**A white rectangular frame with black border

Description automatically generated**

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



BÁO CÁO BÀI TẬP LỚN HỌC PHẦN

Học máy

TÊN ĐỀ TÀI: Dự đoán kết quả trận bóng đá

LỚP: 64ktpm2 - NHóm 6

THÀNH viên: Vũ duy hưng -2251172365

Bùi quốc bình -2251172250

Nguyễn MInh chiến -2251172254

Tạ Tuấn anh -2251172240

Hà Nội, năm 2024

LỜI CÁM ƠN

Nhóm em xin gửi lời cảm ơn chân thành đến Thầy Tạ Quang Chiểu đã tận tình giảng dạy và truyền đạt những kiến thức trong môn học Máy học. Những kiến thức mà thầy chia sẻ đã giúp nhóm em có nền tảng vững chắc để thực hiện bài tập lớn, áp dụng các phương pháp và kỹ thuật đã học vào thực tiễn. Nhờ đó, em đã hiểu rõ hơn về cách xây dựng và đánh giá mô hình, cũng như biết cách giải quyết các vấn đề thực tế.

Một lần nữa, nhóm em xin cảm ơn và mong rằng sẽ tiếp tục nhận được sự hướng dẫn từ thầy trong quá trình học tập và nghiên cứu sắp tới.

MỤC LỤC

[CHƯƠNG 1 GIỚI THIỆU 1](#_Toc181180557)

[CHƯƠNG 2 CƠ SỞ LÝ THUYẾT 3](#_Toc181180558)

[*1.* *Giới thiệu về học máy* 3](#_Toc181180559)

[*2.* *Logistic Regression* 3](#_Toc181180560)

[*3.* *Thuật toán ID3* 4](#_Toc181180561)

[*4.* *Accuracy* 6](#_Toc181180562)

[*5.* *MAE (Mean Absolute Error)* 6](#_Toc181180563)

[*6.* *RR (R-squared) – Hệ số xác định* 6](#_Toc181180564)

[CHƯƠNG 3 XÂY DỰNG VÀ PHÂN TÍCH MÔ HÌNH 7](#_Toc181180565)

[*1.* *Dữ liệu* 7](#_Toc181180566)

[*2.* *Tiền xử lý dữ liệu* 8](#_Toc181180567)

[*3.* *Phân chia dữ liệu* 9](#_Toc181180568)

[*4.* *Biểu đồ phân bố dữ liệu* 9](#_Toc181180569)

[*5.* *Huấn luyện mô hình* 10](#_Toc181180570)

[*6.* *Đánh giá mô hình* 11](#_Toc181180571)

[CHƯƠNG 4 KẾT QUẢ VÀ THẢO LUẬN 11](#_Toc181180572)

[*1.* *Logistic Regression* 11](#_Toc181180573)

[*2.* *ID3* 13](#_Toc181180574)

[*3.* *Kết luận* 14](#_Toc181180575)

[TÀI LIỆU THAM KHẢO 15](#_Toc181180576)

# GIỚI THIỆU

1. **Lý do chọn đề tài**

Dự đoán kết quả bóng đá là một bài toán hấp dẫn, thu hút sự quan tâm của nhiều người đam mê thể thao và các nhà khoa học dữ liệu. Trong lĩnh vực này, việc dự đoán chính xác kết quả của một trận đấu không chỉ mang lại lợi ích về mặt giải trí mà còn có ý nghĩa thực tiễn trong việc phân tích chiến thuật, quản lý đội bóng, và thậm chí là hỗ trợ ra quyết định trong các hoạt động kinh doanh liên quan đến cá cược thể thao.

Bài toán này thường được giải quyết thông qua các mô hình học máy, trong đó Logistic Regression và thuật toán ID3 là hai phương pháp phổ biến. **Logistic Regression** là một phương pháp thống kê dùng để phân loại nhị phân, rất phù hợp với các bài toán có kết quả hai khả năng (như "Thắng" hoặc "Thua/Hòa"). Phương pháp này dựa trên việc tính toán xác suất và cho phép đưa ra kết quả dưới dạng xác suất thắng của đội nhà, qua đó hỗ trợ dự đoán kết quả với độ chính xác cao. Trong khi đó, **thuật toán ID3** là một kỹ thuật xây dựng cây quyết định, phân chia dữ liệu dựa trên các đặc trưng quan trọng để tối đa hóa thông tin thu được. ID3 tạo ra một hệ thống phân loại trực quan, dễ hiểu, giúp người dùng dễ dàng nắm bắt được lý do đằng sau các dự đoán của mô hình.

Sử dụng hai phương pháp này, chúng em sẽ thử nghiệm và so sánh để tìm ra phương pháp hiệu quả nhất trong việc dự đoán kết quả bóng đá dựa trên các đặc trưng như số bàn thắng của đội nhà và đội khách, phong độ, và thành tích gần đây của hai đội. Qua đó, chúng ta có thể đánh giá khả năng dự đoán của Logistic Regression và ID3 trong bối cảnh thực tế, đồng thời tìm hiểu sâu hơn về ưu nhược điểm của từng phương pháp khi áp dụng vào bài toán phân loại nhị phân.

1. **Mục đích nghiên cứu**

Mục đích của đề tài là xây dựng mô hình dự đoán kết quả trận đấu bóng đá bằng cách sử dụng các thuật toán học máy như Logistic Regression, ID3 và các thuật toán khác. Đề tài hướng đến:

* Tìm hiểu và đánh giá hiệu quả của các mô hình dự đoán trong việc phân loại và dự đoán kết quả trận đấu.
* Phân tích các yếu tố quan trọng ảnh hưởng đến kết quả trận đấu, chẳng hạn như phong độ đội bóng, chỉ số trận đấu, số trận thắng, và các yếu tố ngoại cảnh (như sân nhà, điều kiện thời tiết).
* Xây dựng công cụ dự đoán có thể áp dụng trong các ứng dụng thực tiễn như phân tích đối thủ, và đưa ra gợi ý chiến thuật cho các đội bóng.

# CƠ SỞ LÝ THUYẾT

## *Giới thiệu về học máy*

Machine learning đã đóng một vai trò quan trọng trong sự phát triển của trí tuệ nhân tạo và thúc đẩy tiến bộ trong nhiều lĩnh vực khác nhau. Với sự gia tăng về khả năng tính toán dữ liệu và nghiên cứu trong lĩnh vực này, machine learning dự kiến sẽ tiếp tục phát triển và mang lại những ứng dụng mới và tạo ra cá tiến bộ quan trọng trong tương lai.

* 1. **Ưu điểm và hạn chế của học có giám sát và không giám sát**
     1. **Học có giám sát**
        1. **Ưu điểm**
           1. Độ chính xác cao: tạo ra các mô hình có độ chính xác cao đối với dữ liệu kiểm tra, đặc biệt khi có đủ dữ liệu đào tạo
           2. Dễ triển khai
           3. Có khả năng giải thích kết quả dự đoán vì mô hình được đào tạo dựa trên dữ liệu đầu vào
        2. **Nhược điểm**
           1. Đòi hỏi dữ liệu gán nhãn
           2. Dữ liệu phải đại diện: dữ liệu đào tạo phải đại diện cho thực tế, nếu không mô hình có thể dự đoán kém chính xác hoặc bị overfitting
     2. **Học không giám sát**
        1. **Ưu điểm**
           1. Khám phá cấu trúc dữ liệu
           2. Khám phá dữ liệu mới
        2. **Nhược điểm**
           1. Khó đánh giá chất lượng
           2. Thường cần nhiều dữ liệu hơn
           3. Kết quả khó giải thích vì không có đầu ra đã biết để so sánh

## *Logistic Regression*

* 1. **Mục đích của phương pháp**

Mục đích chính của Logistic Regression là:

* **Phân loại nhị phân**: Xác định xác suất của một biến phụ thuộc thuộc về một trong hai nhóm.
* **Dự đoán xác suất**: Tính toán xác suất của một biến đầu ra (thường là 0 hoặc 1) xảy ra dựa trên một hoặc nhiều biến đầu vào.
  1. **Ý tưởng**

Ý tưởng của Logistic Regression dựa trên việc áp dụng hàm **sigmoid (logistic)** để biến đổi đầu ra của hồi quy tuyến tính thành giá trị xác suất trong khoảng từ 0 đến 1 ,đại diện cho khả năng xảy ra của sự kiện được dự đoán.

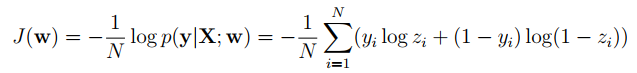
* 1. **Hàm Sigmoid**

**A math equation with numbers and letters

Description automatically generated**

* 1. **Hàm mất mát**

Dùng hàm mất mát log loss để đánh giá sai số giữa dự đoán của mô hình và nhãn thực tế:

Trong đó, yi​ là nhãn thực tế, và zi là xác suất dự đoán.

## *Thuật toán ID3*

* 1. **Mục đích của phương pháp:** xây dựng cây quyết định phân loại dữ liệu tự động dựa trên các thuộc tính của dữ liệu đầu vào
  2. **Ý tưởng**
     1. Chúng ta cần xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước
     2. Tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn nào đó
     3. Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các nút con tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi nút con
     4. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn tham lam (greedy). Cách chọn này có thể không phải là tối ưu, nhưng trực giác cho chúng ta thấy rằng cách làm này sẽ gần với cách làm tối ưu
  3. **Hàm số Entropy**
     1. **Entropy của phân phối xác suất**

**A mathematical equation with numbers and symbols

Description automatically generated**

* + 1. **Entropy của một tập các điểm dữ liệu S**

**A mathematical formula with numbers and symbols

Description automatically generated**

* + 1. **Entropy của 1 thuộc tính**

**A mathematical equation with numbers and symbols

Description automatically generated**

* + 1. **Information gain dựa trên thuộc tính**

**A black and white text

Description automatically generated**

* + 1. **Thuộc tính tốt nhất**

**A mathematical equation with black letters

Description automatically generated**

* 1. **Thuật toán**
     1. Tạo một nút N
     2. Nếu tất cả các mẫu thuộc cùng lớp C thì N được gán nhãn C; dừng thuật toán;
     3. Nếu A là rỗng thì N được gán nhãn C là nhãn phổ biến nhất trong S; dừng thuật toán;
     4. Chọn aA, có độ đo information gain cao nhất; Gán nhãn N theo a;
     5. Với mỗi giá trị v của a:
        1. Phát triển 1 nhánh từ N với điều kiện a=v;
        2. Đặt Sv là tập con của S với a=v;
        3. Nếu Sv là rỗng thì gắn một lá có nhãn phổ biến nhất trong S;
        4. Ngược lại gắn một nút được tạo bởi GenDecTree(Sv , A-a)

## *Accuracy*

Accuracy là chỉ số đo lường tỷ lệ phần trăm các mẫu dữ liệu được phân loại chính xác. Nó được tính bằng công thức sau:

Accuracy = số lượng dự đoán chính xác / tổng số mẫu dữ liệu

## *MAE (Mean Absolute Error)*

MAE là trung bình của các giá trị tuyệt đối của lỗi dự đoán. Nó đo lường mức độ sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế.

A number and numbers with black text

Description automatically generated with medium confidence

* **Ý nghĩa**: MAE cho biết mô hình của bạn dự đoán trung bình sai lệch bao nhiêu so với giá trị thực. MAE càng nhỏ thì mô hình càng chính xác.
* **Ưu điểm**: MAE dễ hiểu vì nó sử dụng sai số tuyệt đối (không tính bình phương) nên không bị ảnh hưởng lớn bởi các lỗi cực lớn.
* **Nhược điểm**: Do MAE không tính đến hướng sai lệch (dự đoán thấp hơn hoặc cao hơn), nó không cho biết thông tin về sự thiên lệch của mô hình.

## *RR (R-squared) – Hệ số xác định*

R² là một chỉ số cho biết tỷ lệ phương sai của biến phụ thuộc (giá trị thực) được giải thích bởi mô hình hồi quy. Nó còn gọi là **Coefficient of Determination**.

A math equations with numbers

Description automatically generated with medium confidence

A white background with black text

Description automatically generated

* **Ý nghĩa**: R² có giá trị từ 0 đến 1 (hoặc có thể âm nếu mô hình rất kém). Giá trị R² càng gần 1 thì mô hình càng giải thích tốt sự biến thiên của dữ liệu.
* **Giải thích**:
  + **R² = 1**: Mô hình dự đoán hoàn hảo giá trị thực.
  + **R² = 0**: Mô hình không giải thích được gì, chỉ dựa vào giá trị trung bình của dữ liệu.
  + **R² < 0**: Mô hình dự đoán kém hơn cả giá trị trung bình của dữ liệu.

# XÂY DỰNG VÀ PHÂN TÍCH MÔ HÌNH

## *Dữ liệu*

Dataset: <https://www.kaggle.com/code/saife245/football-match-prediction/input>

Dữ liệu sử dụng trong nghiên cứu bao gồm 6840 mẫu dữ liệu của các trận đấu bóng đá. Các biến đầu vào được sử dụng bao gồm:

A screenshot of a computer

Description automatically generated

* **FTHG:** (Full Time Home Team ) Số bàn thắng của đội chủ nhà khi kết thúc trận đấu
* **FTAG:** (Full-Time Away Team ) Số bàn thắng của đội khách khi kết thúc trận đấu
* **FTR:** (Full-Time Result) Kết quả cuối cùng
* **HTGS**: Số bàn thắng trung bình của đội nhà (Home Team Goals Scored).
* **ATGS:** Số bàn thắng trung bình của đội khách (Away Team Goals Scored).
* **HTGD:** Hiệu số bàn thắng của đội nhà (Home Team Goal Difference).
* **ATGD:** Hiệu số bàn thắng của đội khách (Away Team Goal Difference).
* **DiffPts:** Chênh lệch điểm số giữa hai đội.
* **DiffFormPts:** Chênh lệch điểm số theo phong độ gần đây của hai đội.

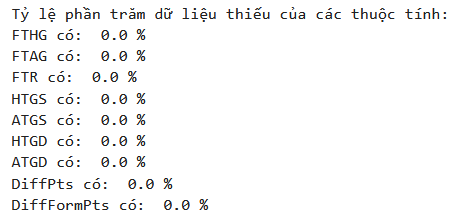
## *Tiền xử lý dữ liệu*

Trước khi huấn luyện mô hình, các bước xử lý dữ liệu sau đã được thực hiện:

* **Chọn các đặc trưng cần thiết**



* **Xử lý giá trị thiếu**: Không có các giá trị thiếu trong các cột.



* **Mã hóa dữ liệu**

Chuyển đổi cột FTR (Full Time Result) kết quả trận đấu thành nhãn (thắng đội nhà = 1, thua hoặc hòa = 0)



## *Phân chia dữ liệu*

1. Logistic Regression

Khởi tạo StratifiedKFold để chia tập dữ liệu thành 5 phần để kiểm định chéo (cross-validation).

1. ID3

Dữ liệu được chia thành hai tập:

* **Tập huấn luyện**: 80% dữ liệu được sử dụng để huấn luyện mô hình.
* **Tập kiểm tra**: 20% dữ liệu còn lại được giữ lại để đánh giá hiệu suất mô hình.

Việc phân chia dữ liệu được thực hiện ngẫu nhiên bằng cách sử dụng hàm **train\_test\_split** từ thư viện scikit-learn, đảm bảo rằng các mẫu được phân phối đều giữa tập huấn luyện và tập kiểm tra.

## *Biểu đồ phân bố dữ liệu*

#### Biểu đồ phân phối các đặc trưng (Histogram)

A group of blue and white graphs

Description automatically generated

Biểu đồ Histogram là một công cụ trực quan giúp hiểu rõ cách các giá trị của từng đặc trưng (hoặc biến) trong tập dữ liệu được phân bố.

* **Phân bố của dữ liệu**:
* **Phân bố chuẩn (Normal Distribution)**: Histogram có dạng hình chuông (bell-shaped), với nhiều giá trị tập trung quanh trung tâm và giảm dần về hai phía.
* **Phân bố lệch trái/phải (Skewed Distribution)**: Nếu biểu đồ có phần đuôi kéo dài về bên trái hoặc bên phải, dữ liệu có thể bị lệch trái hoặc lệch phải.
* **Phân bố đều (Uniform Distribution)**: Khi các thanh trong biểu đồ đều nhau, dữ liệu được phân phối đều (khá hiếm trong thực tế).
* **Giá trị ngoại lệ (Outliers)**:
* Các thanh đơn lẻ nằm cách xa hẳn các thanh còn lại cho thấy sự tồn tại của các giá trị ngoại lệ (outliers) trong dữ liệu.

## *Huấn luyện mô hình*

Các mô hình được huấn luyện bao gồm:

* Logistic Regression
* ID3

## *Đánh giá mô hình*

Mô hình được đánh giá trên tập kiểm tra dựa trên:

* **Accuracy** là chỉ số đo lường tỷ lệ phần trăm các mẫu dữ liệu được phân loại chính xác.
* **MAE (Mean Absolute Error)**: Đo lường sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế. MAE càng thấp thì mô hình càng chính xác.
* **R² (Coefficient of Determination)**: Chỉ số này đo lường mức độ giải thích của mô hình đối với sự biến thiên của dữ liệu. R² càng gần 1 thì mô hình càng tốt.

# KẾT QUẢ VÀ THẢO LUẬN

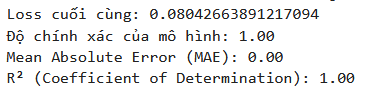
## *Logistic Regression*

1. Dùng thư viện sklearn

A close up of a text

Description automatically generated

1. Không dùng thư viện



* **Loss Cuối Cùng: 0.0804**
* Giá trị loss thấp (gần 0) cho thấy mô hình của bạn hoạt động rất tốt trong việc dự đoán các kết quả. Loss này chỉ ra rằng sai số giữa giá trị dự đoán và giá trị thực tế là rất nhỏ.

A graph with a line

Description automatically generated

* Đồ thị Loss qua các Epochs
* **Độ Chính Xác (Accuracy) của Mô Hình: 1.00 (100%)**
* Một độ chính xác 100% có nghĩa là mô hình của bạn đã dự đoán chính xác tất cả các trường hợp trong tập dữ liệu thử nghiệm. Điều này cho thấy mô hình đã học rất tốt từ dữ liệu huấn luyện.
* Tuy nhiên, độ chính xác 100% cũng có thể là dấu hiệu của hiện tượng **overfitting**, đặc biệt nếu tập dữ liệu không đủ lớn hoặc không đa dạng. Nếu mô hình hoạt động quá tốt trên tập huấn luyện nhưng kém trên dữ liệu mới, điều đó cho thấy nó đã học thuộc lòng dữ liệu thay vì tìm ra các mẫu tổng quát.
* **Mean Absolute Error (MAE): 0.00**
* MAE bằng 0 có nghĩa là không có sai lệch giữa giá trị dự đoán và giá trị thực tế. Điều này càng củng cố thêm cho việc mô hình đã dự đoán chính xác hoàn toàn trên tập dữ liệu.
* **R² (Coefficient of Determination): 1.00**
* Giá trị R² bằng 1 cũng chỉ ra rằng mô hình giải thích hoàn toàn biến thiên của dữ liệu. Điều này có nghĩa là không có phần nào trong dữ liệu không được mô hình giải thích.

## *ID3*

1. Sử dụng thư viện

A close up of a text

Description automatically generated

1. Không dùng thư viện

A black text on a white background

Description automatically generated

* **Accuracy: 0.81** - Độ chính xác là 81%, nghĩa là mô hình dự đoán đúng 81% các trường hợp. Độ chính xác này ở mức khá.
* **Mean Absolute Error (MAE): 0.19** - Sai số trung bình tuyệt đối là 0.19, cho thấy dự đoán của mô hình trung bình chênh lệch 0.19 đơn vị so với giá trị thực tế. MAE thấp cho thấy mô hình dự đoán khá sát với giá trị thực.

A graph with a line

Description automatically generated

* **R²: 0.22** - Giá trị R² khá thấp, chỉ đạt 0.22, nghĩa là mô hình chỉ giải thích được 22% biến thiên của dữ liệu. Điều này cho thấy mô hình có thể chưa bắt được nhiều xu hướng trong dữ liệu.

A diagram of a number

Description automatically generated with medium confidence

* Biểu đồ cây quyết định

## *Kết luận*

* **ID3** cho thấy kết quả dự đoán khá tốt nhưng có khả năng chưa bắt được nhiều xu hướng sâu trong dữ liệu (thể hiện qua R² thấp).
* **Logistic Regression** đạt kết quả hoàn hảo nhưng có thể đã bị quá khớp, làm giảm khả năng dự đoán chính xác trên dữ liệu mới.

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | https://www.kaggle.com/code/saife245/football-match-prediction/notebook |
| [2] | https://machinelearningcoban.com/2018/01/14/id3/ |
| [3] | https://machinelearningcoban.com/2017/01/27/logisticregression/ |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |