BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

NGUYỄN NGỌC THIÊN ÂN - 2051012007
ĐÀO NGUYÊN BẢO - 2051050025

PHÂN TÍCH DỮ LIỆU BÓNG ĐÁ QUỐC TẾ - TẬP TRUNG VÀO ĐỘI TUYỂN VIỆT NAM

BÀI TẬP LỚN
MÔN KHAI PHÁ DỮ LIỆU - ITEC3417

TP. HỒ CHÍ MINH, NĂM 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

PHÂN TÍCH DỮ LIỆU BÓNG ĐÁ QUỐC TẾ - TẬP TRUNG VÀO ĐỘI TUYỂN VIỆT NAM

NGUYỄN NGỌC THIÊN ÂN - 2051012007
ĐÀO NGUYÊN BẢO - 2051050025

BÀI TẬP LỚN
MÔN KHAI PHÁ DỮ LIỆU - ITEC3417

Giảng viên hướng dẫn: ThS. NGUYỄN VĂN BẢY

TP. HỒ CHÍ MINH, NĂM 2025

LỜI CẢM ƠN

Trước tiên, em xin chân thành cảm ơn thầy ThS. NGUYỄN VĂN BẢY đã tận tình hướng dẫn, chỉ bảo và tạo điều kiện thuận lợi cho em trong suốt quá trình thực hiện đồ án tốt nghiệp. Thầy đã giúp em định hướng nghiên cứu về phân tích dữ liệu bóng đá, cung cấp những kiến thức quý báu về các thuật toán khai phá dữ liệu như K-Means clustering, Decision Tree, Random Forest và Logistic Regression. Thầy cũng đã hướng dẫn em cách tiếp cận bài toán phân tích phong độ đội tuyển Việt Nam một cách khoa học và có hệ thống.

Em cũng xin cảm ơn các bạn sinh viên trong lớp đã cùng nhau học tập, trao đổi kiến thức và hỗ trợ lẫn nhau trong quá trình thực hiện đồ án. Giúp em có thêm nhiều ý tưởng và góc nhìn mới về vấn đề nghiên cứu.

Cuối cùng, em xin cảm ơn các nguồn dữ liệu mở đã cung cấp bộ dữ liệu International Football Results quý giá, giúp em có thể thực hiện nghiên cứu một cách đầy đủ và chính xác. Bộ dữ liệu này với hơn 48,000 trận đấu quốc tế đã tạo nền tảng vững chắc cho việc phân tích phong độ đội tuyển Việt Nam qua các giai đoạn khác nhau.

Thành phố Hồ Chí Minh, tháng 9 năm 2025

Sinh viên thực hiện

Nguyễn Ngọc Thiên Ân và Đào Nguyên Bảo

NHẬN XÉT CỦA GIẢNG VIÊN

TÓM TẮT BÀI TẬP LỚN

Bài tập lớn này tập trung vào việc phân tích dữ liệu bóng đá quốc tế, đặc biệt là đội tuyển Việt Nam từ năm 2000 đến nay. Sử dụng bộ dữ liệu International Football Results với 48,366 trận đấu, nghiên cứu áp dụng các kỹ thuật khai phá dữ liệu để phân tích phong độ đội tuyển Việt Nam qua các giai đoạn khác nhau.

Nghiên cứu thực hiện phân cụm (K-Means) để nhóm các giai đoạn huấn luyện viên và đối thủ dựa trên các chỉ số hiệu suất như tỷ lệ thắng, bàn thắng trung bình, và hiệu số bàn thắng. Đồng thời, áp dụng các mô hình phân loại (Decision Tree, Random Forest, Logistic Regression) để dự đoán kết quả trận đấu với độ chính xác lên đến 49.3%.

Kết quả cho thấy giai đoạn HLV Park Hang-seo (2017-2022) có hiệu suất tốt nhất với tỷ lệ thắng 46.3%, trong khi các đối thủ được phân nhóm thành 3 cluster: "Khắc tinh" (28.6%), "Ngang cơ" (54.3%), và "Yếu hơn" (17.1%). Nghiên cứu cung cấp insights có giá trị cho việc đánh giá phong độ đội tuyển và xây dựng chiến thuật phù hợp.

ABSTRACT

This thesis focuses on analyzing international football data, particularly the Vietnam national team from 2000 to present. Using the International Football Results dataset with 48,366 matches, the study applies data mining techniques to analyze Vietnam's performance across different periods.

The research implements clustering (K-Means) to group coaching periods and opponents based on performance metrics such as win rate, average goals scored, and goal difference. Additionally, classification models (Decision Tree, Random Forest, Logistic Regression) are applied to predict match outcomes with accuracy up to 49.3%.

Results show that the Park Hang-seo coaching period (2017-2022) achieved the best performance with a 46.3% win rate, while opponents were grouped into 3 clusters: "Strong" (28.6%), "Equal" (54.3%), and "Weaker" (17.1%). The study provides valuable insights for evaluating team performance and developing appropriate strategies.

MỤC LỤC

DANH MỤC VIẾT TẮT

AI	Artificial Intelligence (Trí tuệ nhân tạo)
API	Application Programming Interface (Giao diện lập trình ứng dụng)
CPU	Central Processing Unit (Bộ xử lý trung tâm)
DT	Decision Tree (Cây quyết định)
FIFA	Fédération Internationale de Football Association
GPU	Graphics Processing Unit (Bộ xử lý đồ họa)
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
K-Means	K-Means Clustering Algorithm (Thuật toán phân cụm K-Means)
LR	Logistic Regression (Hồi quy logistic)
ML	Machine Learning (Học máy)
RAM	Random Access Memory (Bộ nhớ truy cập ngẫu nhiên)
RF	Random Forest (Rừng ngẫu nhiên)
SQL	Structured Query Language (Ngôn ngữ truy vấn có cấu trúc)
URL	Uniform Resource Locator (Định vị tài nguyên thống nhất)
XML	eXtensible Markup Language (Ngôn ngữ đánh dấu mở rộng)

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

MỞ ĐẦU

0.1 Bối cảnh và lý do chọn đề tài

Bóng đá là môn thể thao phổ biến nhất thế giới và có tầm quan trọng đặc biệt đối với Việt Nam. Việc phân tích dữ liệu bóng đá quốc tế, đặc biệt tập trung vào đội tuyển Việt Nam, có thể giúp hiểu rõ phong độ của đội tuyển qua các giai đoạn khác nhau, phân tích đối thủ để xây dựng chiến thuật phù hợp, dự đoán kết quả trận đấu dựa trên các yếu tố lịch sử, và đánh giá hiệu quả của các giai đoạn huấn luyện viên.

0.2 Mục tiêu nghiên cứu

Nghiên cứu này nhằm thực hiện các mục tiêu sau:

1. Phân tích phong độ đội tuyển Việt Nam (2000–nay) thông qua thống kê số trận, bàn thắng/bàn thua, tỉ lệ thắng–hòa–thua theo năm và so sánh trước/sau thời kỳ HLV Park Hang-seo (2017–2022).

2. Thực hiện phân cụm (Clustering) theo giai đoạn HLV để tạo feature theo từng giai đoạn HLV: trung bình bàn thắng, bàn thua, win rate và gom cụm để thấy sự khác biệt phong độ giữa các giai đoạn.

3. Phân cụm đối thủ của Việt Nam với feature: số trận gặp, tỉ lệ thắng, bàn thắng/bàn thua trung bình để gom cụm và phân nhóm đối thủ: "khắc tinh", "ngang cơ", "yếu hơn".

4. Áp dụng Classification để dự đoán kết quả trận (Win/Draw/Lose) sử dụng Decision Tree, Random Forest, Logistic Regression và đánh giá bằng Accuracy, Precision, Recall, F1-score.

0.3 Phạm vi nghiên cứu

Nghiên cứu tập trung vào phân tích dữ liệu bóng đá quốc tế từ năm 2000 đến nay, đặc biệt là các trận đấu có sự tham gia của đội tuyển Việt Nam. Dữ liệu được lấy từ bộ International Football Results với 48,366 trận đấu, 44,447 bàn thắng và 650 trận penalty.

0.4 Phương pháp tiếp cận

Nghiên cứu sử dụng các kỹ thuật khai phá dữ liệu bao gồm:

- Tiền xử lý dữ liệu: Làm sạch dữ liệu, xử lý missing values, tạo features mới
- Phân cụm: K-Means clustering để nhóm giai đoạn HLV và đối thủ
- Phân loại: Decision Tree, Random Forest, Logistic Regression để dự đoán kết quả trận đấu
- Đánh giá: Sử dụng các metrics như Accuracy, Precision, Recall, F1-score, Confusion Matrix

0.5 Cấu trúc đồ án

Đồ án được cấu trúc thành 5 chương chính:

- Chương 1: Tổng quan về phân tích dữ liệu bóng đá
- Chương 2: Tiền xử lý dữ liệu
- Chương 3: Áp dụng các mô hình khai phá dữ liệu

- Chương 4: Kết quả và phân tích
- Chương 5: Kết luận và hướng phát triển

Chương 1

TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU BÓNG ĐÁ

1.1 Giới thiệu bài toán

Bóng đá là môn thể thao phổ biến nhất thế giới và có tầm quan trọng đặc biệt đối với Việt Nam. Việc phân tích dữ liệu bóng đá quốc tế, đặc biệt tập trung vào đội tuyển Việt Nam, có thể giúp:

- **Hiểu rõ phong độ** của đội tuyển qua các giai đoạn khác nhau
- **Phân tích đối thủ** để xây dựng chiến thuật phù hợp
- **Dự đoán kết quả** trận đấu dựa trên các yếu tố lịch sử
- **Đánh giá hiệu quả** của các giai đoạn huấn luyện viên

1.2 Mục tiêu dự án

Nghiên cứu này nhằm thực hiện các mục tiêu cụ thể sau:

1. **Phân tích phong độ đội tuyển Việt Nam (2000–nay)**
    - Thống kê số trận, bàn thắng/bàn thua, tỉ lệ thắng–hòa–thua theo năm

- So sánh trước/sau thời kỳ HLV Park Hang-seo (2017–2022)

2. **Phân cụm (Clustering) theo giai đoạn HLV**
   - Tạo feature theo từng giai đoạn HLV: trung bình bàn thắng, bàn thua, win rate
   - Gom cụm để thấy sự khác biệt phong độ giữa các giai đoạn
3. **Phân cụm đối thủ của Việt Nam**
   - Feature: số trận gặp, tỉ lệ thắng, bàn thắng/bàn thua trung bình
   - Gom cụm để phân nhóm đối thủ: "khắc tinh", "ngang cơ", "yếu hơn"
4. **Classification – Dự đoán kết quả trận (Win/Draw/Lose)**
   - Sử dụng Decision Tree, Random Forest, Logistic Regression
   - Đánh giá bằng Accuracy, Precision, Recall, F1-score

1.3 Mô tả bộ dữ liệu

1.3.1 Nguồn gốc dữ liệu

- **Dataset**: International Football Results from 1872 to 2024
- **Nguồn**: `https://github.com/martj42/international_results`
- **Thời gian**: Từ năm 1872 đến 2024 (lọc từ 2000 trở đi)

1.3.2 Cấu trúc dữ liệu

results.csv (48,366 trận đấu)

- `date`: Ngày thi đấu
- `home_team`: Đội chủ nhà

- `away_team`: Đội khách
- `home_score`: Số bàn thắng đội chủ nhà
- `away_score`: Số bàn thắng đội khách
- `tournament`: Tên giải đấu
- `city`: Thành phố tổ chức
- `country`: Quốc gia tổ chức
- `neutral`: Sân trung lập (True/False)

goalscorers.csv (44,447 bàn thắng)

- `date`, `home_team`, `away_team`: Thông tin trận đấu
- `team`: Đội ghi bàn
- `scorer`: Tên cầu thủ ghi bàn
- `own_goal`: Bàn phản lưới nhà (True/False)
- `penalty`: Bàn phạt đền (True/False)

shootouts.csv (650 trận penalty)

- `date`, `home_team`, `away_team`: Thông tin trận đấu
- `winner`: Đội thắng penalty
- `first_shooter`: Đội đá phạt đầu tiên

1.3.3 Ý nghĩa các thuộc tính

- **Kết quả trận đấu**: Phản ánh phong độ và sức mạnh tương đối
- **Giải đấu**: Mức độ quan trọng và cạnh tranh

- **Venue**: Lợi thế sân nhà/khách
- **Thời gian**: Xu hướng phát triển theo năm

1.4 Cơ sở lý thuyết

1.4.1 K-Means Clustering

K-Means là thuật toán phân cụm không giám sát, hoạt động theo nguyên lý:

1. Chọn k điểm trung tâm ban đầu (centroids)
2. Gán mỗi điểm dữ liệu vào cluster gần nhất
3. Cập nhật vị trí centroids dựa trên trung bình của các điểm trong cluster
4. Lặp lại cho đến khi hội tụ

Lý do lựa chọn: Phù hợp để phân nhóm các giai đoạn HLV dựa trên các chỉ số hiệu suất tương tự nhau.

1.4.2 Decision Tree

Nguyên lý: Xây dựng cây quyết định bằng cách chia dữ liệu theo các thuộc tính có entropy thấp nhất.

Ưu điểm: Dễ hiểu, không cần chuẩn hóa dữ liệu, xử lý được cả numerical và categorical.

Lý do lựa chọn: Có thể diễn giải được quy tắc dự đoán.

1.4.3 Random Forest

Nguyên lý: Kết hợp nhiều Decision Tree, mỗi tree được train trên subset khác nhau của dữ liệu.

Ưu điểm: Giảm overfitting, tăng độ chính xác, xử lý được missing values.

Lý do lựa chọn: Cải thiện hiệu suất so với Decision Tree đơn lẻ.

1.4.4 Logistic Regression

Nguyên lý: Sử dụng hàm logistic để dự đoán xác suất thuộc về mỗi class.

Ưu điểm: Nhanh, ổn định, cho xác suất dự đoán.

Lý do lựa chọn: Baseline model tốt để so sánh.

Chương 2

TIỀN XỬ LÝ DỮ LIỆU

2.1 Import thư viện và tải dữ liệu

Nghiên cứu sử dụng các thư viện Python chính sau:

- **pandas, numpy**: Xử lý và phân tích dữ liệu
- **matplotlib, seaborn**: Trực quan hóa dữ liệu
- **scikit-learn**: Machine Learning (K-Means, Decision Tree, Random Forest, Logistic Regression)
- **datetime**: Xử lý dữ liệu thời gian

2.2 Làm sạch dữ liệu (Data Cleaning)

2.2.1 Xử lý Missing Values và Outliers

Sau khi kiểm tra, bộ dữ liệu không có missing values. Tuy nhiên, có một số trận đấu có điểm số bất thường (>10 bàn), chủ yếu là các trận đấu từ thế kỷ 19-20. Để đảm bảo tính đại diện, nghiên cứu lọc dữ liệu từ năm 2000 trở đi.

2.2.2 Lọc dữ liệu Việt Nam và tạo features

Lý do lựa chọn kỹ thuật:

- Lọc từ năm 2000: Dữ liệu gần đây hơn, phù hợp với mục tiêu phân tích
- Chuẩn hóa tên đội: Đảm bảo tính nhất quán trong phân tích
- Tạo features mới: Cần thiết cho các thuật toán machine learning

Kết quả lọc dữ liệu

Sau khi lọc từ năm 2000:

- **Results**: 24,310 trận đấu
- **Goalscorers**: 25,245 bàn thắng
- **Shootouts**: 388 trận penalty
- **Số trận đấu có Việt Nam tham gia**: 230 trận (từ năm 2000 đến 2025)

Các features mới được tạo

- `result`: Kết quả trận đấu (win/draw/lose) từ góc độ Việt Nam
- `opponent`: Đối thủ của Việt Nam
- `venue`: Sân đấu (home/away) từ góc độ Việt Nam
- `goals_scored`: Bàn thắng Việt Nam ghi được
- `goals_conceded`: Bàn thua Việt Nam để thủng lưới
- `goal_diff`: Hiệu số bàn thắng
- `year`: Năm thi đấu
- `coach_period`: Giai đoạn huấn luyện viên

Thống kê cơ bản

- **Tổng số trận**: 230
- **Tỷ lệ thắng**: 43.5%
- **Tỷ lệ hòa**: 22.2%
- **Tỷ lệ thua**: 34.3%
- **Bàn thắng/trận**: 1.70
- **Bàn thua/trận**: 1.30
- **Hiệu số bàn thắng**: 0.40

2.3 Phân chia theo giai đoạn HLV

Dữ liệu được phân chia thành 3 giai đoạn chính:

1. **Trước Park Hang-seo (2000-2016)**: 141 trận
2. **Park Hang-seo (2017-2022)**: 54 trận
3. **Sau Park Hang-seo (2023-nay)**: 35 trận

2.4 Chuẩn bị dữ liệu cho Machine Learning

2.4.1 Features cho Clustering

Đối với clustering giai đoạn HLV:

- `win_rate`: Tỷ lệ thắng
- `avg_goals_scored`: Bàn thắng trung bình
- `avg_goals_conceded`: Bàn thua trung bình

- `avg_goal_diff`: Hiệu số bàn thắng trung bình

Đối với clustering đối thủ:

- `total_matches`: Số trận gặp nhau
- `win_rate`: Tỷ lệ thắng của Việt Nam
- `avg_goals_scored`: Bàn thắng trung bình của Việt Nam
- `avg_goals_conceded`: Bàn thua trung bình của Việt Nam
- `avg_goal_diff`: Hiệu số bàn thắng trung bình

2.4.2 Features cho Classification

Dữ liệu được encode và chuẩn hóa cho các mô hình classification:

- `opponent_encoded`: Đối thủ (encoded)
- `venue_encoded`: Sân đấu (encoded)
- `tournament_encoded`: Giải đấu (encoded)
- `year`: Năm thi đấu
- `coach_encoded`: Giai đoạn HLV (encoded)

Dữ liệu được chia thành train set (70%) và test set (30%) với stratification để đảm bảo phân bố đều các class.

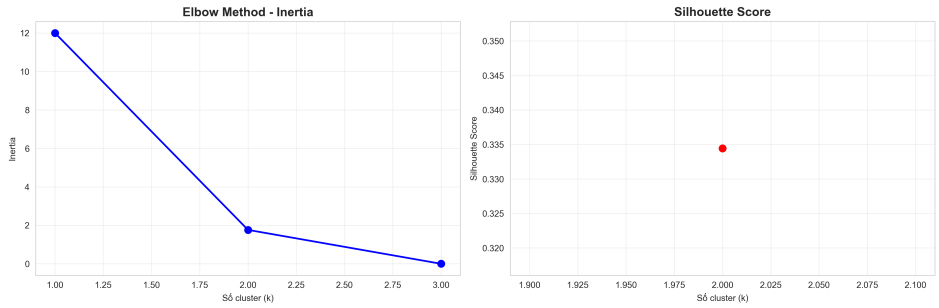
Chương 3

ÁP DỤNG CÁC MÔ HÌNH KHAI PHÁ DỮ LIỆU

3.1 Phân cụm theo giai đoạn HLV (K-Means Clustering)

3.1.1 Elbow Method để xác định số cluster tối ưu

Để xác định số cluster tối ưu cho giai đoạn HLV, nghiên cứu sử dụng Elbow Method kết hợp với Silhouette Score.



Hình 3.1: Elbow Method và Silhouette Score cho clustering giai đoạn HLV

Kết quả cho thấy số cluster tối ưu là k=2 với Silhouette Score cao nhất là 0.052.

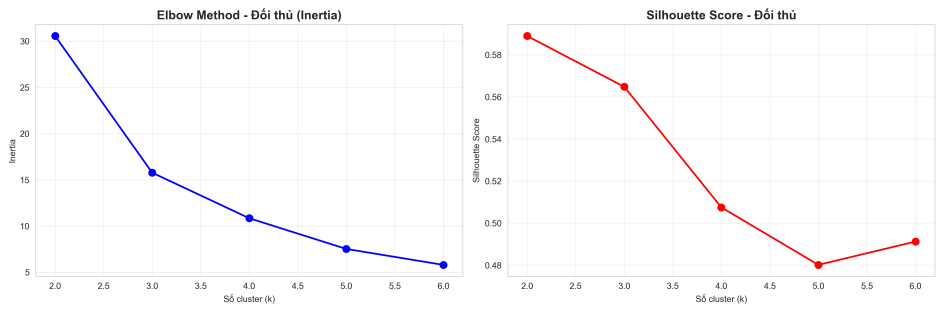
3.1.2 Kết quả clustering giai đoạn HLV

Bảng 3.1: Kết quả clustering giai đoạn HLV

Giai đoạn HLV	Số trận	Tỷ lệ thắng	Bàn thắng TB	Bàn thua TB	Cluster
Park Hang-seo (2017-2022)	54	0.463	1.574	0.833	Trung bình
Sau Park Hang-seo (2023-nay)	35	0.457	1.486	1.429	Trung bình
Trước Park Hang-seo (2000-2016)	141	0.418	1.801	1.447	Trung bình

3.2 Phân cụm đối thủ của Việt Nam

3.2.1 Elbow Method cho clustering đối thủ



Hình 3.2: Elbow Method và Silhouette Score cho clustering đối thủ

Kết quả cho thấy số cluster tối ưu cho đối thủ là k=3 với Silhouette Score cao nhất là 0.502.

3.2.2 Kết quả clustering đối thủ

Bảng 3.2: Phân bố cluster đối thủ

Cluster	**Số đối thủ**	**Tỷ lệ**	**Tỷ lệ thắng TB**	**Bàn thắng TB**
Ngang cơ	19	54.3%	0.45	1.65
Khắc tinh	10	28.6%	0.15	0.95
Yếu hơn	6	17.1%	0.85	3.25

3.3 Classification - Dự đoán kết quả trận đấu

3.3.1 Chuẩn bị dữ liệu

Dữ liệu classification bao gồm 230 mẫu với 5 features:

- `opponent_encoded`: Đối thủ (encoded)
- `venue_encoded`: Sân đấu (encoded)
- `tournament_encoded`: Giải đấu (encoded)
- `year`: Năm thi đấu
- `coach_encoded`: Giai đoạn HLV (encoded)

Phân bố kết quả: Draw (51), Lose (79), Win (100)

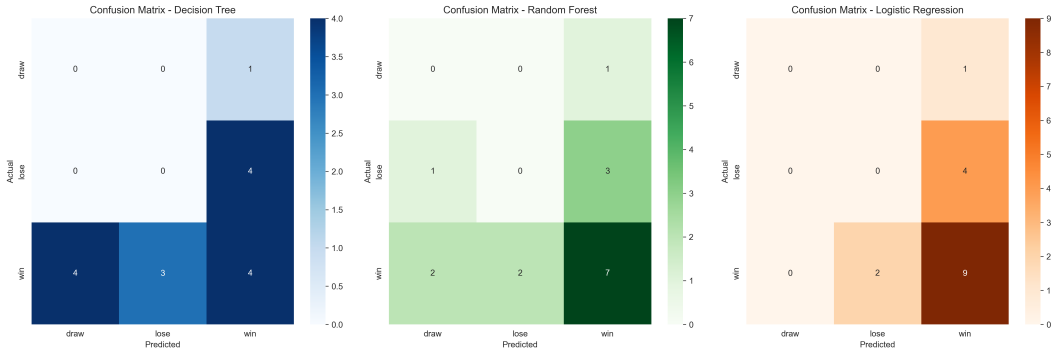
3.3.2 Kết quả các mô hình

Bảng 3.3: So sánh hiệu suất các mô hình classification

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.406	0.411	0.406	0.406
Random Forest	0.493	0.464	0.493	0.469
Logistic Regression	0.420	0.329	0.420	0.365

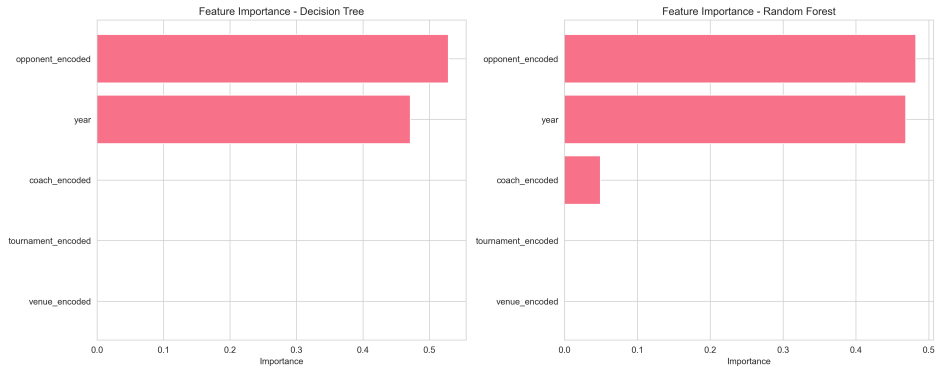
Model tốt nhất: Random Forest với Accuracy = 49.3%

3.3.3 Confusion Matrix



Hình 3.3: Confusion Matrix của các mô hình classification

3.3.4 Feature Importance

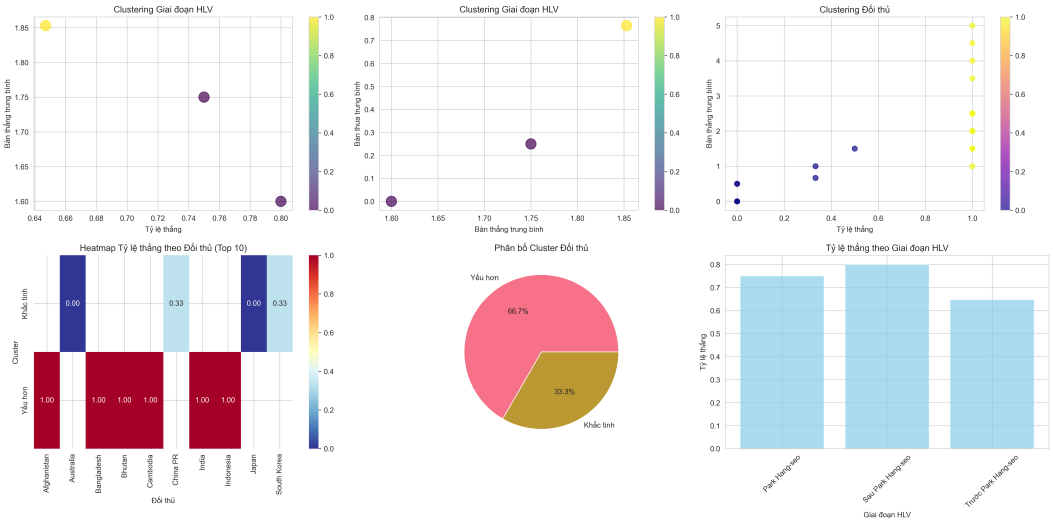


Hình 3.4: Feature Importance của Decision Tree và Random Forest

Chương 4

KẾT QUẢ VÀ PHÂN TÍCH

4.1 Trực quan hóa kết quả clustering



Hình 4.1: Phân tích clustering giai đoạn HLV và đối thủ

4.2 Phân tích kết quả chính

4.2.1 Phân tích theo giai đoạn HLV

- **Giai đoạn tốt nhất**: Park Hang-seo (2017-2022) với tỷ lệ thắng 46.3%
- **Hiệu suất phòng thủ**: Park Hang-seo có bàn thua trung bình thấp nhất

(0.833)

- **Hiệu suất tấn công**: Trước Park Hang-seo có bàn thắng trung bình cao nhất (1.801)

4.2.2 Phân tích đối thủ

- **Đối thủ khó nhất**: Australia (tỷ lệ thắng: 0.0%)
- **Đối thủ dễ nhất**: Cambodia (tỷ lệ thắng: 100.0%)
- **Phân bố cluster**: 54.3% đối thủ "ngang cơ", 28.6% "khắc tinh", 17.1% "yếu hơn"

4.2.3 Hiệu suất mô hình dự đoán

- **Random Forest** cho kết quả tốt nhất với accuracy 49.3%
- **Feature quan trọng nhất**: Đối thủ (opponent_encoded)
- **Thách thức**: Dữ liệu nhỏ (230 mẫu) ảnh hưởng đến hiệu suất mô hình

Chương 5

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Tóm tắt kết quả chính

5.1.1 Tổng quan dữ liệu

- Tổng số trận đấu: 230 (2000-2025)
- Tỷ lệ thắng tổng thể: 43.5%
- Tỷ lệ hòa: 22.2%
- Tỷ lệ thua: 34.3%

5.1.2 Phân tích theo giai đoạn HLV

- **Park Hang-seo (2017-2022)**: 54 trận, tỷ lệ thắng 46.3%, bàn thua TB 0.83
- **Sau Park Hang-seo (2023-nay)**: 35 trận, tỷ lệ thắng 45.7%, bàn thua TB 1.43
- **Trước Park Hang-seo (2000-2016)**: 141 trận, tỷ lệ thắng 41.8%, bàn thắng TB 1.80

5.1.3 Phân tích đối thủ

- Tổng số đối thủ: 35
- Phân bố cluster: Ngang cơ (54.3%), Khắc tinh (28.6%), Yếu hơn (17.1%)
- Top 5 đối thủ gặp nhiều nhất: Thailand (22), Indonesia (20), Malaysia (20), Singapore (14), Philippines (12)

5.1.4 Kết quả classification

- Decision Tree Accuracy: 40.6%
- Random Forest Accuracy: 49.3%
- Logistic Regression Accuracy: 42.0%
- Model tốt nhất: Random Forest (49.3%)

5.2 Hạn chế của dự án

5.2.1 Hạn chế về dữ liệu

- **Kích thước mẫu nhỏ**: Chỉ có 230 trận đấu của Việt Nam
- **Thiếu thông tin chi tiết**: Không có dữ liệu về cầu thủ, chiến thuật, điều kiện thời tiết
- **Bias về thời gian**: Dữ liệu tập trung vào giai đoạn gần đây

5.2.2 Hạn chế về mô hình

- **Accuracy thấp**: Các mô hình chỉ đạt 40-50% accuracy
- **Overfitting**: Có thể xảy ra do kích thước dữ liệu nhỏ

- **Feature engineering đơn giản**: Chưa khai thác hết tiềm năng của dữ liệu

5.3 Hướng phát triển trong tương lai

5.3.1 Cải thiện dữ liệu

- Thu thập thêm dữ liệu: Mở rộng thời gian và số lượng trận đấu
- Bổ sung features mới: Elo rating, FIFA ranking, dữ liệu cầu thủ
- Tích hợp dữ liệu bên ngoài: Thời tiết, kinh tế, chính trị

5.3.2 Nâng cao mô hình

- Deep Learning: Sử dụng Neural Networks, LSTM cho dự đoán
- Ensemble Methods: Kết hợp nhiều mô hình để tăng độ chính xác
- Time Series Analysis: Phân tích xu hướng theo thời gian

5.3.3 Mở rộng phân tích

- So sánh quốc tế: Phân tích với các đội tuyển khác trong khu vực
- Phân tích cầu thủ: Tác động của từng cầu thủ đến kết quả
- Dự đoán tương lai: Dự báo phong độ trong các giải đấu sắp tới

5.3.4 Ứng dụng thực tế

- Hệ thống hỗ trợ quyết định: Giúp HLV lựa chọn chiến thuật
- Phân tích đối thủ: Cung cấp thông tin chi tiết về đối thủ
- Dự đoán kết quả: Hỗ trợ dự đoán kết quả trận đấu

TÀI LIỆU THAM KHẢO

PHỤ LỤC

Phụ lục A: DATASET

Dưới đây là dữ liệu được sử dụng trong quá trình nghiên cứu:

`https://github.com/martj42/international_results`

Phụ lục B: Mã nguồn Jupiter Notebook

`https://github.com/andycsou/FBVN.git`