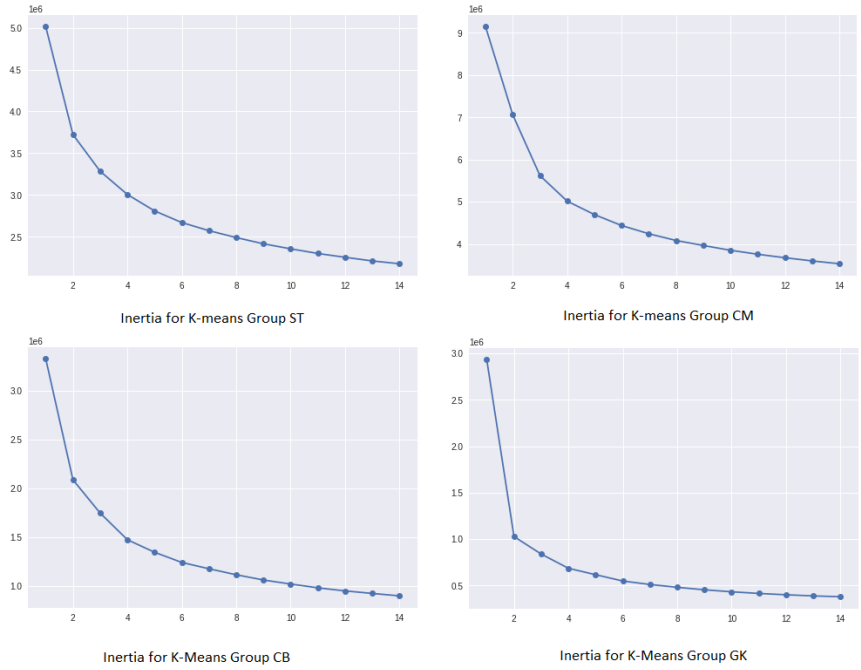
Sau khi phân tích dữ liệu ta có thể nhìn nhận ra được nhiều vấn đề hơn để góp phần phát triển và xây dựng mô hình. Có những vị trí ảnh hưởng đến nhau dẫn đến việc không thể thay thế từ vị trí này thành vị trí khác. Nhưng cũng có những vị trí không ảnh hưởng đến nhau nên hoàn toàn có thể thay thế thành các vị trí khác. Bên cạnh đó những thông tin cá nhân cũng có độ ảnh hưởng nhất định vào sức mạnh, khả chơi bóng của các cầu thủ. Điển hình là giá trị và độ tuổi của các cầu thủ sẽ là thuộc tính đáng lưu ý khi nó hoàn toàn có thể ảnh hưởng đến việc xây dựng đội hình. Còn về phía thông số dữ liệu của các cầu thủ cần phải xem xét những cặp giá trị có mối quan hệ tương quan với nhau để vì có thể ảnh hưởng đến nhu cầu xây dựng đội hình của người dùng và khả năng thay thế những cầu thủ có chỉ số khác cao hơn và sẽ phù hợp hơn.

2.4. Xây dựng mô hình máy học

2.4.1. Mô hình thuật toán K-means

Thuật toán phân cụm k-means là một phương pháp được sử dụng để phân vùng dữ liệu thành k-cụm khác nhau. Mỗi cụm sẽ là các điểm dữ liệu có sự tương đồng vì chúng gần nhau trong không gian. Giải thuật này còn giúp chúng ta xác định được dữ liệu mới đang thuộc về nhóm nào. Chính vì vậy, thuật toán này được nhóm sử dụng để tìm các cầu thủ có điểm tương đồng với các chỉ số mà người chơi mong muốn.

Lựa chọn tham số



Để lựa chọn số cụm phù hợp cho từng nhóm cầu thủ, chúng em sử dụng Inertia *(trục x là số nhóm, trục y là lỗi)*. Số cụm mà từ đó về sau lỗi có xu hướng giảm chậm thì sẽ là tốt nhất cho việc phân chia:

- Với nhóm ST, số lượng phân chia là 4 nhóm nhỏ hơn
- Với nhóm CM, số lượng phân chia là 4 nhóm nhỏ hơn
- Với nhóm CB, số lượng phân chia là 3 nhóm nhỏ hơn
- Với nhóm GK, số lượng phân chia là 3 nhóm nhỏ hơn

Kết quả dự đoán

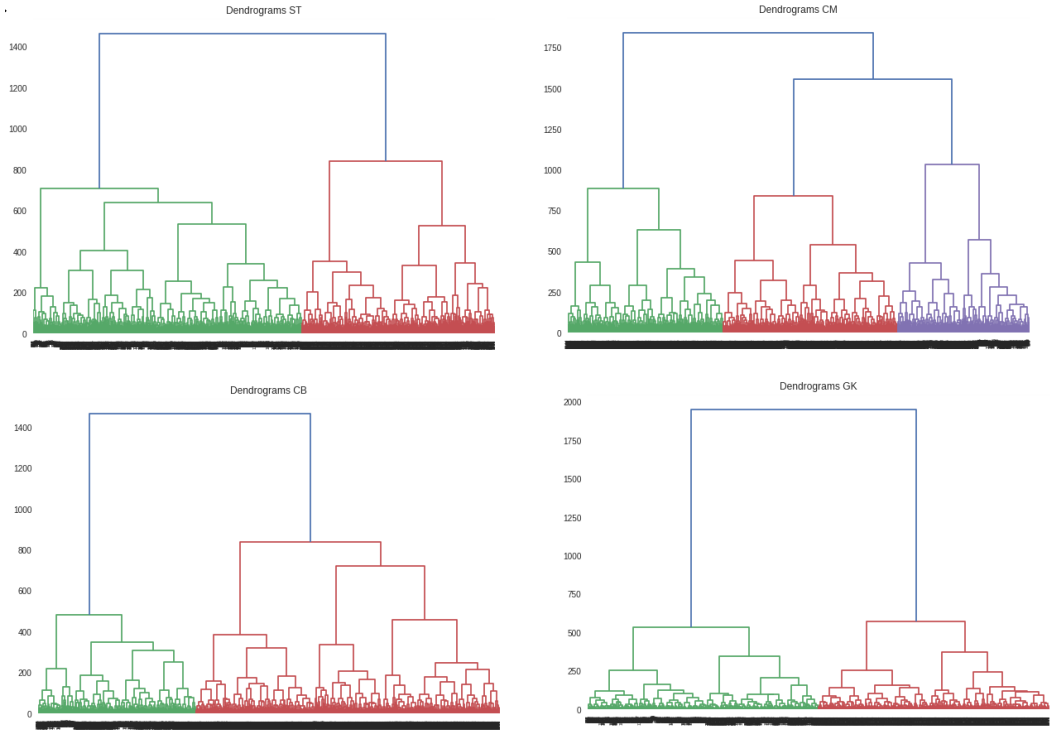
	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Special	...	GKKicking	GKPositioning	GKReflexes	Best Position	Best Overall Rating	Release Clause	DefensiveAwareness	GeneralPosition
1301	208153	L Veldwijk	29	South Africa	70	70	Suwon FC	€1.6M	€4K	1682	...	15.0	15.0	16.0	ST	70.0	€2M	41.0	ST
2483	230940	C Gondo	24	Côte d'Ivoire	59	66	US Salernitana 1919	€450K	€6K	1411	...	13.0	9.0	7.0	ST	61.0	€900K	22.0	ST
1350	184789	A Szalai	33	Hungary	71	71	1. FSV Mainz 05	€1.2M	€17K	1674	...	6.0	15.0	9.0	ST	71.0	€2.2M	39.0	ST

Kết quả gợi ý cầu thủ ST từ mô hình K-means (khi không có tham số đầu vào thì mô hình sẽ gợi ý các cầu thủ tốt nhất trong nhóm)

2.4.2. Mô hình thuật toán Hierarchical + SVM

Hierarchical là một thuật toán phân cụm tương tự như k-means. Nó tạo ra các cụm lớn hơn bằng các sáp nhập các cụm nhỏ hơn gần nhau nhất tại mỗi vòng lặp, nhờ vậy ta có thể tính khoảng cách của 2 nhóm sau mỗi lần lặp để từ đó chọn được số nhóm phân chia mà ta cho rằng khoảng cách giữa các nhóm là phù hợp. Thuật toán này chỉ phân nhóm mà không giúp ta dự đoán được điểm dữ liệu mới thuộc về nhóm nào của các cụm từ Hierachical, chính vì vậy, nhóm đã sử dụng SVM để giải quyết vấn đề này để từ đó có thể gợi ý các cầu thủ.

Lựa chọn tham số



Nhóm đã sử dụng biểu đồ Dendrograms để chọn cụm phù hợp cho thuật toán Hierachical. Chúng em nhận thấy < 600 là có thể chấp nhận được, kết quả như sau:

Với nhóm ST, số lượng phân chia là 5 nhóm

Với nhóm CM, số lượng phân chia là 6 nhóm

Với nhóm CB, số lượng phân chia là 4 nhóm

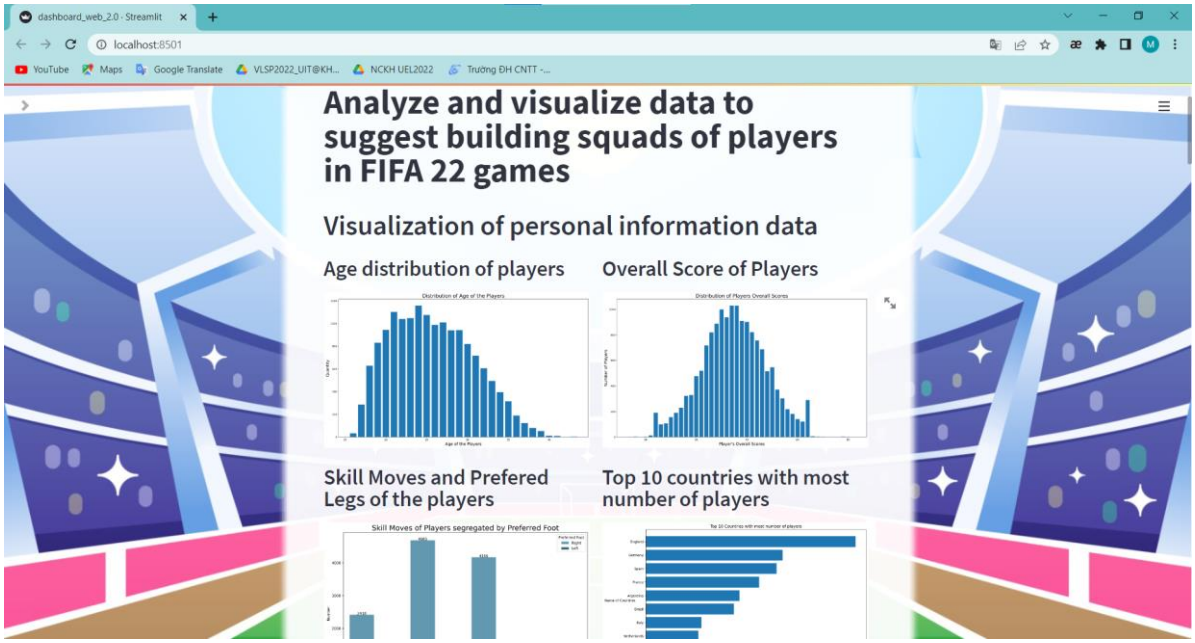
Với nhóm GK, số lượng phân chia là 2 nhóm

Kết quả dự đoán

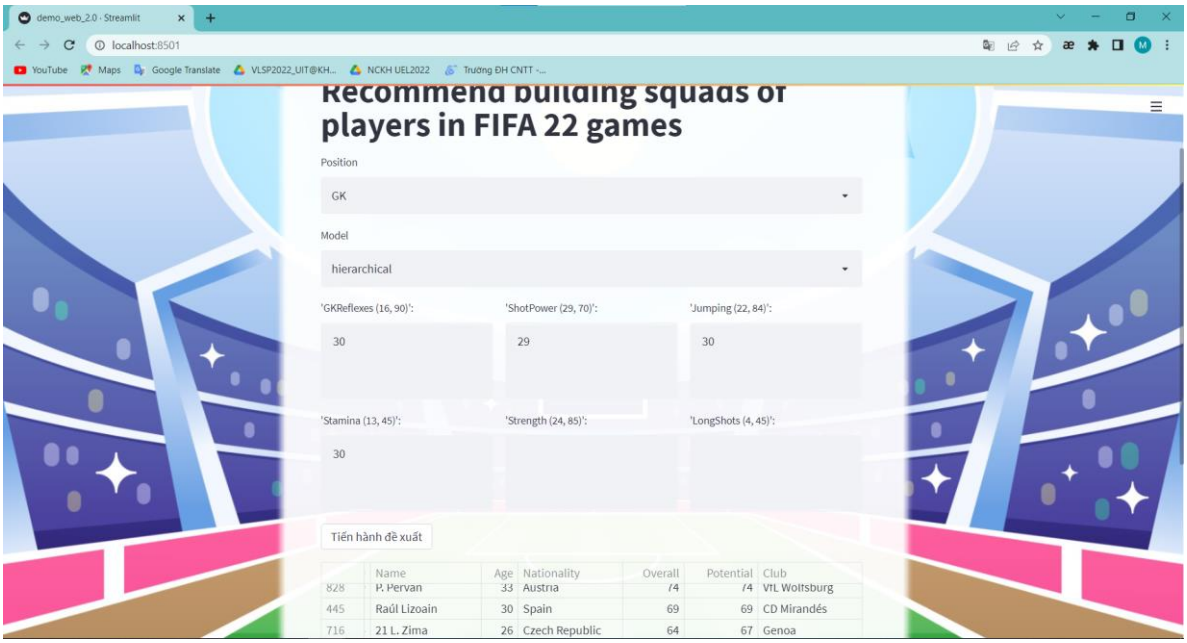
	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Special	...	GKKicking	GKPositioning	GKReflexes	Best Position	Best Overall Rating	Release Clause	DefensiveAwareness	GeneralPosition
35	204838	R Jiménez	30	Mexico	82	83	Wolverhampton Wanderers	€35.5M	€120K	2091	...	13.0	14.0	10.0	ST	83.0	€67.5M	45.0	ST
124	178311	21 Éder	33	Italy	79	79	Jiangsu FC	€11M	€29K	1975	...	11.0	10.0	9.0	ST	80.0	€17.1M	31.0	ST
14	224458	Diogo Jota	24	Portugal	82	86	Liverpool	€46M	€120K	2131	...	15.0	9.0	11.0	ST	84.0	€88.6M	54.0	ST

Kết quả gợi ý cầu thủ ST từ mô hình Hierarchical + SVM (khi không có tham số đầu vào thì mô hình sẽ gợi ý các cầu thủ tốt nhất trong nhóm)

2.5. Triển khai mô hình



Hình ảnh giao diện web dashboard



Hình ảnh giao diện web demo mô hình gợi ý cầu thủ

3. KẾT LUẬN

Sau các bước phân tích và trực quan bộ dữ liệu đã được trình bày ở trên. Từ 65 thuộc tính sau khi thăm dò và trích chọn đặc trưng, ta chia được làm 4 cụm tương ứng với 4 nhóm cầu thủ gồm gồm ST, CM, CB, GK và mỗi cụm sẽ có các đặc trưng, số lượng khác nhau.

Từ các kết quả ấy, Nhóm đã xây dựng được các model từ thuật toán K-means và Hierarchical kết hợp với SVM để recommend ra danh sách cầu thủ với chỉ số phù hợp với nhu cầu của người dùng.

TÀI LIỆU THAM KHẢO

[1] https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset (10/12/2022)

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (15/12/2022)

[3] https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html (15/12/2022)

[4] https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8 (18/12/2022)

[5] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (16/12/2022)

[6] https://docs.streamlit.io/ (19/12/2022)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Hoàng Minh	- Xây dựng phương pháp - Lên cấu trúc của đồ án - Xây dựng model gợi ý, viết báo cáo - Đánh giá báo cáo
2	Nguyễn Minh Tiến	- Hỗ trợ làm báo cáo - Làm slide, thuyết trình
3	Tạ Nhật Minh	- Phân tích thăm dò dữ liệu, viết báo cáo - Tiền xử lý dữ liệu - Hỗ trợ phân tích khai phá dữ liệu
4	Nguyễn Thiện Thuật	- Phân tích khai phá dữ liệu, viết báo cáo - Đóng gói model (viết function) - Xây dựng web dashboard, web demo