TRƯỜNG ĐH GIAO THÔNG VẬN TẢI
PHÂN HIỆU TP.HCM
Bộ Môn Công Nghệ Thông Tin

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập – Tự Do -Hạnh Phúc

ĐỀ CƯƠNG CHI TIẾT ĐỒ ÁN TỐT NGHIỆP

1. **Thông tin Sinh viên:**

Họ tên : Võ Minh Kha — Mã sinh viên : 6251071044

Lớp : CQ.62.CNTT — Hệ : Cử nhân

Ngành đào tạo : Công nghệ thông tin — Khoá : 62

Email : 6251071044@st.utc2.edu.vn — Số điện thoại : 0903654735

2. **Thông tin Giảng viên hướng dẫn:**

Họ tên : Trần Phong Nhã — Học vị : Thạc sĩ

Email : tpnha@utc2.edu.vn — Số điện thoại : 0906 761 014

Đơn vị công tác: Trường Đại học Giao thông Vận tải phân hiệu thành phố Hồ Chí Minh

NỘI DUNG

I. Tên đề tài

Phát triển hệ thống dự đoán kết quả bóng đá tích hợp trên nền tảng website

II. Giới thiệu

Bóng đá không chỉ là một môn thể thao giải trí mà còn là một phần không thể thiếu trong đời sống tinh thần của hàng triệu người trên khắp thế giới. Bên cạnh việc xem và cổ vũ, người hâm mộ ngày càng quan tâm đến các hoạt động phân tích, dự đoán kết quả trận đấu để phục vụ nhiều mục đích khác nhau bình luận thể thao, phân tích số liệu,... Trong vài năm trở lại đây, việc áp dụng các kỹ thuật học máy vào thể thao không còn xa lạ, đặc biệt là trong việc xử lý dữ liệu lịch sử và dự đoán kết quả thi đấu – giúp các hệ thống có thể đưa ra đánh giá có cơ sở hơn thay vì chỉ dựa vào cảm tính.

Một trong những ứng dụng thực tiễn là dự đoán xác suất chiến thắng của các đội bóng dựa trên dữ liệu lịch sử. Tuy nhiên, hiện nay vẫn còn thiếu các hệ thống đơn giản, dễ tiếp cận, có khả năng giải thích kết quả một cách minh bạch, đặc biệt là dành cho người dùng phổ thông. Do đó, đề tài này được lựa chọn nhằm xây dựng một trang web thể thao tích hợp tính năng dự đoán kết quả trận

đấu bóng đá từ dữ liệu lịch sử, nhằm hỗ trợ người dùng theo dõi thông tin trận đấu và đưa ra nhận định về kết quả có cơ sở.

Hệ thống sử dụng mô hình học có giám sát (XGBoost) để ước lượng xác suất thắng, hòa, thua, từ đó cung cấp giao diện trực quan, dễ hiểu và có giá trị sử dụng thực tế.

Yêu cầu:

- Xây dựng hệ thống frontend hiển thị thông tin trận đấu, giải đấu, đội bóng và lịch thi đấu.
- Phân tích các yếu tố ảnh hưởng đến kết quả trận đấu như phong độ gần đây, thứ hạng, đối đầu trực tiếp, sân nhà/sân khách...
- Huấn luyện mô hình học máy (XGBoost) để tính ra xác xuất kết quả trận đấu.
- Kết nối mô hình với giao diện web để hiển thị kết quả dự đoán theo tỷ lệ phần trăm.
- Thiết kế giao diện thân thiện với người dùng.

Phạm vi và đối tượng nghiên cứu

Phạm vi dữ liệu:

Dữ liệu lịch sử trận đấu được thu thập từ các nguồn mở như Kaggle và football-data.co.uk tập trung vào các giải đấu lớn như Premier League, La Liga, Bundesliga, Serie A và Ligue.

Dữ liệu bao gồm: kết quả trận đấu, đội bóng, sân nhà/sân khách, bảng xếp hạng và một số chỉ số phong độ gần đây.

Phạm vi xử lý mô hình:

Mô hình học máy được huấn luyện để dự đoán xác xuất kết quả trận đấu (thắng – hòa – thua**)** dựa trên các yếu tố như: phong độ đội bóng, lịch sử đối đầu, vị trí trên bảng xếp hạng, và lợi thế sân nhà. Thuật toán chính sử dụng là XGBoost, thuộc nhóm học có giám sát.

Phạm vi triển khai hệ thống:

Xây dựng giao diện hiển thị trực quan thông tin trận đấu và kết quả dự đoán dưới dạng phần trăm. Hệ thống hướng đến đối tượng người dùng phổ thông, bao gồm người hâm mộ bóng đá và sinh viên công nghệ thông tin có nhu cầu trải nghiệm hệ thống dự đoán có cơ sở.

III. Cơ sở lý thuyết

1. Bài toán phân loại xác suất kết quả trận đấu bóng đá

Trong đề tài này, hệ thống không chỉ đưa ra dự đoán đơn giản là đội nào sẽ thắng, mà mục tiêu là ước lượng xác suất xảy ra của từng kết quả trận đấu, bao gồm:

- Đội chủ nhà thắng,
- Hai đội hòa,

- Đội khách thắng.

Nói cách khác, đây là bài toán phân loại đa lớp (multi-class classification), nhưng thay vì chọn ra một nhãn duy nhất, mô hình sẽ trả về xác suất cho cả ba khả năng, ví dụ:Manchester United thắng 20%, Manchester City thắng38%, hai đội hoà 42%.

Các xác suất này giúp người dùng hiểu rõ mức độ tin cậy cho mỗi kịch bản có thể xảy ra trong trận đấu, thay vì chỉ biết "kết quả dự đoán là đội A thắng".

Để đưa ra dự đoán, mô hình sẽ sử dụng nhiều yếu tố đặc trưng đầu vào (feature) như:

- Phong độ gần đây của cả hai đội,
- Thứ hạng hiện tại trên bảng xếp hạng,
- Lịch sử đối đầu giữa hai đội,
- Sân thi đấu là sân nhà hay sân khách,
- Hiệu số bàn thắng/thua trung bình…

Việc lựa chọn và xử lý các đặc trưng đầu vào này ảnh hưởng lớn đến độ chính xác của mô hình dự đoán.

2. Công nghệ và công cụ sử dụng

Thành phần	Công nghệ	Vai trò
Mô hình học máy	Python, XGBoost, Pandas, Scikit-learn	Huấn luyện và dự đoán kết quả trận đấu
API dự đoán	Flask	Tạo REST API để gọi mô hình từ backend
Backend	Node.js + Express	Xử lý yêu cầu từ frontend, gọi Flask API
Giao diện	ReactJS, Tailwind CSS	Hiển thị giao diện và kết quả dự đoán
Cơ sở dữ liệu	MongoDB	Lưu thông tin trận đấu và lịch sử dự đoán (tuỳ chọn)
Công cụ phát triển	Postman, Git, VSCode	Kiểm thử API và quản lý mã nguồn

3. Kiến trúc hệ thống web và tích hợp mô hình

Hệ thống được thiết kế theo mô hình ba tầng đơn giản và rõ ràng:

- **Frontend (React):** là phần giao diện người dùng, cho phép xem thông tin trận đấu, giải đấu, thực hiện dự đoán và hiển thị kết quả một cách trực quan (dạng biểu đồ phần trăm).
- **Backend chính (Node.js):** xử lý các yêu cầu từ frontend, gọi mô hình dự đoán để lấy kết quả, và (nếu có) kết nối với cơ sở dữ liệu để lưu thông tin trận đấu hoặc lịch sử dự đoán.
- **Mô hình AI (Flask - Python):** được triển khai riêng, nhận dữ liệu đầu vào từ Node.js, thực hiện dự đoán bằng mô hình XGBoost, sau đó trả kết quả xác suất dự đoán dưới dạng JSON.

Luồng hoạt động:

- Người dùng chọn trận đấu → gửi yêu cầu dự đoán từ giao diện.
- Backend Node.js nhận dữ liệu, gửi qua Flask.
- Flask xử lý bằng mô hình → trả xác suất 3 kết quả.
- Kết quả trả về frontend để hiển thị đẹp mắt cho người dùng.

Việc tách riêng mô hình ra khỏi backend chính giúp hệ thống dễ mở rộng, dễ bảo trì, và hoạt động ổn định hơn.

4. Thuật toán XGBoost

Để giải bài toán trên, đề tài sử dụng thuật toán XGBoost – một thuật toán học máy nổi bật hiện nay nhờ hiệu suất cao và khả năng xử lý dữ liệu bảng rất tốt.

XGBoost thuộc nhóm cây quyết định tăng cường (gradient boosting trees). Thay vì học một mô hình duy nhất, nó kết hợp nhiều cây quyết định nhỏ, mỗi cây cố gắng sửa lỗi của những cây trước. Nhờ đó, mô hình dần cải thiện và đưa ra dự đoán tốt hơn.

Một điểm mạnh khác của XGBoost là khả năng trả về xác suất dự đoán cho từng lớp – chính xác là thứ mà đề tài này cần: xác suất cho thắng, hòa, thua.

Ngoài ra, XGBoost còn hỗ trợ:

- Tự động xử lý giá trị thiếu,
- Cơ chế chống quá khớp (regularization),
- Tính toán độ quan trọng của các đặc trưng.

IV. Phương pháp nghiên cứu

1. Thu thập và xử lý dữ liệu đầu vào

Dữ liệu được lấy từ các nguồn mở như Kaggle và football-data.co.uk, tập trung vào các giải đấu lớn như Ngoại hạng Anh (Premier League), La Liga, Bundesliga, Serie A và Ligue 1. Nội dung dữ liệu bao gồm:

- **Thời gian thi đấu** (Date),
- **Tên đội chủ nhà / đội khách** (HomeTeam, AwayTeam),
- **Tỷ số trận đấu** (FTHG, FTAG),
- **Kết quả trận đấu** (FTR: H – Home win, D – Draw, A – Away win),
- **Các chỉ số trận đấu**: tổng số cú sút (HS, AS), sút trúng đích (HST, AST), thẻ phạt (HY, HR, AY, AR), phạt góc, phạm lỗi,...

Sau khi thu thập, dữ liệu được xử lý qua các bước:

- Làm sạch (xử lý giá trị trống, lỗi định dạng),
- Chuẩn hóa (đồng bộ tên đội, mã hóa kết quả),
- Trích xuất đặc trưng (feature engineering): tạo các cột mới như phong độ đội bóng, thành tích đối đâu để mô hình học và đánh giá được sức mạnh của từng đội từ đó đưa ra kết quả trận đấu.

Mục tiêu của bước này là tạo ra một tập dữ liệu đủ sạch và đủ thông tin để huấn luyện mô hình học máy.

2. Xây dựng mô hình dự đoán kết quả trận đấu

Sau khi hoàn tất bước xử lý và chuẩn hóa dữ liệu, hệ thống sẽ tiến hành xây dựng một mô hình học máy nhằm dự đoán xác suất xảy ra của ba kết quả trong một trận đấu bóng đá: đội chủ nhà thắng, hòa, hoặc đội khách thắng.

Đề tài lựa chọn thuật toán XGBoost (Extreme Gradient Boosting) để huấn luyện mô hình. Đây là một thuật toán phổ biến thuộc nhóm học có giám sát, được đánh giá cao nhờ hiệu quả trong xử lý dữ liệu dạng bảng và khả năng đưa ra kết quả dự đoán dưới dạng xác suất.

Dữ liệu được chia thành hai phần: một phần để huấn luyện mô hình, phần còn lại để kiểm tra hiệu quả dự đoán. Mô hình sẽ học từ các đặc trưng đầu vào và ước lượng xác suất cho từng kết quả có thể xảy ra.

Trong quá trình huấn luyện, mô hình được đánh giá dựa trên các tiêu chí cơ bản nhằm đảm bảo tính chính xác và ổn định. Ngoài XGBoost, một số mô hình khác có thể được thử nghiệm bổ sung để so sánh và đối chiếu kết quả trong quá trình nghiên cứu.

3. Phát triển hệ thống web

Sau khi có mô hình dự đoán hoạt động tốt, hệ thống được tích hợp vào một trang web có giao diện thân thiện, dễ sử dụng, hoạt động tốt trên cả máy tính và điện thoại.

Giao diện người dùng (Frontend):

- Xây dựng bằng ReactJS để tạo bố cục hiện đại.
- Dữ liệu dự đoán được hiển thị bằng biểu đồ thanh phần trăm.

Backend:

- Viết bằng Node.js + Express.
- Nhận dữ liệu từ giao diện, xử lý và gửi tới mô hình để nhận kết quả dự đoán.

API dự đoán:

- Xây dựng bằng Flask (Python).
- Load mô hình đã huấn luyện
- Nhận dữ liệu trận đấu từ backend và trả lại xác suất dự đoán.

V. Kết quả dự kiến

Sản phẩm đầu ra:

- Một hệ thống website thể thao hoàn chỉnh với giao diện trực quan, thân thiện với người dùng, có khả năng hiển thị thông tin trận đấu, thực hiện dự đoán và thống kê kết quả.
- Mô hình học máy XGBoost dự đoán xác suất kết quả trận đấu (thắng/hòa/thua).

Giải quyết được vấn đề gì?

Tăng tính chính xác, minh bạch và khoa học trong việc dự đoán kết quả bóng đá.

Hỗ trợ người dùng phổ thông có thêm công cụ để tham khảo trước khi theo dõi hoặc bình luận trận đấu.

Tạo tiền đề cho việc ứng dụng học máy vào các hệ thống thể thao quy mô nhỏ và vừa, phục vụ nghiên cứu học thuật hoặc demo công nghệ tại các sự kiện giáo dục/kỹ thuật.

VI. Đóng góp của đề tài

Về mặt học thuật:

Minh họa rõ ràng quy trình xây dựng mô hình học máy cho bài toán phân loại trong lĩnh vực thể thao.

Cung cấp một ví dụ thực tế về cách tích hợp mô hình ML vào hệ thống web client-server.

Là tài liệu tham khảo hữu ích cho sinh viên nghiên cứu và phát triển ứng dụng tương tự.

Về mặt thực tiễn:

Tạo ra một công cụ đơn giản, dễ tiếp cận để trải nghiệm khả năng ứng dụng AI trong dự đoán thể thao.

Hỗ trợ sinh viên, người học công nghệ hoặc người yêu thích bóng đá trong việc tiếp cận và hiểu rõ hơn về mô hình học máy thông qua tương tác thực tế.

VII. Cấu trúc đồ án

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1 Bối cảnh và lý do chọn đề tài

1.2 Mục tiêu, đối tượng, phạm vi nghiên cứu

1.3 Phương pháp thực hiện và bố cục đồ án

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Kiến thức nền tảng về học máy

2.2 Kiến thức xây dựng hệ thống web hiện đại

2.3 Công nghệ sử dụng trong hệ thống

CHƯƠNG 3: PHÂN TÍCH THIẾT KẾ HỆ THỐNG

3.1 Phân tích yêu cầu chức năng và phi chức năng

3.2 Thiết kế hệ thống tổng thể

3.2.1 Thiết kế sơ đồ cơ sở dữ liệu

3.2.2 Thiết kế chức năng hệ thống

3.3 Phân tích mô hình dự đoán kết quả trận đấu bóng đá

3.3.1 Phân tích bộ dữ liệu

3.3.2 Xử lý chọn lọc các đặc trưng

CHƯƠNG 4: XÂY DỰNG ỨNG DỤNG

4.1 Huấn luyện và triển khai mô hình học máy

4.2 Đánh giá độ chính xác mô hình và kết quả dự đoán

4.3 Cài đặt hệ thống backend và frontend

4.4 Thiết kế giao diện

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Tổng kết quá trình thực hiện đề tài

5.2 Đánh giá ưu điểm và hạn chế của hệ thống hiện tại

5.3 Đề xuất hướng phát triển trong tương lai

VIII. Tài liệu tham khảo

[1] Juanramrezb (2024) What events can define the result of a match?, Kaggle. Available at: https://www.kaggle.com/code/juanramrezb/what-events-can-define-the-result-of-a-match (Accessed: 28/03/2025).

[2] F. Rodrigues and Á. Pinto, "Prediction of football match results with Machine Learning," Procedia Computer Science, vol. 204, pp. 463–470, 2022. Available: https://doi.org/10.1016/j.procs.2022.08.057 (Accessed: 28/03/2025).

[3] Maximdrejdink (2023) Predicting football match results (classification), Kaggle. Available at: https://www.kaggle.com/code/maximdrejdink/predicting-football-match-results-classification(Accessed: 28/03/2025).

[4] API Football, A. Football - documentation, api. Available at: https://www.api-football.com/documentation-v3 (Accessed: 28/03/2025).

[5] Brownlee, J. (2021) Extreme gradient boosting (XGBoost) ensemble in Python, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/ (Accessed: 28/03/2025).

[6] Getting started Getting Started | Axios Docs. Available at: https://axios-http.com/docs/intro (Accessed: 28/03/2025).

[7] Welcome to Flask - Flask Documentation. Available at: https://flask.palletsprojects.com/en/stable/ (Accessed: 28/03/2025).

[8] Getting started vitejs. Available at: https://vite.dev/guide/ (Accessed: 28/03/2025).

IX. Kế Hoạch thực hiện và tiến độ nghiên cứu

Thời gian	Nội dung công việc	Ghi chú
Tuần 1–2	Tìm hiểu đề tài, xác định hướng triển khai và hoàn thiện đề cương.	
Tuần 3–4	Thu thập dữ liệu, làm sạch và phân tích các yếu	

	tố ảnh hưởng đến kết quả trận đấu.	
Tuần 5–6	Trích xuất đặc trưng đầu vào và xây dựng mô hình học máy dự đoán xác suất thắng – hòa – thua.	
Tuần 7–9	Xây dựng backend và tích hợp mô hình AI vào hệ thống web.	
Tuần 10–13	Phát triển giao diện web (React + Tailwind), tập trung thiết kế bố cục, responsive và phần hiển thị kết quả dự đoán.	
Tuần 14–15	Kiểm thử hệ thống, sửa lỗi giao diện và tối ưu trải nghiệm người dùng.	
Tuần 16	Hoàn thiện báo cáo, chuẩn bị slide và nội dung thuyết trình.	

ngày tháng 04 năm 2025

Trưởng Bộ Môn	Ý kiến của GVHD	Sinh viên thực hiện
ThS.Trần Phong Nhã	ThS. Trần Phong Nhã	Võ Minh Kha