

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN KHOA CÔNG NGHỆ THÔNG TIN**

**Đồ án 3**

**DECISION TREE**

**Nhóm sinh viên thực hiện:**

Lê Gia Bảo **MSSV**:23127325

Vũ Anh **MSSV**:23127321

Hồ Gia Huy **MSSV**:23127376

Nguyễn Phan Thế Vinh

**MSSV**:23127520

**Môn:** Cơ Sở Trí Tuệ Nhân Tạo

**Năm học:** 2024-2025

**TP.HCM, tháng 4 năm 2025**

**Mục Lục**

[I. Bảng phân công 2](#_Toc195988420)

[II. Phân tích các tập dữ liệu 4](#_Toc195988421)

[III. So sánh giữa các bộ dữ liệu và kết luận 10](#_Toc195988422)

[IV. References 11](#_Toc195988423)

### Bảng phân công

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yêu cầu | Phân công | Tiến độ |
| Analysis of the Wine Quality dataset. | Data preparation. | Nguyễn Phan Thế Vinh | Hoàn thành |
|  | Implement decision tree classifiers. | Vũ Anh | Hoàn thành |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. | Vũ Anh | Hoàn thành |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. | Vũ Anh | Hoàn thành |
| Analysis of the Breast Cancer dataset. | Data preparation. | Nguyễn Phan Thế Vinh | Hoàn thành |
|  | Implement decision tree classifiers. | Lê Gia Bảo | Hoàn thành |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. | Lê Gia Bảo | Hoàn thành |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. | Lê Gia Bảo | Hoàn thành |
| Analysis of bank dataset. | Data preparation. | Nguyễn Phan Thế Vinh | Hoàn thành |
|  | Implement decision tree classifiers. | Hồ Gia Huy | Hoàn thành |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. | Hồ Gia Huy | Hoàn thành |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. | Hồ Gia Huy | Hoàn thành |
|  | Comparative analysis of all three datasets. | Hồ Gia Huy | Hoàn thành |
|  | Well-structured and formatted notebooks. |  | Hoàn thành |
|  | Report | Vũ Anh | Hoàn thành |

### Phân tích các tập dữ liệu

Tỷ lệ phân phối lớp cho tập dữ liệu gốc, tập huấn luyện và tập kiểm tra được tính bằng value\_counts(normalize=True) \* 100 trong notebook. Những điều này xác minh rằng phân chia phân tầng bảo toàn tỷ lệ lớp trên tất cả các phân chia.

* 1. **Breast Cancer Wisconsin (Diagnostic) Dataset (Binary Class Dataset):**
* Tổng quan về bộ dữ liệu :

Kích thước: 569 mẫu.

Các lớp: Nhị phân (0: Lành tính, 1: Ác tính).

Tính năng: 30.

* Phân phối lớp ban đầu (giả thuyết, dựa trên số liệu thống kê tập dữ liệu điển hình):

Lành tính (0): 62,74%

Ác tính (1): 37,26%

* Phân bố lớp học trên các phân vùng :
* Phân chia 40/60 (Train: 227 mẫu, Test: 342 mẫu):
* Train: Lành tính (0): 62,73%, Ác tính (1): 37,27%
* Test: Lành tính (0): 62,74%, Ác tính (1): 37,26%
* Phân chia 60/40 (Train: 341 mẫu, Test: 228 mẫu):
* Train: Lành tính (0): 62,74%, Ác tính (1): 37,26%
* Test: Lành tính (0): 62,74%, Ác tính (1): 37,26%
* Phân chia 80/20 (Train: 455 mẫu, Test: 114 mẫu):
* Train: Lành tính (0): 62,74%, Ác tính (1): 37,26%
* Test: Lành tính (0): 62,73%, Ác tính (1): 37,27%
* Phân chia 90/10 (Train: 512 mẫu, Test: 57 mẫu):
* Train: Lành tính (0): 62,74%, Ác tính (1): 37,26%
* Test: Lành tính (0): 63,16%, Ác tính (1): 36,84%
* Nhận xét:

 Tỷ lệ lớp gần như giống hệt nhau trên tất cả các lần chia, với độ lệch nhỏ (ví dụ: ±0,4% trong bộ kiểm tra 90/10) do quy mô bộ kiểm tra nhỏ (57 mẫu).

 Phân tầng đảm bảo cả hai lớp đều được biểu diễn tốt, ngay cả trong tập kiểm tra nhỏ nhất.

* Huấn luyện cây:

Sử dụng thư viện Graphviz để trực quan cây

Tỷ lệ 60/40

* Đánh giá mô hình

Tỷ lệ 60/40:

* Classification report

A screenshot of a computer

AI-generated content may be incorrect.

* Nhận xét:

+ Accuracy đạt 95% cho thấy mô hình có khả năng phân loại rất tốt

+ Precision cao (95%-96%) thể hiện mô hình ít dự đoán sai dương tính -> giảm báo động nhầm

+ Recall cao (98% cho Benign, 91% cho Malignant) là tín hiệu tích cực. Tuy nhiên, 14 mẫu u ác bị bỏ sót là điều cần lưu ý vì có thể dẫn đến chẩn đoán sai

+ F1-score đạt 0.96 và 0.93 chứng tỏ mô hình cân bằng tốt giữa precision và recall

+ Kết quả cho thấy Decision Tree có thể là lựa chọn khả thi trong phân tích sơ bộ, tuy nhiên để dùng trong thực tế y tế, cần cải thiện độ nhạy với u ác (recall của Malignant)

* Bảng so sánh các tỷ lệ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tỷ lệ Train/Test** | **Acccurary** | **Precision** | **Recall** | **F1-score** |
| 40/60 | 0.93 | 0.93 | 0.93 | 0.93 |
| 60/40 | 0.95 | 0.95 | 0.95 | 0.95 |
| 80/20 | 0.93 | 0.93 | 0.93 | 0.93 |
| 90/10 | 0.93 | 0.93 | 0.93 | 0.93 |

* Ảnh hưởng của độ sâu cây

A screenshot of a computer

AI-generated content may be incorrect.

* + - Nhận xét:

+ Việc điều chỉnh Max\_depth ảnh hưởng rõ rệt đến hiệu suất mô hình

+ Mô hình quá nông (depth 2-3) có xu hướng underfitting, còn mô hình quá sâu có thể gây overfitting nếu không kiểm soát tốt

+ Độ sâu tối trong trường hợp này là 7, vừa đảm bảo độ chính xác cao , vừa tránh cây quá phức tạp.

* 1. **Wine Quality Dataset (Multi-class Dataset)**
* Tổng quan về bộ dữ liệu :

Quy mô: 6.497 mẫu (rượu vang đỏ và trắng kết hợp).

Các lớp: Nhiều lớp (0: Low, 1: Standard, 2: High, dựa trên điểm chất lượng được phân loại thành 0–4, 5–6, 7–10).

Tính năng: 12 (bao gồm wine\_type ).

* Phân bổ lớp ban đầu (dựa trên kết quả đầu ra của sổ ghi chép ban đầu):

Low (0): 3,79% (246 mẫu)

Tiêu Standard (1): 76,53% (4.974 mẫu)

High (2): 19,68% (1.277 mẫu)

* Phân bố lớp học trên các phân vùng :
* Phân chia 40/60 (Train: 2.599 mẫu, Test: 3.898 mẫu):
* Train: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Test: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Phân chia 60/40 (Train: 3.898 mẫu, Test: 2.599 mẫu):
* Train: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Test: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Phân chia 80/20 (Train: 5.198 mẫu, Test: 1.299 mẫu):
* Train: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Test: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Phân chia 90/10 (Train: 5.847 mẫu, Test: 650 mẫu):
* Train: Low (0): 3,79%, Standard (1): 76,53%, High (2): 19,68%
* Test: Low (0): 3,84%, Standard (1): 76,46%, High (2): 19,70%
* Nhận xét :
* Tỷ lệ lớp rất nhất quán trên tất cả các phần chia, với độ lệch ±0,05% trong bộ thử nghiệm 90/10 do kích thước nhỏ hơn (650 mẫu).
* Lớp Low (3,79%) không được đại diện đầy đủ, nhưng sự phân tầng đảm bảo sự hiện diện của lớp này trong tất cả các tập hợp (ví dụ: ~25 mẫu trong tập thử nghiệm 90/10).
* Đánh giá mô hình

Tỷ lệ 90/10:

* Classification report

A screenshot of a computer screen

AI-generated content may be incorrect.

* Nhận xét:

+ Mô hình đạt accurary 80% - cao nhất trong các tỷ lệ đã thử

+ Lớp Standard vẫn là lớp mô hình hoạt động tốt nhất (F1 = 0.88), do chiếm phần lớn dữ liệu

+ Lớp High (F1 = 0.61) có cải thiện so với các tỷ lệ chia trước, cho thấy mô hình học được tốt hơn khi có nhiều dữ liệu

+ Lớp Low có kết quả thấp (F1 = 0.31) do số mẫu rất ít (chỉ 24 mẫu) -> dễ bị mô hình bỏ qua

+ Chênh lệch giữa macro avg và weighted avg phản ánh mất cân bằng nhãn trong dữ liệu

* Bảng so sánh các tỷ lệ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tỷ lệ Train/Test** | **Acccurary** | **Precision** | **Recall** | **F1-score** |
| 40/60 | 0.74 | 0.74 | 0.74 | 0.74 |
| 60/40 | 0.77 | 0.78 | 0.77 | 0.77 |
| 80/20 | 0.80 | 0.79 | 0.80 | 0.80 |
| 90/10 | 0.80 | 0.81 | 0.80 | 0.81 |

* Ảnh hưởng của độ sâu cây

A screenshot of a computer

AI-generated content may be incorrect.

* + - Nhận xét:

+ Khi không giới hạn độ sâu, mô hình đạt độ chính xác cao nhất 0.80

+ Khi giới hạn độ sâu nhỏ depth = 2-3, độ sâu giảm xuống còn 0.782 cho thất mô hình bị underfitting, quá đơn giản không đủ khả năng phân loại tốt các lớp

+ Từ độ sâu 4 – 6, accurary tăng nhẹ, cho thấy mô hình dần cải thiện hiệu suất khi được phép phân chia nhiều hơn

+ Depth = 7, accurary giảm xuống lại cho thấy mô hình bị overfitting khi cây học quá kỹ dữ liệu huấn luyện làm mất khả năng tổng quát hóa.

* 1. **Additional Dataset: Bank Marketing (Binary Class Dataset)**
* Tổng quan về bộ dữ liệu :

Kích thước: 4.521 mẫu.

Các lớp: Nhị phân (0: no, 1: yes, dành cho đăng ký gửi tiết kiệm có kỳ hạn).

Tính năng: 16.

Phân phối lớp ban đầu (giả thuyết, dựa trên số liệu thống kê tập dữ liệu điển hình):

no (0): 88,23%

yes (1): 11,77%

* Phân bố lớp học trên các phân vùng :
* Phân chia 40/60 (Train: 1.808 mẫu, Test: 2.713 mẫu):
* Train: no (0): 88,23%, yes (1): 11,77%
* Test: no (0): 88,23%, yes (1): 11,77%
* Phân chia 60/40 (Train: 2.713 mẫu, Test: 1.808 mẫu):
* Train: no (0): 88,23%, yes (1): 11,77%
* Test: no (0): 88,23%, yes (1): 11,77%
* Phân chia 80/20 (Train: 3.617 mẫu, Test: 904 mẫu):
* Train: no (0): 88,23%, yes (1): 11,77%
* Test: no (0): 88,23%, yes (1): 11,77%
* Phân chia 90/10 (Train: 4.069 mẫu, Test: 452 mẫu):
* Train: no (0): 88,23%, yes (1): 11,77%
* Test: no (0): 88,22%, yes (1): 11,78%
* Nhận xét:
* Tỷ lệ lớp học hầu như giống hệt nhau ở tất cả các lần chia, với độ lệch no đáng kể (±0,01% trong bộ Test 90/10).
* Lớp Yes (11,77%) là lớp thiểu số, nhưng phân tầng đảm bảo tính đại diện của lớp này (ví dụ: ~53 mẫu trong tập Test 90/10)
* Đánh giá mô hình

Tỷ lệ 60/40:

* Classification report

A screenshot of a computer screen

AI-generated content may be incorrect.

* Nhận xét:

+ Mô hình đạt accurary 89%, nhưng con số này che giấu hiệu suất rất thấp ở lớp thiểu số (lớp 1 – khách hàng đăng kí gửi tiết kiệm)

+ Precision và Recall của lớp 1 lần lượt chỉ đạt 0.52 và 0.26, dẫn đến F1-score chỉ 0.35 -> cho thấy mô hình gần như bỏ qua phần lớn khách hàng tiềm năng thực sự

+ Lớp 0 được phân loại cực kỳ tốt (F1 = 0.94, recall = 0.97) vì đây là lớp chiếm đa số (gần 89% dữ liệu)

+ Macro avg thấp hơn nhiều so với weighted avg -> mô hình mất cân bằng, chủ yếu học theo xu hướng của lớp chiếm ưu thế

* Bảng so sánh các tỷ lệ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tỷ lệ Train/Test** | **Acccurary** | **Precision** | **Recall** | **F1-score** |
| 40/60 | 0.89 | 0.86 | 0.89 | 0.85 |
| 60/40 | 0.89 | 0.86 | 0.89 | 0.87 |
| 80/20 | 0.88 | 0.85 | 0.88 | 0.86 |
| 90/10 | 0.87 | 0.84 | 0.87 | 0.85 |

* Ảnh hưởng của độ sâu cây

A screen shot of a number

AI-generated content may be incorrect.

* + - Nhận xét:

+ Khi không giới hạn độ sâu, mô hình đạt accurary thấp nhất cho thấy cây bị overfitting

+ Khi tăng độ sâu từ 2 – 6, độ chính xác tăng đều , đạt đỉnh tại depth = 6, điều này cho thấp việc giới hạn độ sâu hợp lý giúp cải thiện hiệu suất, tránh tình trạng mô hình quá phức tạp mà không cần thiết

+ Tại depth = 7, accurary giảm nhẹ -> giấu hiện ban đầu của overfitting khi cây bắt đầu học theo nhiễu của dữ liệu

### So sánh giữa các bộ dữ liệu và kết luận

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Bộ dữ liệu** | **Số lớp** | **Số mẫu** | **Acccurary** | **Precision** | **Recall** | **F1-score** |
| Breast Cancer | 2 | 569 | 0.95 | 0.95 | 0.94 | 0.95 |
| Wine | 3 (Low/Std/High) | 4898 | 0.77 | 0.78 | 0.77 | 0.77 |
| Bank | 2 | 4521 | 0.89 | 0.86 | 9.89 | 0.87 |

* Nhận xét:

+ Breast Cancer là bộ dữ liệu dễ phân loại và cân bằng -> đạt độ chính xác và F1-score cao nhất (~0.95). Do đặc trung phân lớp rõ ràng và số lượng lớp chỉ là 2.

+ Wine Quality có 3 lớp với sự mất cân bằng nhẹ -> mô hình gặp khó khăn hơn, accurary chỉ đạt ~0.77, thấp nhất trong 3 bộ, lớp Lơ và High bị dự đoán sai nhiều.

+ Bank bị mất cân bằng nghiêm trọng(~88%:12%), nhưng mô hình vẫn đạt accurary 89% và F1-score 0.87, tuy nhiên chủ yếu do đoan đúng lớp chiếm đa số. Khi xét riêng lớp thiểu số (khách đăng kí), recall rất thấp -> accurary cao không đảm bảo mô hình tốt nếu dữ liệu mất cân bằng, cần đánh giá toàn diện bằng F1-score và recall từng lớp

* Kết luận:

+ Mô hình Dicision Tree phù hợp với các bài toán có dữ liệu cân bằng và phân biệt đặc trung rõ ràng như Breast Cancer

+ Đối với bài toán có nhiều lớp hoặc mất cân bằng, cần kết hợp các phương pháp khác như:

* Xử lý mất cân bằng dữ liệu
* Tối ưu hyperparameters(max\_depth, min\_samples\_leaf)
* Thử nghiệm mô hình mạnh hơn như Random Forest hoặc Gradient Boosting

### References

[Tài liệu Scikit-learn](https://scikit-learn.org/%0c)