TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG – HCM

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN:

CHỦ ĐỀ:

"Decision Tree"

Giảng viên: NGUYỄN TIẾN HUY

Họ và tên	MSSV
THIỀU QUANG VINH	23127143
BÙI ĐỨC ĐẠT	23127337
LỮ BẢO ĐẠT	23127338
ĐINH TUẤN DUY	23127356

TP. Hồ Chí Minh – 04/2025

MỤC LỤC

1. MỞ ĐẦU

- Mục đích thực hiện: Xây dựng cây quyết định (Decision Tree) dựa trên bộ dữ liệu đã cho và đánh giá tính chính xác của mô hình.
- Bộ dữ liệu được sử dụng:
    + Bộ dữ liệu UCI Heart Disease gồm 303 mẫu khảo sát tại Cleveland về số người có bị bệnh tim hay không dựa vào các đặc điểm được cho
    + Bộ dữ liệu Palmer Penguins gồm 344 mẫu khảo sát phân loại giữa 3 giống chim cánh cụt: Adelie, Chinstrap và Gentoo dựa vào các đặc điểm được cho.
- Phương pháp huấn luyện và kiểm tra: Ta sẽ chia tỷ lệ train/ test trong bộ dữ liệu thành 4 tỉ lệ sau (đối với mỗi bộ dữ liệu): 40/60, 60/40, 80/20, 90/10 (sử dụng xáo trộn ngẫu nhiên và đảm bảo train_samples/test_samples giữa các label sai lệch ít nhất)
- Cây quyết định lấy số liệu đo lường từ hàm Entropy và Information Gain.
- Các số liệu và kết quả đã được trực quan hoá trong file .ipynb

2. PHÂN CÔNG CÔNG VIỆC

Họ và tên	MSSV	Công việc	Mức độ hoàn thành
THIỀU QUANG VINH	23127143	Analyzing statistics and writing report	100%
BÙI ĐỨC ĐẠT	23127337	Data Preparation	100%
LỮ BẢO ĐẠT	23127338	Depth and Accuracy of Decision Tree	100%
ĐINH TUẤN DUY	23127356	Decision Tree Implementation and Performance	100%

3. TỰ ĐÁNH GIÁ

3.1. Về tổng quát

STT	Tiêu chí	Điểm
1	Phân tích bộ dữ liệu Heart Disease.	30%
2	Phân tích bộ dữ liệu Palmer Penguins.	30%
3	Phân tích thêm một bộ dữ liệu khác.	30%
4	Phân tích so sánh giữa ba bộ dữ liệu.	5%
5	Notebook được trình bày và định dạng rõ ràng.	5%
	Tổng cộng	100%

3.2. Đối với mỗi bộ dữ liệu:

STT	Tiêu chí	Điểm
1	Chuẩn bị dữ liệu.	30%
2	Triển khai mô hình cây quyết định (Decision Tree).	20%
3	Đánh giá hiệu năng mô hình cây quyết định.	
	- Báo cáo phân loại và ma trận nhầm lẫn.	10%
	- Nhận xét, giải thích kết quả.	10%
4	Chiều sâu và độ chính xác của cây quyết định.	
	- Trực quan hóa (cây, bảng, biểu đồ).	20%
	- Nhận xét, giải thích kết quả.	10%
	Tổng cộng	100%

4. DATA PROCESSING AND PREPARATION

4.1. Overview of Datasets

4.1.1. UCI Heart Disease Dataset

- **Số lượng mẫu**: 303
- **Đặc trưng (Features)**: Bao gồm các đặc trưng về sức khỏe như tuổi (age), giới tính (sex), loại đau ngực (chest pain type), huyết áp khi nghỉ (resting blood pressure), cholesterol, đường huyết lúc đói (fasting blood sugar), kết quả điện tâm đồ khi nghỉ (resting ECG), nhịp tim tối đa (max heart rate), đau thắt ngực do vận động (exercise-induced angina), độ chênh ST (ST depression), dốc của đoạn ST (slope of peak), số lượng mạch máu chính được nhuộm màu (number of vessels), và thalassemia.
- **Nhãn (Label)**: Nhị phân (0: không mắc bệnh, 1: mắc bệnh tim).

4.1.2. Palmer Penguins Dataset

- **Số lượng mẫu**: 344
- **Đặc trưng:** Bao gồm tuổi (age), nghề nghiệp (job), tình trạng hôn nhân (marital), học vấn (education), số dư tài khoản (balance), thông tin liên hệ (contact), thời gian gọi (duration), số lần tiếp xúc (campaign), v.v.
- **Nhãn**: Loài chim cánh cụt (Adelie, Chinstrap, Gentoo – phân loại đa lớp).

4.1.3. Dữ liệu bổ sung Bank marketing campaigns dataset | Opening Deposit

- **Số lượng mẫu**: 11162
- **Đặc trưng**: Tuổi (age), Nghề nghiệp (job), Tình trạng hôn nhân (marital), Trình độ học vấn (education), tình trạng nợ xấu (default), số dư (balance), tình trạng nhà ở (housing), Có vay tiêu dùng (loan), Phương thức liên lạc (contact), Ngày gọi (day), Tháng gọi (month), Thời lượng (duration), Số lần tiếp xúc trong chiến dịch hiện tại (campaign), Số ngày từ lần liên hệ gần nhất (pdays), Số lần liên hệ trước đây (previous), Kết quả chiến dịch trước (poutcome);
- **Nhãn:** deposit (yes/no – phân loại nhị phân).

4.2. Data Preprocessing

4.2.1. Kiểm tra và xử lý dữ liệu thiếu (missing data)

- **Heart Disease**: Phát hiện một số đặc trưng như cholesterol, thal có giá trị thiếu. Với các dữ liệu còn thiếu, trực tiếp xóa dòng thiếu dữ liệu do dữ liệu còn thiếu chiếm khá ít trong tổng số dữ liệu.
- **Penguins**: trực tiếp loại bỏ missing data.
- **Bank marketing campaigns dataset | Opening Deposit**: Các dữ liệu “unknow" chiếm tỉ lệ khá lớn trong tổng số toàn bộ dữ liệu, vẫn dữ lại unknow coi như là một phần data.

4.2.2. Chuẩn hóa dữ liệu phân loại (Encoding Categorical Variables)

- **Heart Disease**: Dữ liệu đã sẵn ở dạng số thực, không cần chuẩn hóa
- **Penguins**:
+ Encode ‘sex’ về dạng bool.
+ One-hot encode đối với ‘island'.
- **Bank marketing campaigns dataset | Opening Deposit:**
+ Encode : deposit, loan, default, housing về bool.
+ One-hot encode: marital, education, contact, month, poutcome, job.

4.2.3. Chuẩn hóa dữ liệu số (Feature Scaling)

Cây quyết định không yêu cầu chuẩn hóa dữ liệu số nên bước này được bỏ qua.

4.2.4. Chia tập huấn luyện và kiểm tra (Train/Test Splitting)

- Sử dụng train_test_split của scikit-learn với tham số stratify theo nhãn để đảm bảo tỷ lệ nhãn giữa các tập không bị lệch.
- Các tỷ lệ chia: 40/60, 60/40, 80/20 và 90/10 (huấn luyện/kiểm tra). Tổng cộng tạo ra 16 tập con.
- Visual hóa phân bố nhãn bằng biểu đồ cột để xác thực việc chia dữ liệu đã hợp lý.

5. XÂY DỰNG MÔ HÌNH (MODEL BUILDING)

Sử dụng DecisionTreeClassifier của thư viện scikit-learn với criterion là 'entropy' (tức là dùng chỉ số thông tin - information gain).

5.1. Quá trình huấn luyện

- Mỗi tỷ lệ train/test, huấn luyện một mô hình cây quyết định mới.
- Sử dụng export_graphviz và Graphviz để trực quan hóa cây quyết định, giúp dễ dàng giải thích mô hình.

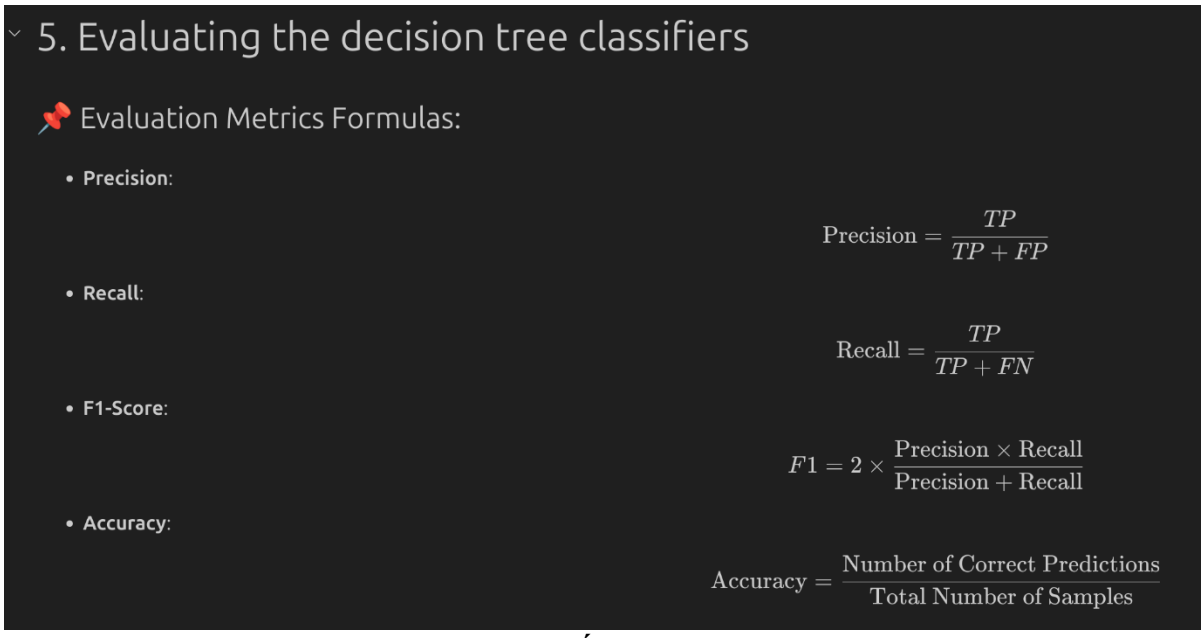
5.2. Thiết lập tham số mô hình

- Các tham số giữ nguyên mặc định, chỉ thay đổi max_depth khi khảo sát độ sâu cây.
- Đặt random_state cố định để đảm bảo có thể tái lập kết quả.

6. ĐÁNH GIÁ MÔ HÌNH (MODEL EVALUATION)

6.1. Các chỉ số sử dụng

- **Classification Report**: Báo cáo chi tiết độ chính xác (precision), độ bao phủ (recall), F1-score cho từng lớp.
- **Confusion Matrix**: Ma trận thể hiện số lượng dự đoán đúng/sai trên từng lớp.
- **Accuracy Score**: Tỷ lệ mẫu dự đoán đúng trên tổng số mẫu.
- Công thức đánh giá:



Hình 1. Tham số đánh giá mô hình

6.2. Trực quan hóa

- Sử dụng seaborn heatmap để vẽ confusion matrix.
- Vẽ biểu đồ so sánh accuracy khi thay đổi max_depth.

7. NHẬN XÉT & GIẢI THÍCH KẾT QUẢ (MODEL INSIGHTS & INTERPRETATION)

7.1. Tổng quan kết quả

- Heart Disease: Nếu không giới hạn độ sâu (max_depth), mô hình có thể overfit, đạt accuracy rất cao trên tập huấn luyện nhưng giảm trên tập kiểm tra. Khi giới hạn độ sâu, mô hình khái quát tốt hơn.
- Penguins: Mô hình phân biệt tốt giữa ba loài chim.
- Bank: Mô hình đạt tỉ lệ đúng gần như ngang nhau khi độ sau cao, cho thấy tập dữ liệu không bị nhiễu quá nhiều.

7.2. Ảnh hưởng của độ sâu cây

- Heart Disease:

max_depth	None	2	3	4	5	6	7
Accuracy	0.7667	0.7167	0.8167	0.7667	0.7667	0.7333	0.7667

- Nhận xét:

Hiện tượng Underfitting: Khi max_depth quá nhỏ (ví dụ max_depth = 2), cây không đủ phức tạp để học được các mối quan hệ quan trọng giữa các đặc trưng, dẫn đến accuracy thấp do underfitting (mô hình học chưa đủ).

Hiện tượng Overfitting: Khi max_depth tăng quá cao (6, 7 hoặc None), accuracy trên tập kiểm tra không tăng mà còn giảm. Điều này cho thấy mô hình bắt đầu học cả nhiễu và chi tiết không quan trọng trong tập huấn luyện, dẫn đến overfitting (quá khớp).

Độ sâu vừa phải (max_depth = 3): Mức độ sâu này giúp mô hình đủ khả năng "học" các mối quan hệ quan trọng mà không bị quá phức tạp hoặc học nhiễu, mang lại accuracy cao nhất trên tập kiểm tra.

Sự dao động của accuracy: Việc accuracy dao động nhẹ khi tăng độ sâu (ở các giá trị 4, 5, 6, 7) là điều thường thấy do đặc điểm của mẫu dữ liệu và cách cây quyết định phân chia các nút. Nếu tập dữ liệu nhỏ hoặc có nhiễu, các giá trị này có thể biến động nhẹ.

- Palmer Penguins:

max_depth	None	2	3	4	5	6	7
Accuracy	0.9254	0.9403	0.9552	0.9701	0.9254	0.9254	0.9254

- Nhận xét:

Độ sâu cây quyết định và khả năng tổng quát hóa:

- Ở max_depth thấp (None, 2), cây chưa đủ phức tạp để học hết các đặc trưng quan trọng → accuracy chưa cao.
- Khi tăng max_depth lên 3, 4, mô hình học được các mối quan hệ phức tạp hơn giữa các đặc trưng, cải thiện khả năng phân biệt giữa các loài chim cánh cụt → accuracy tăng mạnh.
- Độ sâu tối ưu ở đây là 4, có thể xem là mức phức tạp vừa đủ để mô hình học tốt mà chưa bị overfit.
- Overfitting khi tăng max_depth quá cao:
- Khi max_depth ≥ 5, accuracy lại giảm nhẹ và giữ ổn định. Điều này cho thấy mô hình bắt đầu học thêm cả những chi tiết nhỏ, nhiễu trong dữ liệu huấn luyện (quá khớp), khiến khả năng tổng quát hóa trên tập kiểm tra giảm xuống.

Dữ liệu Palmer Penguins có sự tách biệt tốt giữa các lớp:

Đặc trưng dữ liệu này (các loài chim cánh cụt khác nhau về kích thước, hình dạng, v.v.) giúp mô hình đạt accuracy rất cao ngay cả khi dùng cây quyết định đơn giản.

Nhận định tổng quát:

Việc lựa chọn max_depth hợp lý rất quan trọng. Quá nhỏ sẽ underfit, quá lớn sẽ overfit.

Với bộ dữ liệu Palmer Penguins, max_depth = 4 là tối ưu cho độ chính xác cao nhất trên tập kiểm tra.

- Bank:

max_depth	None	2	3	4	5	6	7
Accuracy	0.7980	0.7152	0.7730	0.7721	0.8070	0.8110	0.8101

- Nhận xét:

Dưới độ sâu nhỏ (max_depth = 2):

Cây quyết định quá nông nên không đủ khả năng phân chia dữ liệu phức tạp của bộ bank marketing, dẫn đến accuracy thấp (underfitting).

Độ sâu trung bình (max_depth = 3, 4):

Khi tăng độ sâu, mô hình học được nhiều đặc trưng hơn, accuracy cải thiện rõ rệt.

Độ sâu lớn (max_depth ≥ 5):

Khi max_depth tăng lên 5, 6, 7, accuracy tiếp tục tăng nhẹ và đạt mức cao nhất, sau đó hầu như không thay đổi (tăng rất ít hoặc giảm không đáng kể). Điều này cho thấy ở mức độ sâu này mô hình đã đủ phức tạp để mô tả tốt dữ liệu, nhưng chưa bị overfit rõ rệt.

Không có dấu hiệu overfitting mạnh:

Không giống như một số bộ dữ liệu khác, với bank marketing, khi tăng max_depth lên cao, accuracy trên tập test vẫn giữ ổn định hoặc tăng nhẹ, cho thấy dữ liệu có thể có nhiều đặc trưng quan trọng cần quyết định phân chia phức tạp hơn.

7.3. Ưu, nhược điểm mô hình

- Ưu điểm: Dễ hiểu, trực quan hóa tốt, xử lý được dữ liệu số lẫn phân loại, không cần chuẩn hóa dữ liệu số.
- Nhược điểm: Dễ overfit, không ổn định khi dữ liệu biến động lớn, kém hiệu quả khi có nhiều giá trị thiếu hoặc nhiễu.

8. SO SÁNH KẾT QUẢ GIỮA CÁC BỘ DỮ LIỆU 80/20

Dataset	Số lớp	Số đặc trưng	Số mẫu	Accuracy cao nhất	Nhận xét nổi bật
Heart Disease	2	13	303	82%	Dữ liệu hơi mất cân bằng, có giá trị thiếu
Penguins	3	5	344	97%	Dữ liệu sạch
Bank	2	16	11162	81%	Dữ liệu sạch, ít bị overfitting.

- Bộ dữ liệu nhiều mẫu, phân bố nhãn cân bằng cho kết quả tốt hơn.
- Mô hình cây quyết định phù hợp cả bài toán nhị phân và đa lớp, nhưng nhạy cảm với dữ liệu thiếu và đặc trưng nhiễu.

9. KẾT LUẬN

- Xử lý dữ liệu thiếu và biến đổi đặc trưng phân loại là rất quan trọng.
- Chia tập satisfy giúp đánh giá công bằng.
- Độ sâu cây cần được tối ưu để tránh overfit.
- Cây quyết định có tính giải thích cao, phù hợp với các bài toán cần minh bạch.
- Đặc trưng bộ dữ liệu (số mẫu, cân bằng nhãn, số đặc trưng) ảnh hưởng mạnh đến kết quả mô hình.

TÀI LIỆU THAM KHẢO

[1] UCI Heart Disease: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Palmer Penguins: https://github.com/allisonhorst/palmerpenguins

[3] scikit-learn documentation: https://scikit-learn.org/

[4] Bank marketing campaigns dataset | Opening Deposit: https://www.kaggle.com/datasets/volodymyrgavrysh/bank-marketing-campaigns-dataset

[5] https://github.com/RussH-code/DecisionTree-Heart-Disease-Classifier

[6] https://github.com/ashutoshmakone/Bank-Marketing-Dataset-Machine-Learning/tree/main