|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN**  **CHUYÊN ĐỀ CHUYÊN SÂU KỸ THUẬT**  **CÔNG NGHỆ THÔNG TIN 1**  **PHÂN TÍCH VÀ DỰ ĐOÁN HÀNH VI KHÁCH HÀNG TRONG MARKETING**  **Giảng viên hướng dẫn : ThS. VƯƠNG XUÂN CHÍ**  **Sinh viên thực hiện : TRẦN HUỲNH TRỌNG**  **MSSV : 2200002205**  **Khoá : 2022-2026**  **Ngành : CÔNG NGHỆ THÔNG TIN**  **Chuyên ngành : KỸ THUẬT CÔNG NGHỆ THÔNG TIN**  **Tp.HCM, tháng 11 năm 2024**  **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN**  **CHUYÊN ĐỀ CHUYÊN SÂU KỸ THUẬT**  **CÔNG NGHỆ THÔNG TIN 1**  **PHÂN TÍCH VÀ DỰ ĐOÁN HÀNH VI KHÁCH HÀNG TRONG MARKETING**  **Giảng viên hướng dẫn : ThS. VƯƠNG XUÂN CHÍ**  **Sinh viên thực hiện : TRẦN HUỲNH TRỌNG**  **MSSV : 2200002205**  **Khoá : 2022-2026**  **Ngành : CÔNG NGHỆ THÔNG TIN**  **Chuyên ngành : KỸ THUẬT CÔNG NGHỆ THÔNG TIN**  **Tp.HCM, tháng 11 năm 2024** |

**MỤC LỤC**

**Lời cảm ơn i**

**Lời giới thiệu: ii**

**Chương 1: LÝ THUYẾT KHAI PHÁ DỮ LIỆU................................................... 12**

**1.1 Mô tả các thuộc tính của bộ dữ liệu................................................................ 12**

1.1.1 Các thuộc tính định lượng: 12

1.1.2 Các thuộc tính định tính: 13

1.1.3 Thuộc tính mục tiêu 14

**1.2 Chuẩn bị dữ liệu 14**

1.2.1 Tải dữ liệu 15

1.2.2 Phân tích các trường dữ liệu 15

1.2.3 Loại bỏ các giá trị thiếu 16

1.2.4 Xử lý ngoại lệ 16

1.2.5 Chuẩn hóa dữ liệu 17

**1.3 Tiền sử xử lý dữ liệu 17**

1.3.1 Chuẩn hóa dữ liệu 18

1.3.2 Mã hóa các thuộc tính phân loại 19

1.3.3 Xử lý dữ liệu thiếu 20

**1.4 Cơ sở lý thuyết 21**

**Chương 2: PHÂN TÍCH DỮ LIỆU** 30

**2.1 Mô tả mục tiêu** 23

**2.2 Chuẩn bị và tiền xử lý dữ liệu** 23

2.2.1 Tải và kiểm tra dữ liệu 24

2.2.2 Loại bỏ giá trị thiếu hoặc ngoại lệ 24

2.2.3 Chuẩn hóa dữ liệu 24

2.2.4 Mã hóa các thuộc tính phân loại 24

2.2.5 Tách tập dữ liệu 25

2.2.6 Kiểm tra lại dữ liệu đã xử lý 25

**2.3 Ứng dụng Weka** 25

2.3.1 Nhập dữ liệu vào Weka 25

2.3.2 Tiền xử lý dữ liệu trong Weka 26

2.3.3 Chạy thuật toán học máy 26

2.3.4 Phân tích kết quả 26

**2.4. Mô tả thuật toán**

2.4.1 Decision Trees (Cây quyết định) 27

2.4.2 Random Forest (Rừng ngẫu nhiên) 27

2.4.3 Naive Bayes 28

**Chương 3: ÁP DỤNG KẾT HỢP PHÂN LOẠI** 29

**3.1 Giới thiệu mô hình 29**

**3.2 Mô tả chức năng 30**

3.2.1 **Cách mô hình dự đoán hoặc phân loại dữ liệu** 30

3.2.2 **Mô hình áp dụng vào bộ dữ liệu telecom marketing** 31

**Chương 4: KẾT LUẬN** 20

4.1 Kết luận 44

4.2 Hướng phát triển 44

4.2.1 Cân bằng dữ liệu 44

4.2.2 Thử nghiệm với các thuật toán khác 45

4.2.3 Kiểm tra trên tập dữ liệu khác 45

**PHỤ LỤC 122**

**DANH MỤC TÀI LIỆU TAM KHẢO 127**

**MỤC LỤC HÌNH ẢNH**

[Hình 1.1. Mô Tả 12](#_Toc21041)

[Hình 1.2. Weka 15](#_Toc11891)

[Hình 1.3. Ảnh Cấu trúc dữ liệu 15](#_Toc4019)

[Hình 1.4. Tiền xử lý dữ liệu 18](#_Toc1997)

[Hình 1.5. K-Nearest Neighbors (KNN) 18](#_Toc4637)

[Hình 1.6. Min Max Scolling 19](#_Toc2246)

[Hình 1.7. Z-score Standardization 19](#_Toc23469)

[Hình 1.8.Label Encoding 20](#_Toc5956)

[Hình 1.9 One-Hot-Encoding 20](#_Toc31947)

[Hình 1.10. Thuật toán KNN Imputation 21](#_Toc31947)

[Hình 1.11. Ảnh Weka 21](#_Toc31947)

[Hình 1.12. Ảnh Excel 22](#_Toc31947)

[Hình 2.0. Phân loại khách hàng tiềm năng 23](#_Toc9151)

[Hình 2.1. Weka 25](#_Toc1786)

[Hình 2.2. Naive Bayes 28](#_Toc30960)

[Hình 3.1. Decision Tree 29](#_Toc25086)

[Hình 3.2. Chọn Explorer 31](#_Toc24396)

[Hình 3.3. Preprocess và nhấn Open file 32](#_Toc8977)

[Hình 3.4. Chọn tập dữ liệu 33](#_Toc18406)

[Hình 3.5 Hiển thị các thuộc tính 33](#_Toc24430)

[Hình 3.6. Chuyển sang tab Classify 34](#_Toc17264)

[Hình 3.7. Chọn thuật toán J48 (Decision Tree) 35](#_Toc8589)

[Hình 3.8. Chọn thuộc tính y 35](#_Toc2812)

[Hình 3.9. Kết quả thuật toán J48 36](#_Toc7369)

[Hình 3.10. Kết quả thuật toán J48 37](#_Toc12485)

[Hình 3.11. Điều chỉnh tham số 39](#_Toc15765)

[Hình 3.12. Kết quả 40](#_Toc3596)

[Hình 3.13. Kết quả thuật toán Random Forest 41](#_Toc1119)

**KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| Chữ viết tắt | Ý nghĩa |
| CNTT | Công nghệ thông tin |
| NTTU | Nguyễn Tất Thành University |
|  |  |

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 1 NĂM HỌC 2024 - 2025** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Chuyên đề chuyên sâu CNTT 1 Lớp học phần: 22DTH1B

Nhóm sinh viên thực hiện: Nhóm 2

1. Nguyễn Hoàng Khá Tham gia đóng góp: 100%

2. Đỗ Minh Thắng Tham gia đóng góp: 100%

3. Lê Văn An Tham gia đóng góp: 100%

4. Trần Huỳnh Trọng Tham gia đóng góp: 100%

5. Bùi Đại Bill Tham gia đóng góp: 100%

Ngày thi: 03/01/2025 Phòng thi: : L.512

Đề tài tiểu luận/báo cáo của sinh viên: PHÂN TÍCH VÀ DỰ ĐOÁN HÀNH VI KHÁCH HÀNG TRONG MARKETING

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**LỜI CẢM ƠN**

Trước tiên, em xin gửi lời cảm ơn chân thành đến Thầy Vương Xuân Chí người đã tận tình chỉ bảo, hướng dẫn và hỗ trợ em trong suốt quá trình thực hiện tiểu luận chuyên đề này. Sự tận tâm, nhiệt tình và những ý kiến đóng góp quý báu của thầy đã giúp em hoàn thành bài tiểu luận một cách tốt nhất.

Em cũng xin bày tỏ lòng biết ơn đến Khoa Công nghệ thông tin cùng các thầy cô giáo trong Trường đại học Nguyễn Tất Thành, những người đã truyền đạt kiến thức và tạo điều kiện thuận lợi để em thực hiện tiểu luận này.

Bên cạnh đó, em xin gửi lời cảm ơn đến bạn bè, đồng nghiệp và gia đình – những người luôn động viên, hỗ trợ tinh thần cũng như chia sẻ những khó khăn trong suốt thời gian nghiên cứu và hoàn thiện chuyên đề.

Cuối cùng, mặc dù đã cố gắng hết mình, bài tiểu luận vẫn không thể tránh khỏi những thiếu sót. Em mong nhận được những ý kiến đóng góp từ thầy cô và các bạn để hoàn thiện hơn trong tương lai.

Xin chân thành cảm ơn!

Trần Huỳnh Trọng

**GIỚI THIỆU**

Trong bối cảnh cạnh tranh ngày càng gay gắt của thị trường hiện nay, việc áp dụng công nghệ vào các chiến lược marketing đã trở thành một yếu tố không thể thiếu đối với các doanh nghiệp. Một trong những hình thức marketing truyền thống nhưng vẫn rất hiệu quả là tiếp thị qua điện thoại. Tiếp thị qua điện thoại không chỉ giúp các công ty kết nối trực tiếp với khách hàng mà còn là một công cụ mạnh mẽ để tiếp cận khách hàng tiềm năng, đặc biệt là trong các ngành tài chính, ngân hàng. Tuy nhiên, để tối ưu hóa hiệu quả của các chiến dịch tiếp thị qua điện thoại, các công ty cần phải hiểu rõ về hành vi và sở thích của khách hàng, từ đó đưa ra các chiến lược phù hợp. Để làm được điều này, việc phân tích và khai thác dữ liệu khách hàng là một bước đi vô cùng quan trọng.

Trong những năm gần đây, học máy (machine learning) đã trở thành một công cụ hữu ích trong việc phân tích và dự đoán hành vi người dùng, đặc biệt là trong lĩnh vực tiếp thị. Việc áp dụng học máy vào phân tích dữ liệu tiếp thị qua điện thoại sẽ giúp các công ty phân tích một lượng dữ liệu khổng lồ mà thủ công không thể xử lý được. Các mô hình học máy có thể giúp dự đoán khả năng khách hàng sẽ tham gia vào một chiến dịch tiếp thị, từ đó hỗ trợ các chiến lược marketing trở nên hiệu quả hơn. Hơn nữa, học máy còn giúp tự động hóa quá trình phân tích và tối ưu hóa các chiến dịch tiếp thị, giảm thiểu sự can thiệp của con người và nâng cao độ chính xác trong việc đưa ra các quyết định marketing.

Mục tiêu nghiên cứu của tiểu luận này là sử dụng các mô hình học máy để phân tích dữ liệu khách hàng trong các chiến dịch tiếp thị qua điện thoại của một ngân hàng. Cụ thể, em sẽ xây dựng và áp dụng các mô hình học máy để phân loại dữ liệu và dự đoán khả năng khách hàng tham gia vào các chiến dịch tiếp thị tài chính. Tiểu luận sẽ tập trung vào việc lựa chọn các mô hình học máy phù hợp với đặc điểm của bộ dữ liệu và đánh giá hiệu quả của từng mô hình trong việc phân loại khách hàng.

Bộ dữ liệu mà em sử dụng là bộ dữ liệu tiếp thị qua điện thoại của một ngân hàng Bồ Đào Nha, bao gồm 4521 mẫu dữ liệu với 17 thuộc tính, chẳng hạn như độ tuổi, nghề nghiệp, tình trạng hôn nhân, tình trạng vay mượn, mức thu nhập, lịch sử giao dịch, và thông tin về các chiến dịch tiếp thị trước đó. Mục tiêu là dự đoán khả năng khách hàng sẽ tham gia vào một chiến dịch tiếp thị tài chính, từ đó giúp ngân hàng tối ưu hóa các chiến dịch marketing và giảm thiểu chi phí. Các thuộc tính trong bộ dữ liệu này không chỉ cung cấp cái nhìn tổng thể về khách hàng mà còn giúp nhận diện các yếu tố quan trọng ảnh hưởng đến quyết định tham gia dịch vụ tài chính. Việc khai thác thông tin này sẽ giúp ngân hàng đưa ra các chiến lược marketing được cá nhân hóa hơn, phù hợp với từng đối tượng khách hàng.

Trong nghiên cứu này, em sẽ giải quyết vấn đề dự đoán kết quả chiến dịch tiếp thị qua điện thoại thông qua việc áp dụng các thuật toán học máy, bao gồm các mô hình phân loại như cây quyết định (Decision Tree), hồi quy logistic (Logistic Regression), máy vector hỗ trợ (Support Vector Machine - SVM), và các thuật toán học máy khác. Bằng cách áp dụng và so sánh hiệu quả của từng mô hình, nghiên cứu sẽ chỉ ra đâu là mô hình hiệu quả nhất trong việc phân loại khách hàng tham gia chiến dịch tiếp thị, đồng thời chỉ ra các yếu tố quan trọng nhất trong việc dự đoán quyết định của khách hàng.

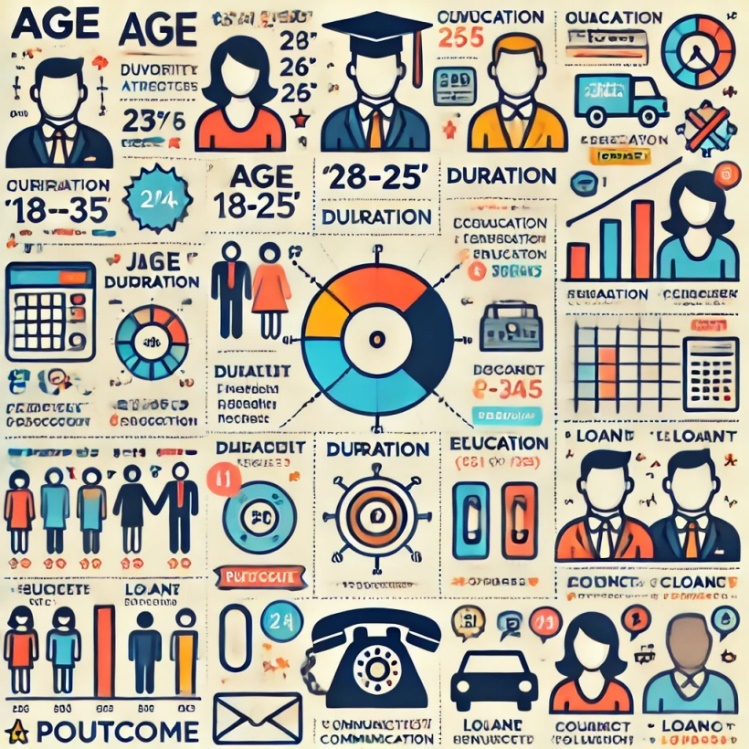
Việc nghiên cứu và áp dụng học máy trong phân tích dữ liệu tiếp thị qua điện thoại không chỉ có giá trị lý thuyết mà còn mang lại những ứng dụng thực tiễn đáng kể. Đối với các doanh nghiệp, việc sử dụng mô hình học máy để phân tích dữ liệu giúp tăng cường hiệu quả các chiến lược tiếp thị, tiết kiệm chi phí quảng cáo và tối ưu hóa tỷ lệ chuyển đổi khách hàng. Việc cải thiện chiến lược tiếp thị không chỉ giúp nâng cao hiệu quả kinh doanh mà còn nâng cao trải nghiệm khách hàng, tạo ra mối quan hệ lâu dài và bền vững giữa khách hàng và doanh nghiệp.

**CHƯƠNG I**

**LÝ THUYẾT KHAI PHÁ DỮ LIỆU**

**1.1 MÔ TẢ CÁC THUỘC TÍNH CỦA TRANG DỮ LIỆU**

Bộ dữ liệu tiếp thị qua điện thoại bao gồm nhiều thuộc tính quan trọng, được phân loại thành các dạng định lượng (số liệu) và định tính (danh mục). Dưới đây là phân tích chi tiết từng thuộc tính:



**Hình 1.1:** Mô tả các thuộc tính

**1.1.1Các thuộc tính định lượng**

* **age (tuổi):**
* Mô tả: Tuổi của khách hàng, dạng số nguyên.
* Ý nghĩa: Tuổi ảnh hưởng lớn đến hành vi tiêu dùng và quyết định tài chính.
* Phân tích: Có thể phân loại thành nhóm tuổi để dễ dàng phân tích, ví dụ: "18-25", "26-35", "36-50", "51+".
* **Balance (số dư tài khoản):**
* Mô tả: Số dư trung bình hàng năm của tài khoản ngân hàng, đơn vị là euro.
* Ý nghĩa: Phản ánh khả năng tài chính của khách hàng. Số dư cao có thể tăng khả năng chấp nhận sản phẩm.
* **duration (thời lượng cuộc gọi):**
* Mô tả: Thời gian của cuộc gọi tiếp thị, tính bằng giây.
* Ý nghĩa: Đây là một trong những thuộc tính quan trọng nhất vì thời lượng cuộc gọi càng dài, khả năng thành công càng cao.
* **campaign (số lần liên lạc):**
* Mô tả: Số lần liên lạc với khách hàng trong chiến dịch hiện tại.
* Ý nghĩa: Số lần liên lạc nhiều có thể ảnh hưởng đến sự chấp nhận hoặc từ chối sản phẩm.
* **pdays (số ngày kể từ lần liên lạc trước đó):**
* Mô tả: Số ngày kể từ khi khách hàng được liên lạc trong chiến dịch trước đó. Giá trị -1 chỉ ra rằng khách hàng chưa được liên lạc trước đó.
* Ý nghĩa: Giúp đánh giá tần suất liên lạc và mức độ tiếp cận của chiến dịch.
* **previous (số liên lạc trước đó):**
* Mô tả: Tổng số lần liên lạc trong các chiến dịch trước.
* Ý nghĩa: Chỉ ra mối quan hệ giữa khách hàng và ngân hàng trong quá khứ.

**1.1.2 Các thuộc tính định tính**

* **job (nghề nghiệp):**
* Mô tả: Nghề nghiệp hiện tại của khách hàng. Các giá trị bao gồm: "admin.", "technician", "services", "unemployed", v.v.
* Ý nghĩa: Nghề nghiệp liên quan mật thiết đến mức thu nhập và khả năng chi tiêu.
* **marital (tình trạng hôn nhân):**
* Mô tả: Tình trạng hôn nhân, bao gồm "single", "married", và "divorced".
* Ý nghĩa: Ảnh hưởng đến quyết định tài chính và thói quen chi tiêu của khách hàng.
* **education (trình độ học vấn):**
* Mô tả: Trình độ học vấn của khách hàng, các giá trị bao gồm: "primary", "secondary", "tertiary", "unknown".
* Ý nghĩa: Trình độ học vấn thường liên quan đến nhận thức tài chính và quyết định đầu tư.
* **default (nợ xấu):**
* Mô tả: Khách hàng có lịch sử nợ xấu hay không ("yes" hoặc "no").
* Ý nghĩa: Dự đoán khả năng chấp nhận sản phẩm dựa trên lịch sử tín dụng.
* **housing (vay thế chấp):**
* Mô tả: Khách hàng có khoản vay thế chấp hay không ("yes" hoặc "no").
* Ý nghĩa: Phản ánh khả năng tài chính và rủi ro tín dụng.
* **loan (vay cá nhân):**
* Mô tả: Khách hàng có khoản vay cá nhân hay không ("yes" hoặc "no").
* Ý nghĩa: Ảnh hưởng đến mức độ sẵn sàng đầu tư vào sản phẩm tài chính khác.
* **contact (phương thức liên lạc):**
* Mô tả: Phương thức liên lạc với khách hàng, gồm "cellular" (di động) và "telephone" (cố định).
* Ý nghĩa: Cho biết cách tiếp cận hiệu quả nhất với khách hàng.
* **poutcome (kết quả chiến dịch trước):**
* Mô tả: Kết quả của chiến dịch tiếp thị trước đó, gồm "success", "failure", "unknown".
* Ý nghĩa: Đánh giá hiệu quả của chiến dịch tiếp thị trước đó.

**1.1.3 Thuộc tính mục tiêu**

* **Y (kết quả):**

Mô tả: Biến mục tiêu cho biết khách hàng có đồng ý đăng ký sản phẩm tiết kiệm kỳ hạn hay không ("yes" hoặc "no").

**1.2 CHUẨN BỊ DỮ LIỆU**

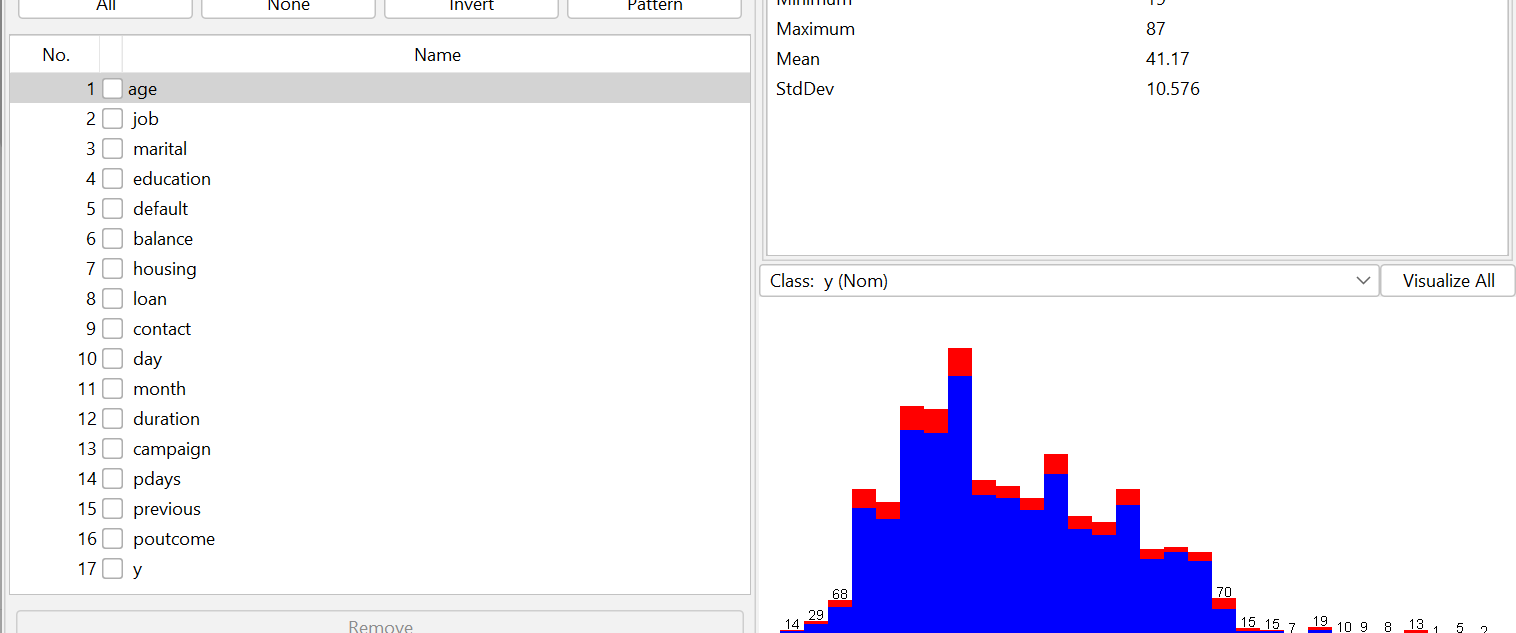
**1.2.1 Tải dữ liệu**

* **Mục tiêu:** Tải dữ liệu từ nguồn gốc đã xác định, đảm bảo dữ liệu được nhập vào hệ thống một cách đầy đủ và chính xác.
* **Thực hiện:**
  + Sử dụng các công cụ như Python (pandas) hoặc Weka để tải dữ liệu từ tệp CSV, Excel, hoặc cơ sở dữ liệu SQL.



**Hình 1.2** Ảnh Weka

* + Kiểm tra định dạng và cấu trúc dữ liệu (ví dụ: số lượng cột, kiểu dữ liệu của từng cột).



**Hình 1.3** Ảnh cấu trúc dữ liệu

**1.2.2 Phân tích các trường dữ liệu**

* **Mục tiêu:** Hiểu rõ nội dung và ý nghĩa của từng thuộc tính trong bộ dữ liệu.
* **Thực hiện:**
  + Kiểm tra lại dữ liệu
  + Phân loại thành dữ liệu định lượng (số học) và định tính (danh mục).Xác định kiểu dữ liệu (int, float, string).
* **Thống kê mô tả:**
  + Tính toán các thông số như giá trị trung bình, trung vị, phương sai, và độ lệch chuẩn cho dữ liệu định lượng.
  + Đếm tần suất xuất hiện của các giá trị trong dữ liệu định tính.
* **Phát hiện bất thường:**

Tìm kiếm các giá trị không hợp lệ, như số âm trong cột "age" hoặc giá trị "unknown" trong cột "job".

**1.2.3 Loại bỏ giá trị thiếu**

* **Mục tiêu:** Xử lý các giá trị thiếu để tránh ảnh hưởng đến hiệu quả của mô hình.
* **Thực hiện:**

**Xác định giá trị thiếu:**

* + Sử dụng các công cụ kiểm tra các giá trị NULL hoặc các giá trị đặc biệt như "unknown".
* **Phương pháp xử lý:**
  + Loại bỏ các bản ghi có quá nhiều giá trị thiếu.
  + Điền giá trị thay thế, như:
  + Giá trị trung bình (cho dữ liệu định lượng).
  + Giá trị phổ biến nhất (cho dữ liệu định tính).

**1.2.4 Xử lý ngoại lệ**

* **Mục tiêu:**Xử lý các giá trị bất thường có thể làm lệch kết quả phân tích hoặc mô hình.
* **Thực hiện:**

**Xác định ngoại lệ:**

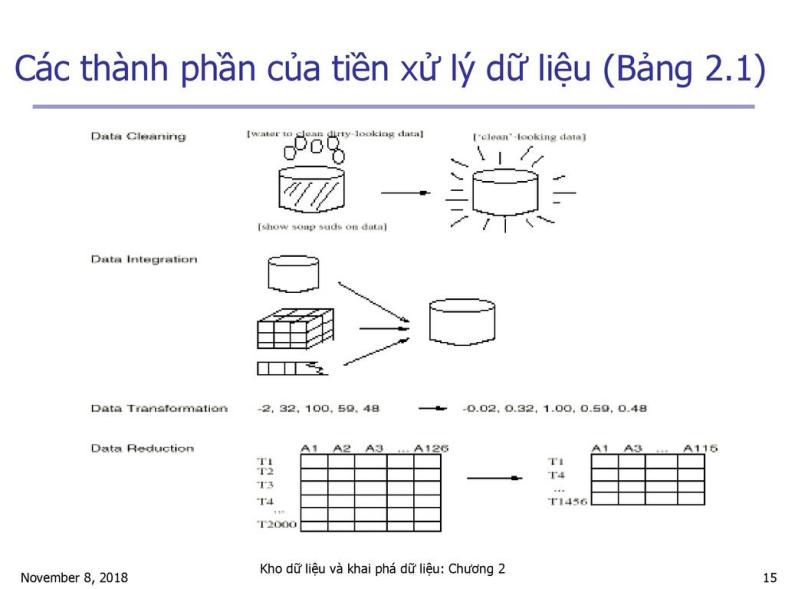
* Sử dụng biểu đồ (boxplot, histogram) để tìm các điểm bất thường.
* Kiểm tra các giá trị nằm ngoài khoảng [Q1 - 1.5IQR, Q3 + 1.5IQR] (IQR: khoảng cách giữa tứ phân vị).
* **Phương pháp xử lý:**
  + Loại bỏ các giá trị ngoại lệ (nếu chúng là lỗi nhập liệu).
  + Chuyển đổi hoặc chuẩn hóa giá trị ngoại lệ để giảm tác động tiêu cực.

**1.2.5 Chuẩn hóa dữ liệu**

* **Mục tiêu:**
  + Đưa dữ liệu về cùng thang đo để sử dụng trong các thuật toán phân tích và học máy.
* **Thực hiện:**
  + Chuẩn hóa giá trị định lượng bằng cách sử dụng Min-Max Scaling hoặc Standardization (Z-score).
  + Mã hóa dữ liệu định tính thành dạng số liệu (One-hot encoding hoặc Label encoding).

**1.3 TIỀN XỬ LÝ DỮ LIỆU**

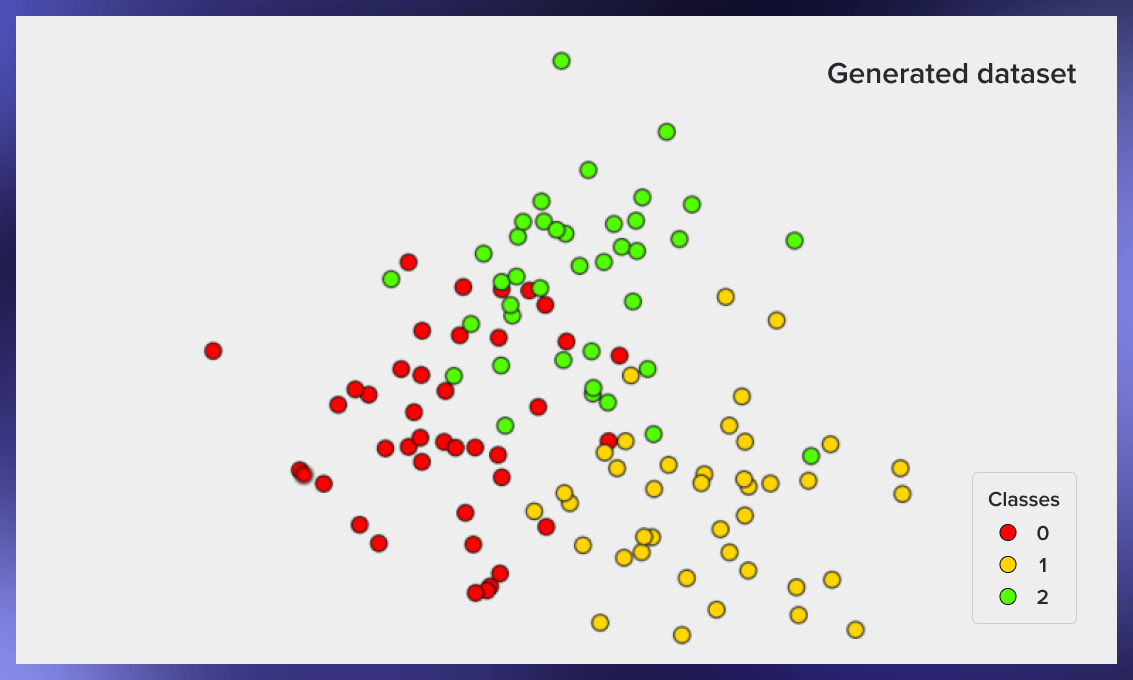
Tiền xử lý dữ liệu là một bước không thể thiếu trong quá trình khai phá dữ liệu và xây dựng mô hình học máy. Mục tiêu của giai đoạn này là biến đổi dữ liệu thô thành dạng dễ xử lý và phù hợp với các thuật toán. Các kỹ thuật phổ biến trong tiền xử lý bao gồm chuẩn hóa dữ liệu, mã hóa các thuộc tính phân loại, và xử lý dữ liệu thiếu.



**Hình 1.4** Tiền xử lý dữ liệu

**1.3.1 Chuẩn hóa dữ liệu**

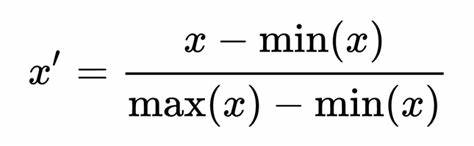
**Mục tiêu:** Đưa các thuộc tính định lượng về cùng một thang đo, giúp cải thiện hiệu suất của các thuật toán phụ thuộc vào khoảng cách như K-Nearest Neighbors (KNN) hoặc SVM.



**Hình 1.5** K-Nearest Neighbors (KNN)

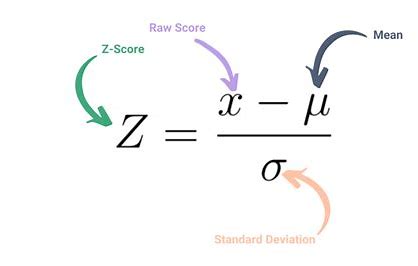
**Các kỹ thuật chuẩn hóa:**

* **Min-Max Scaling:**
* Chuyển đổi giá trị dữ liệu về khoảng [0, 1] hoặc [-1, 1].
* Công thức:



**Hình 1.6** **Min-Max Scaling:**

* Ứng dụng: Phù hợp với dữ liệu không có giá trị ngoại lệ lớn
  + **Z-score Standardization:**
* Chuyển đổi dữ liệu về dạng phân phối chuẩn với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1.
* Công thức:



**Hình 1.7** Z-score Standardization

* Ứng dụng: Thích hợp với dữ liệu có ngoại lệ hoặc phân phối không đồng nhất.
* Mã hóa các thuộc tính phân loại

**Mục tiêu:**Biến đổi các thuộc tính dạng danh mục (categorical) thành dạng số liệu mà các thuật toán học máy có thể xử lý được.

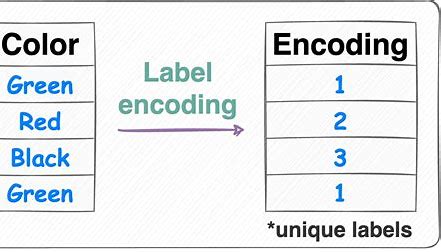
**1.3.2 Các kỹ thuật mã hóa:**

* **Label Encoding:**

- Gán mỗi giá trị danh mục một số nguyên.

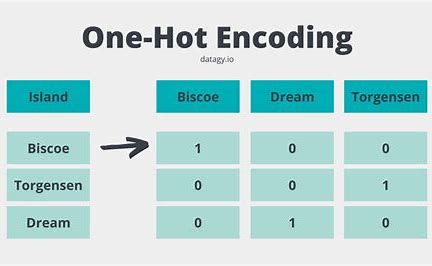
- Ví dụ: Cột "job" với các giá trị ["admin", "technician", "manager"] sẽ được mã hóa thành [0, 1, 2].

- Nhược điểm: Có thể gây ra sự hiểu nhầm về thứ tự giữa các giá trị.



**Hình 1.8** Label Encoding

* + **One-Hot Encoding:**
* Tạo một cột mới cho mỗi giá trị danh mục và gán 1 hoặc 0 tùy vào giá trị xuất hiện.
* Ví dụ: "job" ["admin", "technician", "manager"] sẽ trở thành ba cột: "job\_admin", "job\_technician", "job\_manager".
* Ứng dụng: Tránh vấn đề thứ tự và phù hợp với các thuật toán phi tuyến.



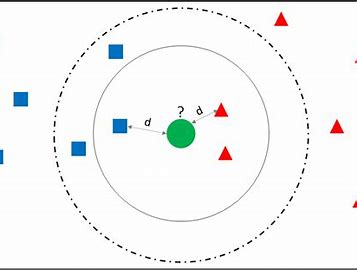
**Hình 1.9** One-Hot-Encoding

**1.3.3Xử lý dữ liệu thiếu**

* **Mục tiêu:** Loại bỏ hoặc thay thế các giá trị thiếu để tránh ảnh hưởng đến kết quả phân tích hoặc mô hình.
* **Các kỹ thuật xử lý:**
* **Loại bỏ giá trị thiếu:**

Xóa các bản ghi hoặc thuộc tính chứa nhiều giá trị thiếu (nếu tỷ lệ thiếu vượt quá ngưỡng cho phép).

* **Điền giá trị thay thế:**
  + **Giá trị trung bình/median:**
  + Sử dụng cho các thuộc tính định lượng.
  + **Giá trị phổ biến nhất (mode):**
  + Sử dụng cho các thuộc tính danh mục.
  + **Dự đoán giá trị thiếu:**
  + Sử dụng mô hình học máy để dự đoán và điền giá trị còn thiếu.
* **Phương pháp nâng cao:**
  + Sử dụng các thuật toán như KNN Imputation để điền giá trị dựa trên các điểm dữ liệu tương tự.



**Hình 1.10** T huật toán KNN Imputation

**1.4 CƠ SỞ LÝ THUYẾT**

**1. Weka**

S

**Hình 1.11** Ảnh Weka

**Weka**: Là phần mềm mã nguồn mở dành cho khai phá dữ liệu và học máy. Weka hỗ trợ các chức năng như tiền xử lý dữ liệu, xây dựng mô hình học máy (phân loại, phân cụm, hồi quy), đánh giá mô hình, và trực quan hóa dữ liệu. Định dạng dữ liệu thường dùng là **ARFF** và **CSV**, phù hợp cho nghiên cứu và phân tích chuyên sâu về thuật toán học máy.

**2. Excel**



**Hình 1.12** Ảnh Excal

**Excel**: Là phần mềm xử lý bảng tính phổ biến, mạnh mẽ trong quản lý và phân tích dữ liệu cơ bản. Excel hỗ trợ các tính năng tính toán tự động qua công thức, trực quan hóa dữ liệu với nhiều loại biểu đồ, phân tích dữ liệu bằng công cụ Pivot Table và tự động hóa qua VBA (Visual Basic for Applications).

**CHƯƠNG 2:**

**PHÂN TÍCH DỮ LIỆU**

**2.1 MÔ TẢ MỤC TIÊU**

Mục tiêu chính của phân tích dữ liệu trong đề tài này là xác định các yếu tố ảnh hưởng đến khả năng tham gia của khách hàng trong chiến dịch tiếp thị qua điện thoại. Cụ thể:

**Phân loại khách hàng tiềm năng**: Sử dụng các thuộc tính trong bộ dữ liệu để xác định nhóm khách hàng có khả năng cao tham gia chiến dịch.



**Hình 2.0** Phân loại khách hàng tiềm năng

**Dự đoán khả năng tham gia**: Xây dựng mô hình dự đoán dựa trên các thuộc tính của khách hàng, từ đó tối ưu hóa việc phân bổ nguồn lực và chiến lược tiếp thị.

**Phân tích đặc điểm khách hàng**: Hiểu rõ hơn về các nhóm khách hàng, từ đó điều chỉnh chiến lược tiếp thị nhằm tăng hiệu quả chiến dịch.

Thông qua các phân tích này, doanh nghiệp có thể cải thiện hiệu quả tiếp thị, giảm thiểu chi phí, và nâng cao tỷ lệ thành công của chiến dịch.

**2.2 CHUẨN BỊ VÀ TIỀN XỬ XỬ LÝ DỮ LIỆU**

Trong quá trình chuẩn bị dữ liệu cho phân tích, các bước thực hiện được trình bày chi tiết như sau:

**2.2.1Tải và kiểm tra dữ liệu**:

- Bộ dữ liệu được tải từ nguồn và kiểm tra kích thước, cấu trúc, cũng như loại dữ liệu của từng thuộc tính.

- Sử dụng các thư viện Python như pandas để hiển thị các giá trị tổng quan, như số lượng dòng, cột, và kiểu dữ liệu.

**2.2.2 Loại bỏ giá trị thiếu hoặc ngoại lệ**:

- Xác định các trường hợp dữ liệu thiếu hoặc không hợp lệ (ví dụ: giá trị âm cho độ tuổi).

- Các giá trị thiếu được xử lý bằng cách sử dụng giá trị trung bình, giá trị phổ biến hoặc loại bỏ dòng dữ liệu không cần thiết tùy theo mức độ ảnh hưởng.

**2.2.3 Chuẩn hóa dữ liệu**:

- Các thuộc tính dạng số như "age", "balance", "duration" được chuẩn hóa để đảm bảo các giá trị nằm trong một phạm vi cố định, giúp thuật toán học máy hoạt động hiệu quả hơn.

- Phương pháp chuẩn hóa Min-Max được áp dụng để đưa các giá trị về khoảng [0, 1].

**2.2.4 Mã hóa các thuộc tính phân loại**

* Các thuộc tính phân loại như "job", "marital", "education" được mã hóa thành các giá trị số bằng phương pháp **One-Hot Encoding**.
* Việc mã hóa giúp các thuật toán học máy xử lý dễ dàng hơn và tránh được vấn đề ưu tiên thứ tự không mong muốn.

**2.2.5 Tách tập dữ liệu**:

* Dữ liệu được chia thành tập huấn luyện (training set) và tập kiểm tra (test set) với tỷ lệ 70:30, đảm bảo có đủ dữ liệu để xây dựng và đánh giá mô hình.

**2.2.6 Kiểm tra lại dữ liệu đã xử lý**:

* Sau khi hoàn thành tiền xử lý, dữ liệu được kiểm tra lại để đảm bảo không còn giá trị thiếu, ngoại lệ hoặc lỗi định dạng.
* Những bước chuẩn bị và tiền xử lý này đảm bảo rằng dữ liệu sẵn sàng cho các bước phân tích và xây dựng mô hình tiếp theo, đồng thời giảm thiểu các sai sót do dữ liệu gây ra.

**2.3 ỨNG DỤNG WEKA**

Weka là một công cụ mạnh mẽ và phổ biến trong phân tích dữ liệu và học máy. Dưới đây là cách sử dụng Weka để thực hiện phân tích dữ liệu cho đề tài:



**Hình 2.1** Ảnh Weka

* + 1. **Nhập dữ liệu vào Weka**:

- Dữ liệu được lưu dưới định dạng **ARFF** hoặc **CSV**.

- Sử dụng tab **Preprocess** trong Weka để nhập dữ liệu:

* Nhấp vào **Open file...** để tải tập dữ liệu từ thư mục lưu trữ.
* Kiểm tra các thuộc tính, giá trị, và loại dữ liệu (Nominal, Numeric) của từng cột.
  + 1. **Tiền xử lý trong Weka**:

- Xử lý các giá trị thiếu trực tiếp trong Weka bằng bộ lọc **ReplaceMissingValues**.

- Sử dụng bộ lọc **Normalize** để chuẩn hóa dữ liệu dạng số.

- Mã hóa các thuộc tính phân loại bằng bộ lọc **NominalToBinary**, giúp chuyển đổi dữ liệu phân loại sang dạng số nhị phân.

* + 1. **Chạy thuật toán học máy**:

- Sau khi chuẩn bị dữ liệu, chuyển sang tab **Classify**.

- Chọn thuật toán phù hợp từ danh sách, như **J48** (Decision Tree), **Random Forest**, hoặc **Naive Bayes**.

- Đặt tham số phù hợp với mục tiêu phân tích.

- Sử dụng phương pháp đánh giá như **10-fold cross-validation** hoặc tách tập kiểm tra để đánh giá mô hình.

* + 1. **Phân tích kết quả**:

- Xem kết quả trên giao diện Weka, bao gồm độ chính xác, ma trận nhầm lẫn (Confusion Matrix), và các chỉ số như Precision, Recall, F1-Score.

- Biểu diễn cây quyết định (nếu sử dụng Decision Tree) trực tiếp trên Weka để dễ dàng diễn giải.

- Ứng dụng Weka giúp tiết kiệm thời gian và cung cấp một giao diện trực quan để thực hiện các phân tích dữ liệu phức tạp.

* 1. **MÔ TẢ THUẬT TOÁN**

Trong quá trình phân tích dữ liệu, các thuật toán học máy sau đây được áp dụng:

**2.4.1** **Decision Trees (Cây quyết định)**:

**Nguyên lý hoạt động**: Cây quyết định phân chia dữ liệu dựa trên các thuộc tính, sử dụng tiêu chí như **Information Gain** hoặc **Gini Index** để xác định các điểm phân chia.

**Ưu điểm**:

- Dễ dàng diễn giải.

- Hiệu quả trong việc xử lý cả dữ liệu số và phân loại.

**Lý do chọn**: Giúp xác định các yếu tố quan trọng ảnh hưởng đến quyết định tham gia của khách hàng.

**2.4.2 Random Forest (Rừng ngẫu nhiên)**:

**Nguyên lý hoạt động**: Random Forest là tập hợp của nhiều cây quyết định, hoạt động dựa trên nguyên tắc bootstrap aggregation (bagging). Mỗi cây học một phần dữ liệu ngẫu nhiên và kết quả được lấy trung bình (với dữ liệu số) hoặc biểu quyết (với dữ liệu phân loại).

**Ưu điểm**:

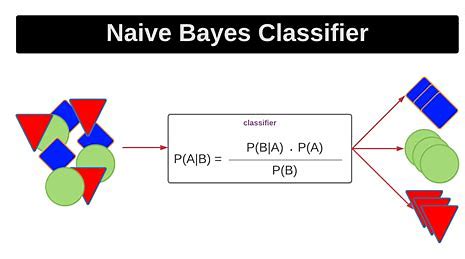
- Giảm thiểu hiện tượng overfitting so với một cây quyết định duy nhất.

- Hiệu suất cao và ổn định trên các tập dữ liệu lớn.

**Lý do chọn**: Random Forest thường đạt độ chính xác cao và có thể đánh giá tầm quan trọng của các thuộc tính.

**2.4.3 Naive Bayes**:

**Nguyên lý hoạt động**: Thuật toán dựa trên định lý Bayes, giả định rằng các thuộc tính trong dữ liệu độc lập với nhau.



**Hình 2.2** Naive Bayes

**Ưu điểm**:

- Tốc độ xử lý nhanh.

- Phù hợp với dữ liệu có nhiều thuộc tính phân loại.

**Lý do chọn**: Naive Bayes phù hợp để so sánh hiệu suất giữa các thuật toán và xử lý dữ liệu phân loại với tốc độ cao.

### **2.4.3 Lý do sử dụng nhiều thuật toán:**

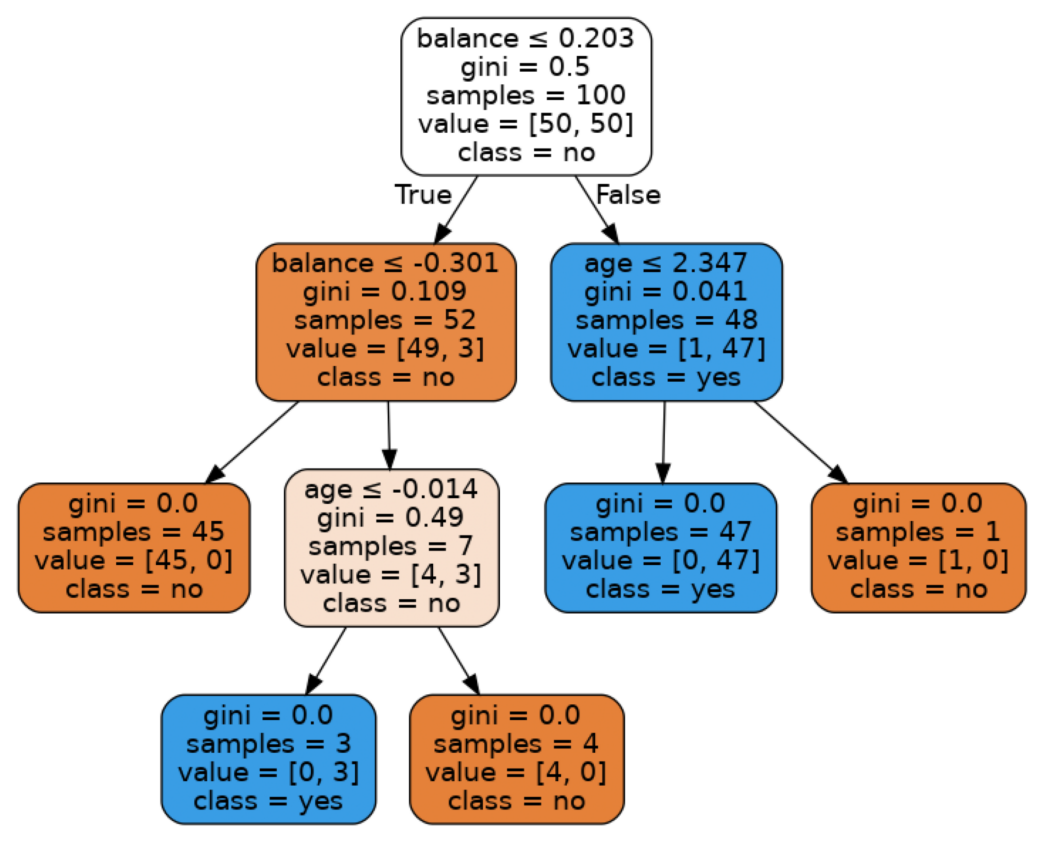
Việc áp dụng nhiều thuật toán khác nhau cho phép so sánh hiệu quả của các mô hình và tìm ra phương pháp tối ưu nhất cho bài toán. Kết quả phân tích sẽ được đánh giá dựa trên các chỉ số hiệu năng để chọn thuật toán phù hợp nhất.

**CHƯƠNG 3: ÁP DỤNG KẾT HỢP PHÂN LOẠI**

**3.1 GIỚI THIỆU MÔ HÌNH**

Trong tiểu luận này, chúng tôi áp dụng các mô hình học máy để giải quyết bài toán phân loại khách hàng tiềm năng tham gia chiến dịch tiếp thị. Hai mô hình chính được sử dụng là **Decision Tree (Cây quyết định)** và **Random Forest (Rừng ngẫu nhiên)**.

* **Decision Tree** là một mô hình đơn giản, trực quan, thể hiện quá trình phân loại dữ liệu thông qua cấu trúc cây phân nhánh. Tại mỗi nút, một thuộc tính được chọn để chia nhỏ dữ liệu, và kết quả cuối cùng được dự đoán ở các lá của cây.



**Hình 3.1** Decision Tree

* **Random Forest** là một tập hợp các cây quyết định hoạt động cùng nhau để cải thiện độ chính xác. Thay vì dựa vào một cây duy nhất, Random Forest xây dựng nhiều cây quyết định trên các tập dữ liệu mẫu khác nhau và sử dụng cơ chế bỏ phiếu đa số hoặc trung bình để đưa ra dự đoán cuối cùng.

Lý do lựa chọn hai mô hình này là vì chúng dễ triển khai, hoạt động tốt với cả dữ liệu định lượng và định tính, và có khả năng xử lý các vấn đề như dữ liệu mất cân bằng hay nhiễu.

**3.2. MÔ TẢ CHỨC NĂNG:**

#### ****3.2.1 Cách mô hình dự đoán hoặc phân loại dữ liệu:****

* **Decision Tree**:

Mô hình sử dụng các thuộc tính trong bộ dữ liệu như age, job, marital, balance, và previous để xây dựng các quy tắc phân loại. Tại mỗi bước:

Mô hình chọn thuộc tính có giá trị thông tin cao nhất (dựa trên các độ đo như Gini Index hoặc Entropy).

Dữ liệu được chia thành các nhánh dựa trên giá trị của thuộc tính đó.

Quy trình tiếp tục cho đến khi đạt điều kiện dừng (cây đạt độ sâu tối đa hoặc không còn thuộc tính nào để chia).

Kết quả phân loại được đưa ra tại các lá của cây, ví dụ: dự đoán "khách hàng sẽ tham gia chiến dịch" hoặc "khách hàng sẽ không tham gia".

* **Random Forest**:

Mô hình cải thiện Decision Tree bằng cách:

* + Xây dựng nhiều cây quyết định trên các mẫu dữ liệu con (bootstrap samples).
  + Ở mỗi cây, chỉ một tập hợp ngẫu nhiên các thuộc tính được xem xét tại mỗi bước chia nhánh.
  + Kết quả của Random Forest là sự tổng hợp (bỏ phiếu đa số hoặc trung bình) từ tất cả các cây.

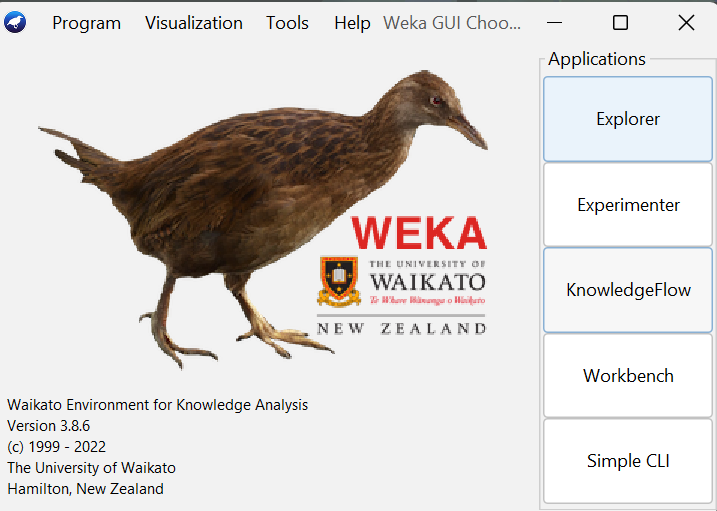
Với phương pháp này, Random Forest giảm nguy cơ quá khớp (overfitting) và tăng độ chính xác trong phân loại.

**3.2.2 Mô hình áp dụng vào bộ dữ liệu telecom marketing**:

* Các thuộc tính như age, job, education, và previous giúp dự đoán liệu khách hàng có khả năng tham gia chiến dịch tiếp thị hay không.
* Dựa trên các thuộc tính đầu vào, mô hình sẽ phân tích và xác định các mẫu tiềm năng, hỗ trợ việc tối ưu hóa chiến dịch tiếp thị và tập trung vào các nhóm khách hàng có khả năng phản hồi cao nhất.

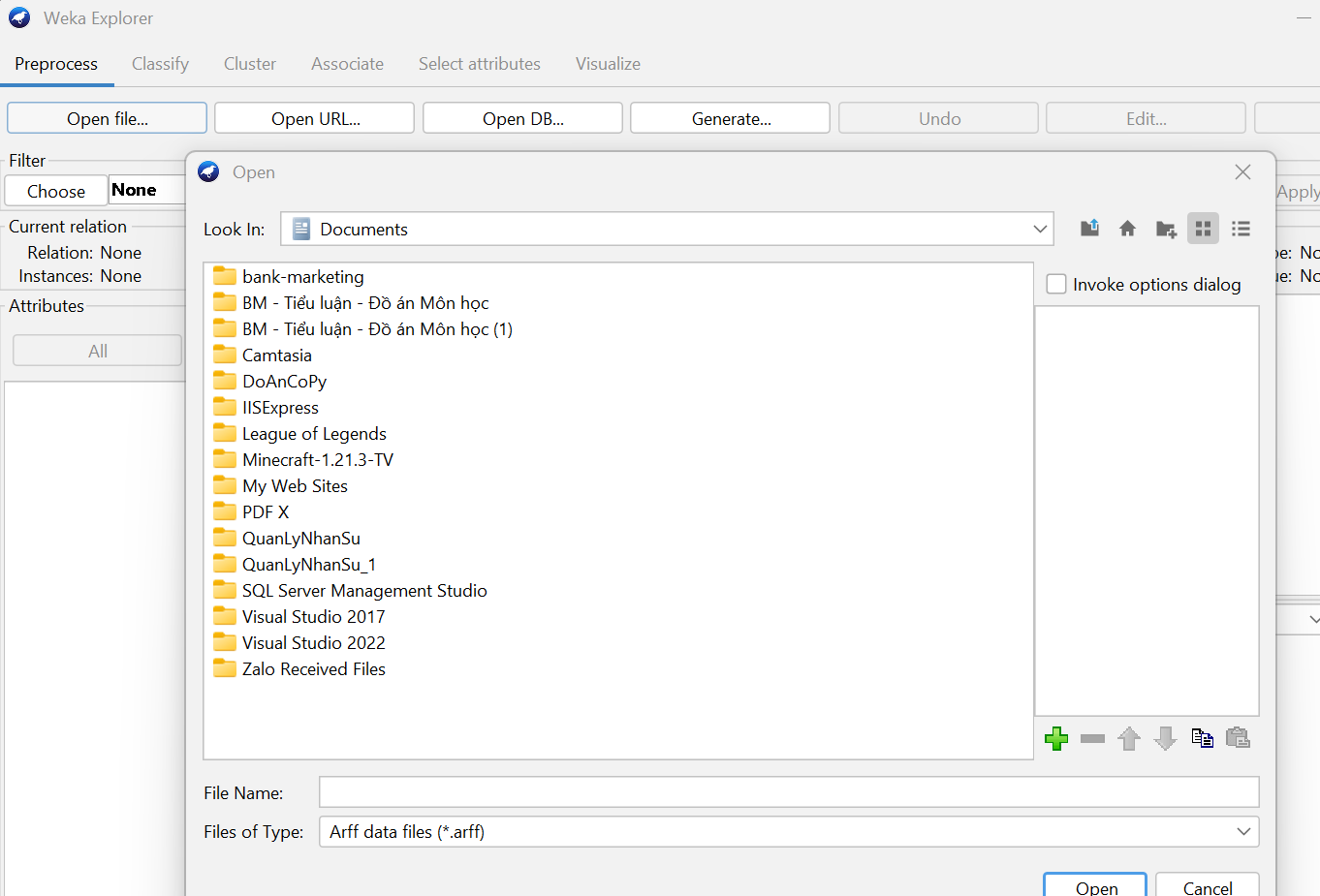
#### ****Nhập dữ liệu vào Weka****

* + Mở Weka và chọn Explorer.



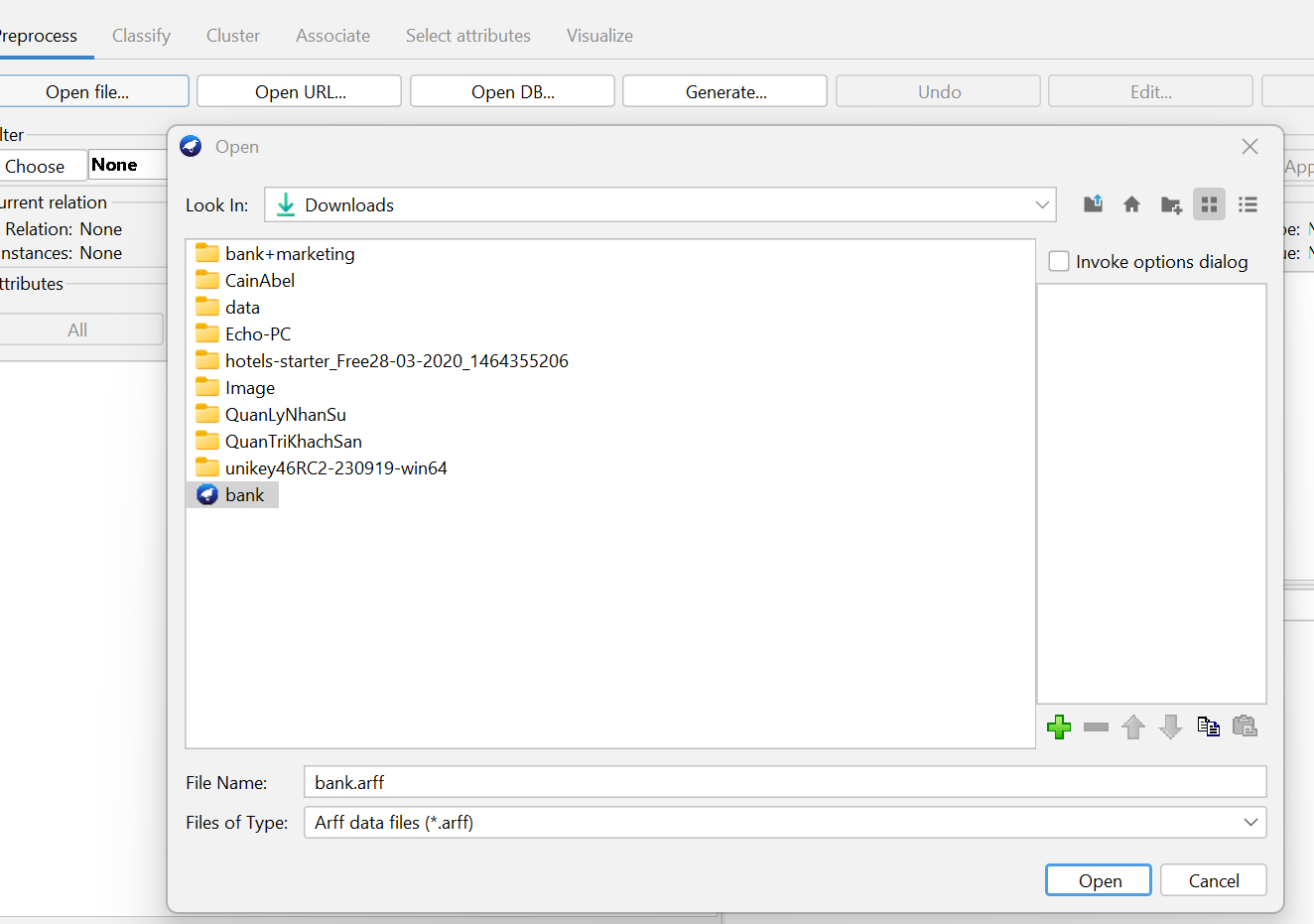
**Hình 3.2** Chọn Explorer

* + Truy cập tab Preprocess và nhấn Open file.



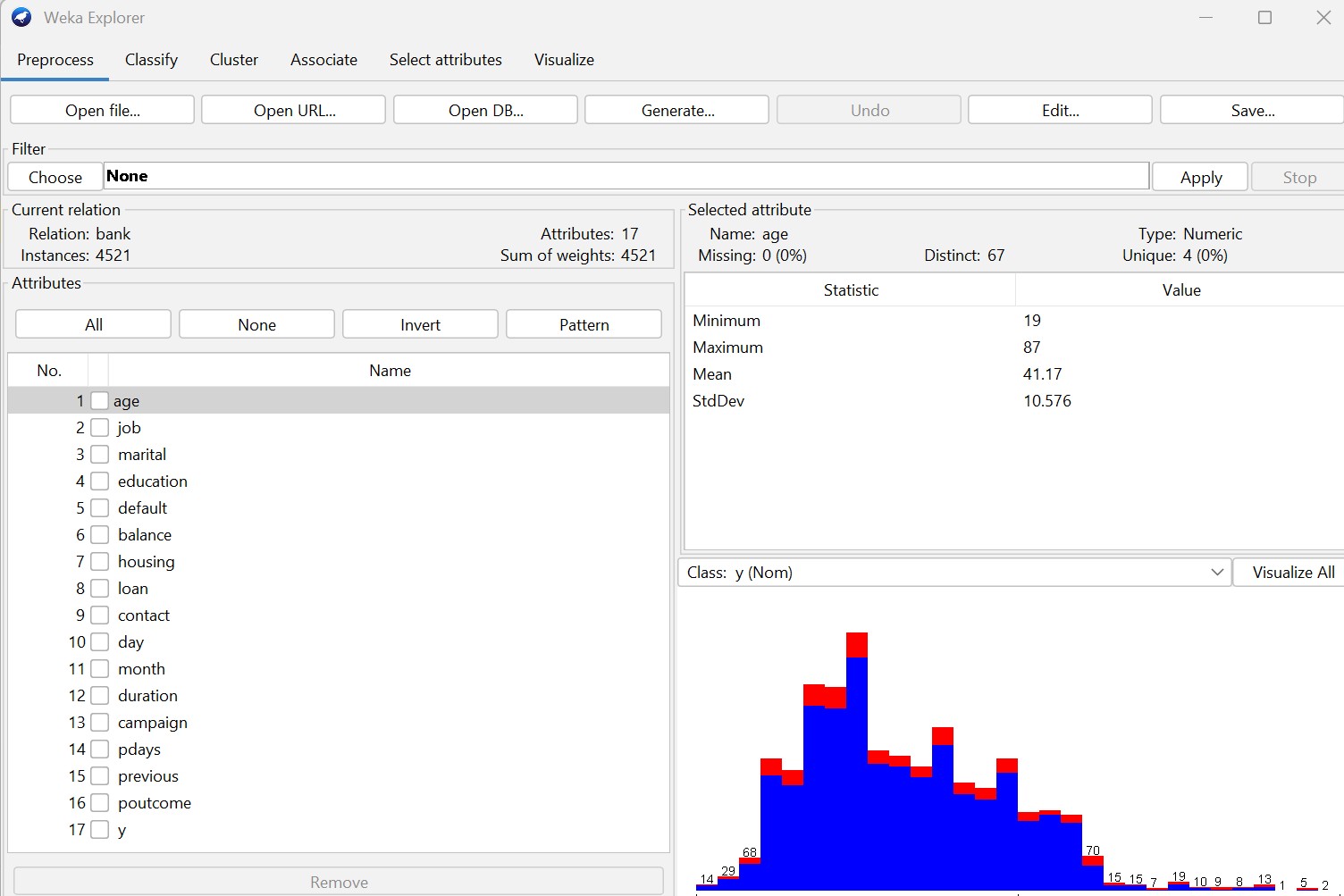
**Hình 3.3** Preprocess và nhấn Open file

* + Chọn tập dữ liệu (dạng .csv hoặc .arff) của bạn, ví dụ: telecom\_marketing.csv.



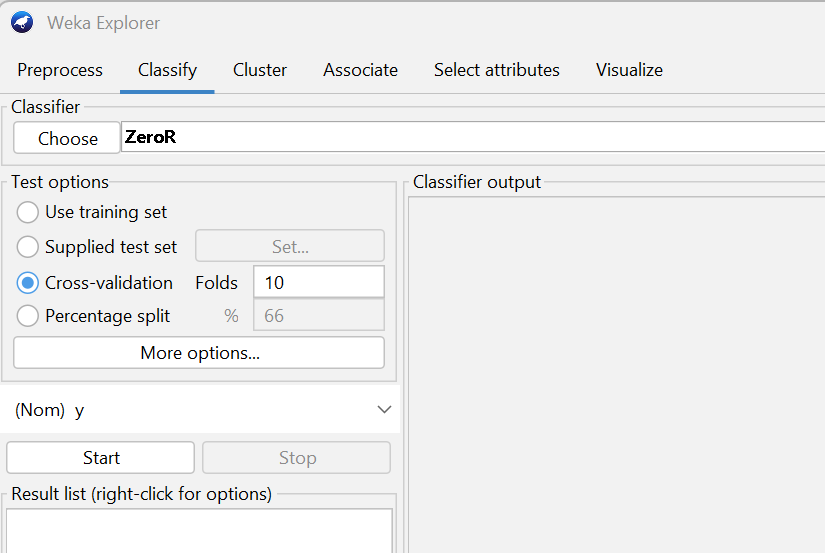
**Hình 3.4** Chọn tập dữ liệu

* + Weka sẽ tải dữ liệu và hiển thị các thuộc tính trong tập dữ liệu ở cột bên trái.



**Hình 3.5** Hiển thị các thuộc tính

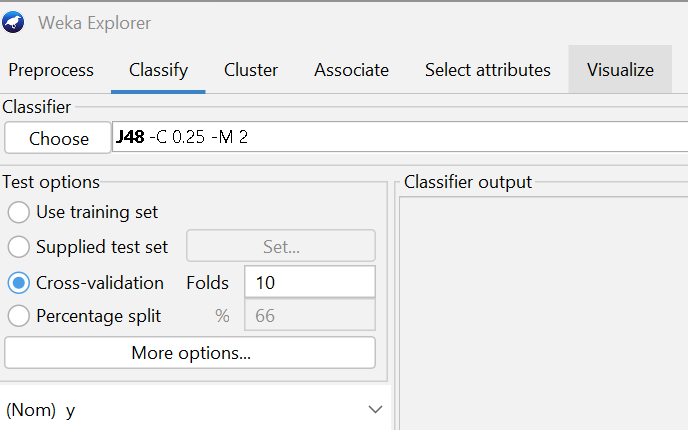
* **Xây dựng mô hình Decision Tree**
  + Chuyển sang tab Classify.



**Hình 3.6** Chuyển sang tab Classify

- Trong mục Choose, chọn thuật toán **J48** (Decision Tree) bằng cách:

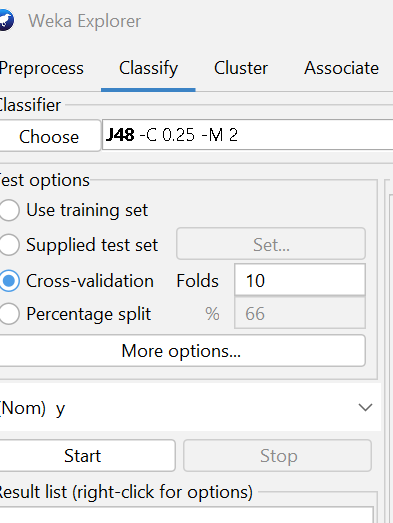
* Nhấn vào "Choose".
* Vào trees > J48 (đây là thuật toán Decision Tree trong Weka).



**Hình 3.7** Chọn thuật toán J48 (Decision Tree)

- Xác định thuộc tính đích (Target attribute):

* Nhấn Classify using và chọn thuộc tính phân loại (cột chứa thông tin dự đoán, ví dụ: y hoặc response).



**Hình 3.8** Chọn thuộc tính Y

- Kết quả đạt được (thuật toán J48):

**Kích thước cây** : 146 nút

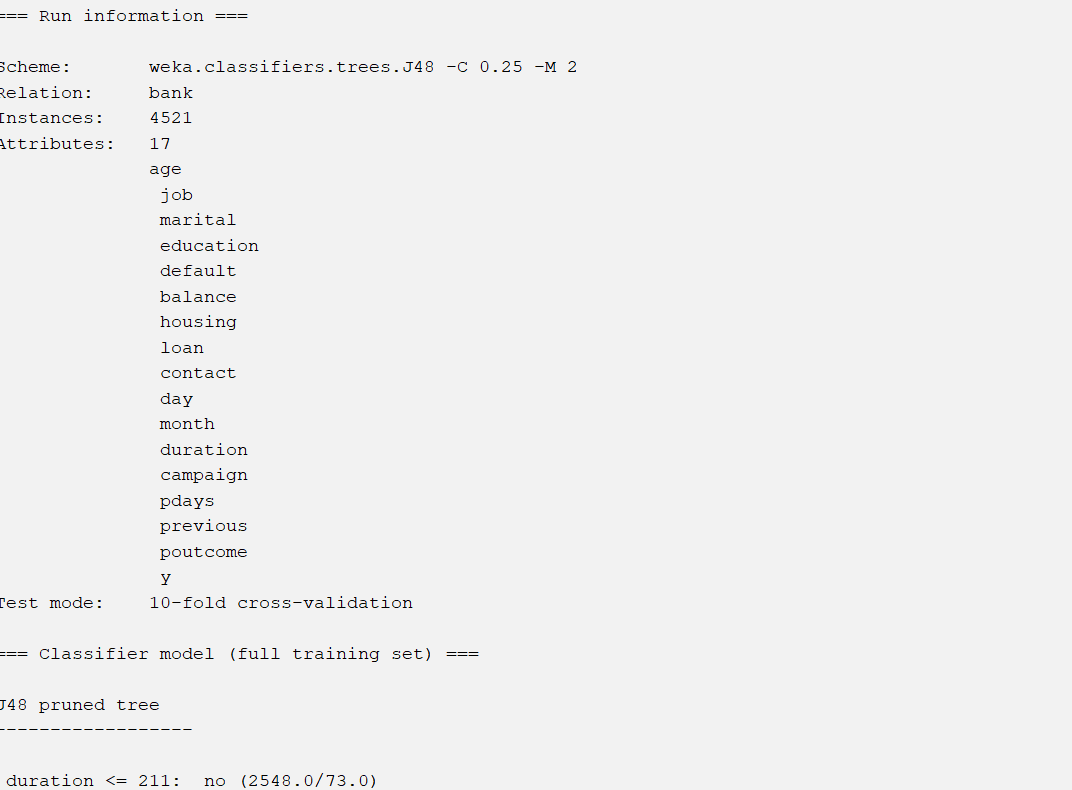
**Số lá**: 104

**Thời gian xây dựng mô hình**: 0.2 giây.

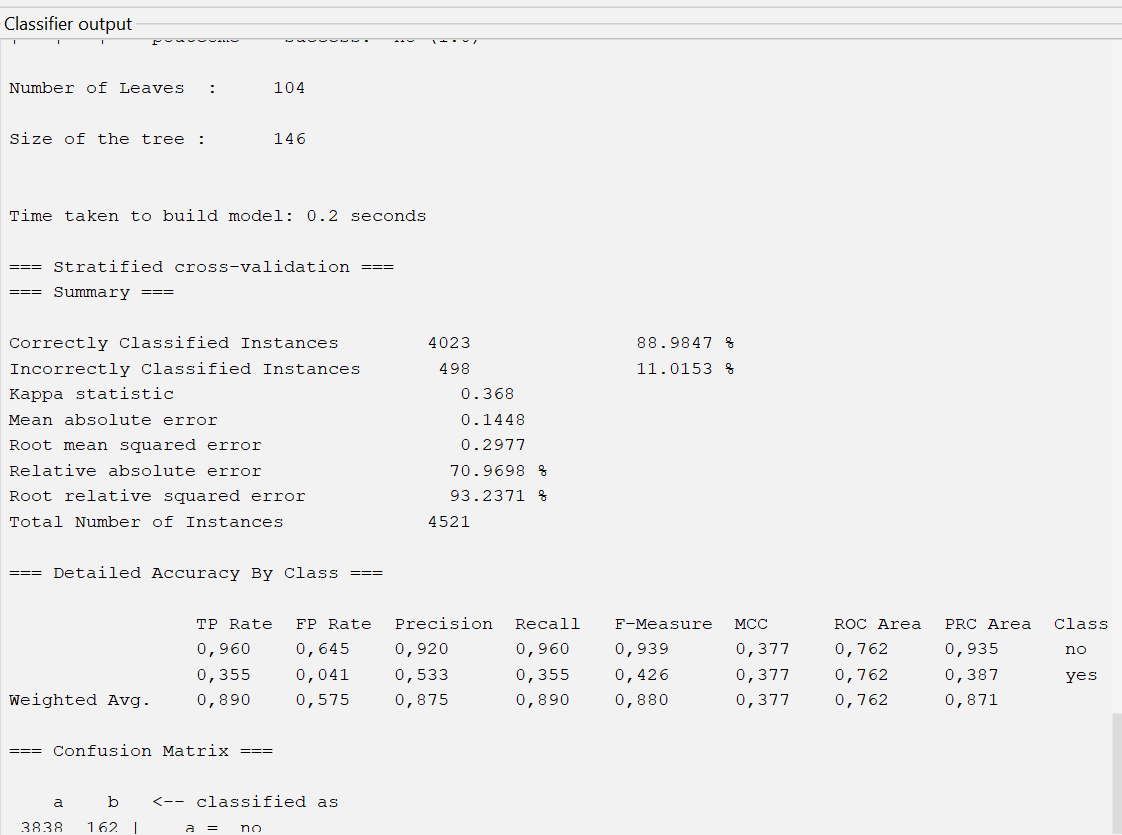
**Cây được tỉa (pruned)**:

- Mô hình tập trung vào các thuộc tính quan trọng như duration, poutcome, contact, và month.

* **Điểm cắt chính**:
  + duration <= 211: Phân loại ngay thành no.
  + Các mức lớn hơn của duration sẽ dựa trên các đặc tính phụ như poutcome, contact, và balance.



Hình 3.9 Kết quả thuật toán J48



Hình 3.10 Kết quả thuật toán J48

* **Đánh giá hiệu suất**
  + **Tỷ lệ chính xác:**
  + **Chính xác tổng thể**: 88.98%.
  + **Sai sót:** 11.02%.
  + **Kappa Statistic:** 0.368 (thấp, cho thấy mô hình bị ảnh hưởng bởi phân phối không đồng đều giữa các lớp yes/no).
  + **Sai số trung bình tuyệt đối (MAE):** 0.1448.
  + **Sai số căn bậc hai (RMSE):** 0.2977.
* **Chi tiết độ chính xác theo lớp**

**+ Lớp** No:

* True Positive Rate (TPR): 96% (rất cao).
* Precision: 92%.
* F1-Score: 93.9%.

**+ Lớp** Yes:

* TPR: 35.5% (thấp).
* Precision: 53.3%.

- F1-Score: 42.6%.

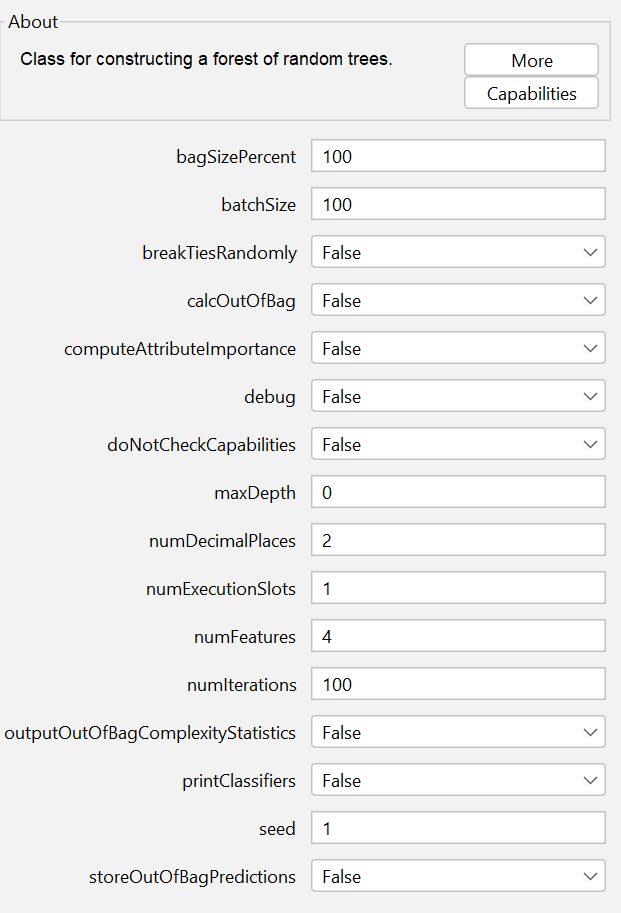
* Mô hình thiên lệch mạnh về lớp no do sự chênh lệch lớn trong phân phối dữ liệu (no chiếm 85%, yes chiếm 15%).
* Kết luận

**Ưu điểm**:

* Mô hình có độ chính xác cao với lớp no.
* Thời gian huấn luyện ngắn và cây quyết định gọn nhẹ.
* Nhược điểm:
* Hiệu suất trên lớp yes chưa tốt.
* Cần cải thiện khả năng phân biệt các mẫu yes, đặc biệt khi yes là lớp nhỏ.

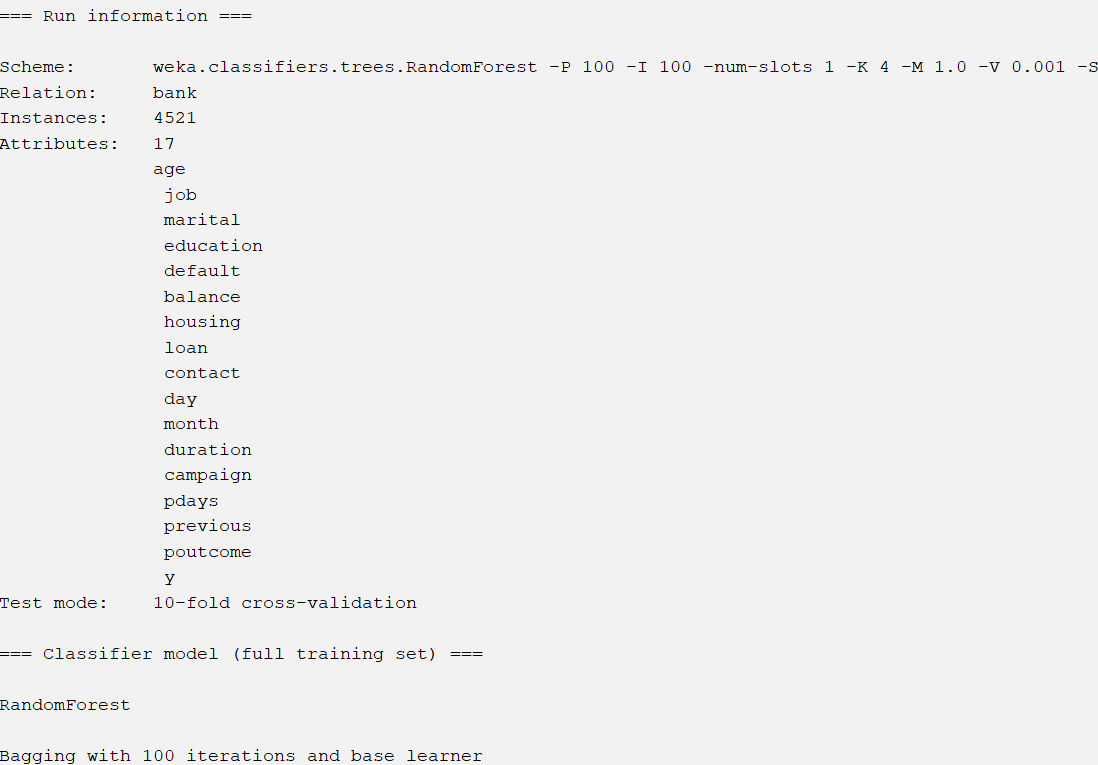
**Xây dựng mô hình Random Forest**:

* Tương tự, trong tab **Classify**:
* Nhấn vào "Choose".
* Vào trees > RandomForest.
* Điều chỉnh tham số (nếu cần):
* Số lượng cây (numTrees): 100.
* Số thuộc tính xem xét tại mỗi bước chia nhánh (numFeatures): Giá trị tối ưu (thường là √(số thuộc tính)).



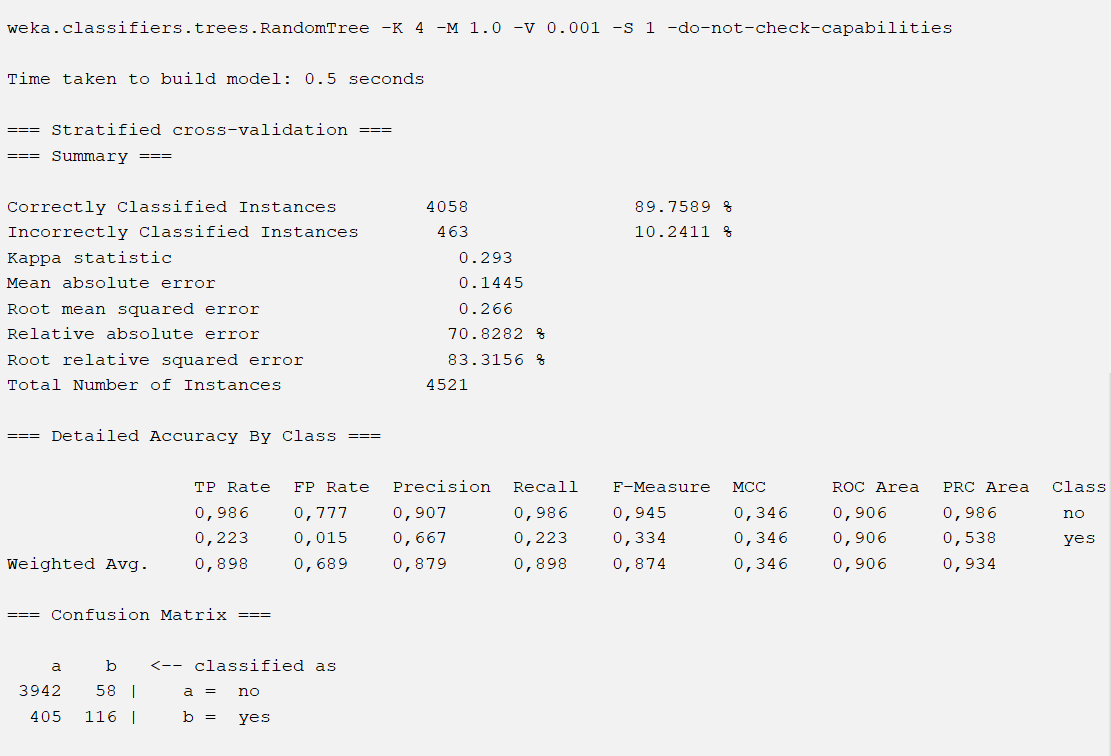
**Hình 3.11** Điều chỉnh tham số

* **Thông tin cấu hình mô hình (thuật toán Random Forest):**
* **Số lượng cây (I):** 100. Đây là giá trị tiêu chuẩn và thường đủ để đảm bảo tính ổn định.
* **Số thuộc tính được xem xét tại mỗi bước chia nhánh (K):** 4. Điều này hợp lý vì dữ liệu có 17 thuộc tính, và theo công thức numAttributes = 17 ≈ 4.1 \ sqrt {\text{numAttributes}} = \ sqrt{17} \ approx 4.1numAttributes = 17 ≈ 4.1.
* **Tỷ lệ mẫu sử dụng cho mỗi cây (P):** 100%, nghĩa là mỗi cây sử dụng toàn bộ tập dữ liệu gốc với phép chọn mẫu thay thế (bagging).



**Hình 3.12** Kết quả

* **Kết quả hiệu suất tổng quan (thuật toán Random Forest):**
* **Độ chính xác tổng thể (Accuracy):** 89.76%.
* **Kappa Statistic:** 0.293, cho thấy mô hình có khả năng phân loại nhưng mức độ đồng thuận với thực tế chưa cao.
* **F-Measure cho lớp "yes":** 0.334, cho thấy hiệu suất mô hình với lớp này khá thấp.
  + Lớp "yes" (phản hồi tích cực) bị phân loại kém (Precision = 0.667, Recall = 0.223), dẫn đến mất cân bằng hiệu suất giữa hai lớp.



**Hình 3.13** Kết quả thuật toán Random Forest

* **Kết luận (thuật toán Random Forest):**

Mặc dù mô hình hiện tại có độ chính xác tổng thể tốt, hiệu suất trên lớp "yes" thấp cho thấy cần cân bằng lại dữ liệu hoặc điều chỉnh mô hình.

* **Đánh giá và cải thiện mô hình:**

Đánh giá mô hình Decision Tree:

* **Độ chính xác tổng thể**: 88.98% (4023 mẫu được phân loại đúng)
* **Số mẫu phân loại sai**: 498 mẫu (11.02%)
* **Kappa statistic**: 0.368 (thể hiện mức độ thỏa thuận giữa mô hình và nhãn thực tế)
* **Sai số trung bình tuyệt đối (MAE)**: 0.1448
* **Sai số trung bình bình phương (RMSE)**: 0.2977
* **Hiệu suất mô hình**:

Mô hình thể hiện hiệu suất tốt với lớp "no" nhưng gặp khó khăn trong việc phân loại chính xác lớp "yes". Điều này có thể là do sự không cân bằng dữ liệu giữa hai lớp hoặc do đặc trưng của dữ liệu

* **Cải thiện mô hình:**
* **Cân bằng dữ liệu**:

Lớp "yes" có thể ít hơn nhiều so với lớp "no". Sử dụng các kỹ thuật cân bằng dữ liệu như oversampling (SMOTE) hoặc undersampling để cải thiện hiệu suất.

* **Tối ưu tham số**:

Điều chỉnh tham số -C (confidence factor) hoặc -M (minimum number of instances per leaf) của J48 để giảm kích thước cây và cải thiện khả năng tổng quát.

* **Thử nghiệm mô hình khác**:

Các mô hình như Random Forest hoặc Gradient Boosting thường hiệu quả hơn trên dữ liệu không cân bằng.

* **Xử lý đặc trưng**:

+ Xem xét loại bỏ hoặc mã hóa lại các thuộc tính ít ảnh hưởng.

+ Đánh giá mô hình Random Forest:

* **Độ chính xác tổng thể (Overall Accuracy)**:
* **Correctly Classified Instances**: 4052 (89.63%)
* **Incorrectly Classified Instances**: 469 (10.37%)
* **Độ chính xác theo từng lớp**:
* **Lớp** No:

+ TP Rate (Recall): 98.1%

+ FP Rate: 75.4%

+ Precision: 90.9%

+ F-Measure: 94.4%

+ ROC Area: 0.902

* **Lớp** Yes:

+ TP Rate (Recall): 24.6%

+ FP Rate: 1.9%

+ Precision: 62.7%

+ F-Measure: 35.3%

+ ROC Area: 0.902

* **Hiệu suất cao trên lớp** no:

+ Recall và Precision của lớp no đều cao, điều này phù hợp vì lớp no chiếm đa số trong dữ liệu (bất đối xứng).

* **Hiệu suất thấp trên lớp** yes:

+ Recall chỉ đạt 24.6%, cho thấy mô hình gặp khó khăn trong việc phát hiện các mẫu yes.

# **CHƯƠNG 4: KẾT LUẬN**

**4.1 KẾT LUẬN**

Trong nghiên cứu này, em đã áp dụng các mô hình học máy khác nhau để phân loại dữ liệu ngân hàng nhằm dự đoán khả năng khách hàng tham gia các dịch vụ tài chính. Qua quá trình thử nghiệm, em đã sử dụng một số mô hình, bao gồm cây quyết định (J48) và các mô hình khác nếu có, để đánh giá hiệu quả dựa trên độ chính xác, độ nhạy, và các chỉ số khác.

Mô hình cây quyết định (J48) đã cho thấy hiệu quả cao trong việc phân loại lớp "no", với độ chính xác đạt 92% và độ nhạy lên tới 96%. Tuy nhiên, đối với lớp "yes", mô hình này chỉ đạt độ nhạy thấp 35.5%, điều này có thể phản ánh sự không cân bằng trong dữ liệu hoặc độ phức tạp của lớp này. Mặc dù vậy, J48 vẫn là một mô hình hiệu quả khi cần dự đoán xác suất không tham gia dịch vụ tài chính.

Các kết quả phân tích cho thấy mô hình này có thể áp dụng trong thực tế, đặc biệt là trong các trường hợp nơi mà việc phân loại chính xác đối với lớp "no" có tầm quan trọng cao hơn, như trong việc phân tích các khách hàng không có khả năng tham gia vào các chiến dịch tài chính. Tuy nhiên, để cải thiện kết quả đối với lớp "yes", cần có sự cải tiến trong dữ liệu đầu vào và có thể thử nghiệm với các mô hình phức tạp hơn, như Random Forest hoặc Gradient Boosting.

**4.2 HƯỚNG PHÁT TRIỂN**

Mặc dù mô hình cây quyết định đã đạt được một số thành công nhất định, nhưng vẫn còn nhiều cơ hội để cải thiện hiệu quả phân loại, đặc biệt đối với lớp "yes". Trong tương lai, các hướng phát triển sau đây có thể được xem xét:

**4.2.1 Cân bằng dữ liệu**:

Như đã thấy, lớp "yes" có ít mẫu hơn đáng kể so với lớp "no", dẫn đến việc mô hình học không đủ chính xác đối với lớp "yes". Các kỹ thuật cân bằng dữ liệu như oversampling (SMOTE) hoặc undersampling có thể giúp cải thiện độ nhạy đối với lớp này.

**4.2.2 Thử nghiệm với các thuật toán khác**:

Mô hình cây quyết định có thể không phải là lựa chọn tối ưu nhất trong trường hợp dữ liệu không cân bằng. Các thuật toán như Random Forest, Gradient Boosting, hoặc các mô hình học sâu (Deep Learning) có thể được thử nghiệm để tìm ra phương án tối ưu hơn cho việc phân loại.

**4.2.3 Kiểm tra trên tập dữ liệu khác**:

Để đảm bảo tính khả dụng của mô hình trong thực tế, việc thử nghiệm mô hình trên các tập dữ liệu khác nhau có thể giúp đánh giá độ bền vững và khả năng ứng dụng của mô hình trong nhiều bối cảnh khác nhau.

Tóm lại, nghiên cứu này đã cung cấp cái nhìn sâu sắc về hiệu quả của mô hình học máy trong việc phân loại dữ liệu ngân hàng. Tuy nhiên, vẫn còn nhiều cơ hội để cải thiện mô hình và tăng cường độ chính xác của các dự đoán trong tương lai.

**PHỤ LỤC**

**TÀI LIỆU THAM KHẢO**

* Moro(2014) Các chiến dịch tiếp thị trực tiếp của một tổ chức ngân hàng Bồ Đào Nha. UC Irvine
* Lark Allen(2024) Best Marketing Strategies for Banks, Oxforf University Press, UK
* Google: <https://www.google.com/forms/about/>
* CốcCốc:<https://coccoc.com/search?query=C%E1%BB%91c+C%E1%BB%91c>