

**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

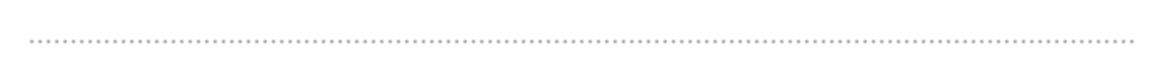
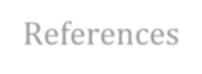
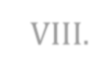
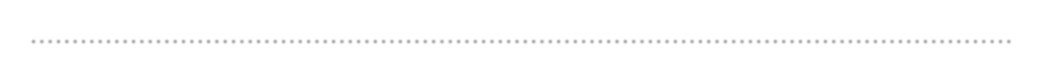
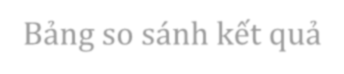
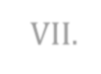
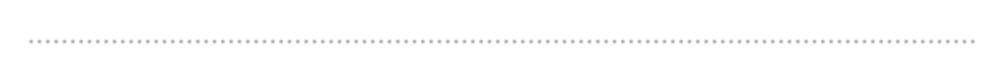
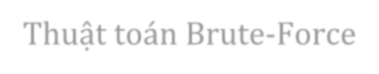
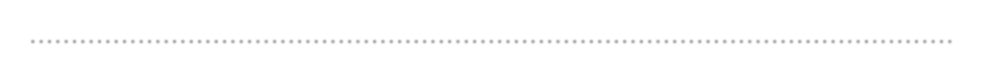
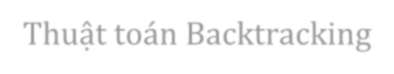
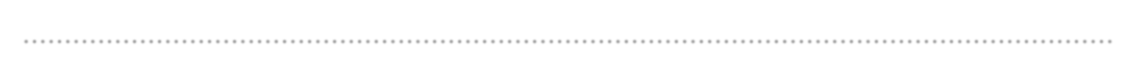
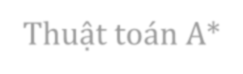
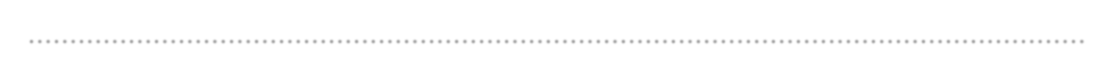
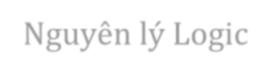
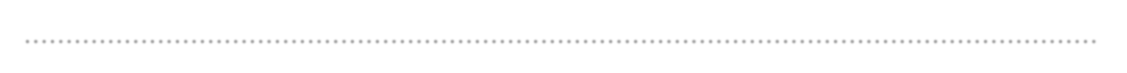
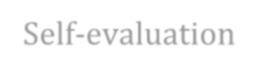
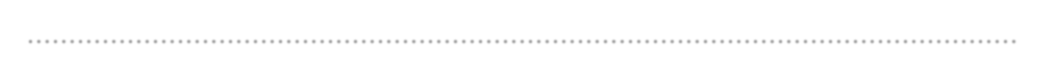
**MSSV**:23127520

**Môn:** Cơ Sở Trí Tuệ Nhân Tạo

**Năm học:** 2024-2025

**TP.HCM, tháng 4 năm 2025**

1. **Bảng phân công**



. [2](#_bookmark0)

1. Self-evaluation [2](#_bookmark1)
2. Nguyên lý Logic [4](#_bookmark2)
3. Thuật toán A\* [5](#_bookmark3)
4. Thuật toán Backtracking [7](#_bookmark4)
5. Thuật toán Brute-Force [8](#_bookmark5)
6. Bảng so sánh kết quả [8](#_bookmark6)
7. References [9](#_bookmark7)

|  |  |
| --- | --- |
| **Thành Viên** | **Công việc** |
| **Vũ Anh** | Tham gia làm CNF, viết báo cáo, so sánh thời gian  chạy, hàm ghi output ra file. |
| **Lê Gia Bảo** | Trình bày nguyên lí logic sinh ra các mệnh đề CNF,  viết chương trình tự động sinh CNF |
| **Hồ Gia Huy** | Thuật toán A\*, Backtracking, giải các mệnh đề CNF  bằng pySAT, hàm đọc input từ file |
| **Nguyễn Phan Thế**  **Vinh** | Thuật toán Brute Force, làm các file input, quay  video |

1. **Self-evaluation**
   1. Mức độ hoàn thành:
      * Các thuật toán hoạt động đúng với yêu cầu đồ án đưa ra.
      * Có đầy đủ các test case của các trường hợp có thể xảy ra.
      * Link hoàn thành đồ án:

<https://github.com/Vincenthevinh/Hashiwokakero.git>

* + - Video hướng dẫn đồ án: <https://youtu.be/3-7EKNwq8KE>

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yêu cầu | Phân công | Tiến độ |
| Analysis of the Wine Quality dataset. | Data preparation. |  | Hoàn thành |
|  | Implement decision tree classifiers. |  |  |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. |  |  |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. |  |  |
| Analysis of the Breast Cancer dataset. | Data preparation. |  | Hoàn thành |
|  | Implement decision tree classifiers. |  |  |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. |  |  |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. |  |  |  |
| Analysis of an additional dataset. | Data preparation. |  | Hoàn thành |  |
|  | Implement decision tree classifiers. |  |  |  |
|  | Performance evaluation of decision tree:  - Classification report and confusion matrix.  - Insights. |  |  |  |
|  | Depth and accuracy of decision trees:  - Visualization (trees, tables, charts).  - Insights. |  |  |  |
|  | Comparative analysis of all three datasets. |  |  |  |
|  | Well-structured and formatted notebooks. |  |  |  |

* 1. Khó khăn:

Machine learning models không thể xử lý chuỗi trực tiếp → cần mã hóa thành số:

Tạo một dictionary tên là le\_dict, trong đó:

* Key là tên của từng cột phân loại.
* Value là một đối tượng LabelEncoder() ứng với cột đó.

1. **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.

* 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).
  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)