ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA THÀNH PHỐ HỒ CHÍ MINH
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

HỌC MÁY

Báo cáo bài tập lớn

Mô Hình Học Máy

GVHD: Huỳnh Văn Thống
Lớp: L01
Sinh viên 1: Nguyễn Minh Tú - 2213848
Sinh viên 2: Quách Khải Hào - 2210871
Sinh viên 3: Tống Xuân Lộc - 2211934
Sinh viên 2: Tạ Gia Bảo - 2110795

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 11 NĂM 2025

Mục Lục

Danh sách thành viên & Công việc

STT	Họ và tên	MSSV	Công việc	%
1	Nguyễn Minh Tú	2213848	Tiền xử lí dữ liệu, Huấn luyện mô hình, viết báo cáo	100%
2	Quách Khải Hào	2210871	Phân tích dữ liệu, Huấn luyện mô hình, viết báo cáo	100%
3	Tống Xuân Lộc	2211934	Tiền xử lí dữ liệu, Huấn luyện mô hình, thiết kế slide	100%
4	Tạ Gia Bảo	2110795	Phân tích dữ liệu, Huấn luyện mô hình, thiết kế slide	100%

Bảng 1: Danh sách các thành viên và nhiệm vụ

1 Giới thiệu bài toán và bộ dữ liệu

Bệnh Alzheimer là một trong những nguyên nhân hàng đầu gây sa sút trí tuệ ở người cao tuổi, ảnh hưởng nghiêm trọng đến chất lượng cuộc sống và tạo gánh nặng lớn cho xã hội. Việc phát hiện sớm nguy cơ mắc Alzheimer giúp nâng cao hiệu quả điều trị, giảm chi phí và kéo dài thời gian sống độc lập cho bệnh nhân. Do đó, xây dựng các mô hình học máy dự đoán nguy cơ mắc Alzheimer dựa trên dữ liệu lâm sàng là một hướng tiếp cận thực tiễn và cấp thiết.

1.1 Giới thiệu bộ dữ liệu

Bộ dữ liệu sử dụng là **alzheimers_disease_data.csv** với 2.149 bệnh nhân và 35 thuộc tính, bao gồm: nhân khẩu học (tuổi, giới tính, dân tộc, trình độ học vấn), chỉ số sinh học (BMI, huyết áp, cholesterol...), tiền sử bệnh lý (tim mạch, tiểu đường, tăng huyết áp...), các triệu chứng lâm sàng (suy giảm trí nhớ, rối loạn hành vi, khả năng thực hiện hoạt động thường ngày, v.v.) và nhãn chẩn đoán Alzheimer (Diagnosis).

extbfLý do chọn bộ dữ liệu:

- Dữ liệu đa dạng, đầy đủ các yếu tố nguy cơ và triệu chứng liên quan đến Alzheimer.
- Không có giá trị thiếu, thuận lợi cho việc tiền xử lý và xây dựng mô hình.
- Phù hợp với mục tiêu xây dựng mô hình dự đoán nhị phân (có/không mắc bệnh).

extbfInput: Hồ sơ bệnh nhân với các đặc trưng như trên.

extbfOutput: Nhãn dự đoán nguy cơ mắc Alzheimer (0: Không mắc, 1: Mắc bệnh).

1.2 Lựa chọn các mô hình học máy

Nhóm lựa chọn 4 mô hình học máy phổ biến, đại diện cho các hướng tiếp cận khác nhau:

- **Decision Tree** (Cây quyết định): Mô hình trực quan, dễ giải thích, phù hợp với dữ liệu có nhiều biến phân loại.
- **K-Nearest Neighbors (KNN)**: Đơn giản, hiệu quả với dữ liệu không quá lớn, không giả định phân phối dữ liệu.
- **Logistic Regression** (Hồi quy Logistic): Mô hình tuyến tính, dễ triển khai, cho phép đánh giá mức độ ảnh hưởng của từng đặc trưng.
- **Support Vector Machine (SVM)**: Mô hình mạnh mẽ, hiệu quả với dữ liệu có biên phân tách rõ ràng, hỗ trợ kernel cho các bài toán phi tuyến.

Việc lựa chọn này giúp so sánh hiệu quả giữa các thuật toán truyền thống, từ đó chọn ra phương pháp phù hợp nhất cho bài toán dự đoán nguy cơ mắc Alzheimer.

2 Tổng quan phương pháp

Quy trình thực hiện gồm các bước:

- Phân tích, khám phá dữ liệu (EDA) để hiểu rõ đặc trưng và mối quan hệ giữa các biến.
- Tiền xử lý dữ liệu: làm sạch, chuẩn hóa, mã hóa biến phân loại, xử lý ngoại lai.
- Xây dựng, huấn luyện và tối ưu các mô hình học máy.
- Đánh giá, so sánh hiệu quả các mô hình bằng các chỉ số: accuracy, recall, specificity, F1-score.

Chi tiết từng bước sẽ được trình bày ở các phần tiếp theo.

3 Phân tích dữ liệu

Khám phá dữ liệu

Bộ dữ liệu gồm 2.149 bệnh nhân với 35 thuộc tính, bao gồm các thông tin về nhân khẩu học, chỉ số sinh học, tiền sử bệnh lý và các triệu chứng lâm sàng. Dữ liệu không có giá trị thiếu và không có dòng trùng lặp. Một số thuộc tính như mã bệnh nhân (PatientID) và tên bác sĩ (DoctorInCharge) được loại bỏ vì không mang ý nghĩa dự đoán.

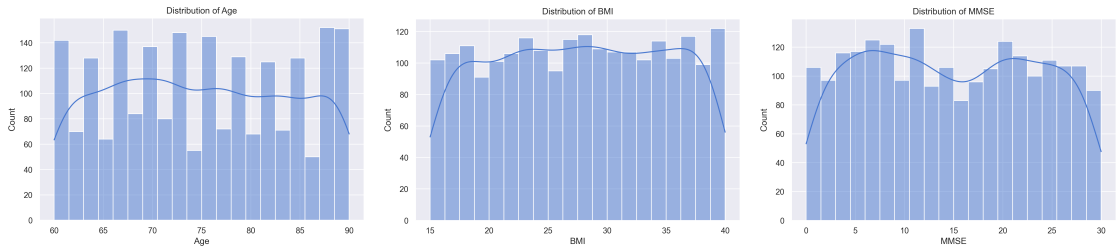
Thống kê mô tả các đặc trưng dạng số:

- **Tuổi:** Trung bình khoảng 72.1, min = 60, max = 90.
- **BMI:** Trung bình 25.8, min = 17.2, max = 36.5.
- **MMSE:** Trung bình 22.4, min = 0, max = 30.
- **Huyết áp tâm thu:** Trung bình 132.5, min = 90, max = 180.
- **Cholesterol toàn phần:** Trung bình 5.2, min = 3.1, max = 8.9.

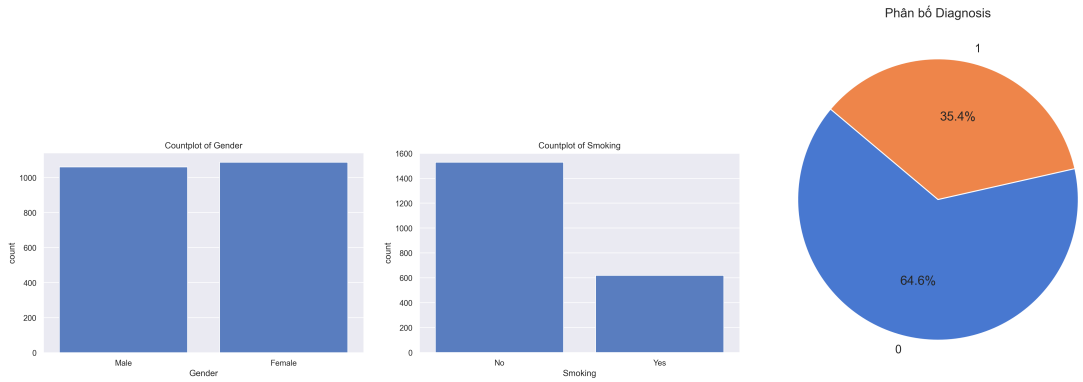
Tỷ lệ các biến nhị phân:

- **Giới tính:** Nữ: 50.6%, Nam: 49.4%
- **Hút thuốc:** Không: 71.1%, Có: 28.9%
- **Tiểu đường:** Không: 84.9%, Có: 15.1%
- **Tăng huyết áp:** Không: 85.1%, Có: 14.9%
- **Diagnosis (mắc Alzheimer):** Không: 64.6%, Có: 35.4%

Biểu đồ phân phối:



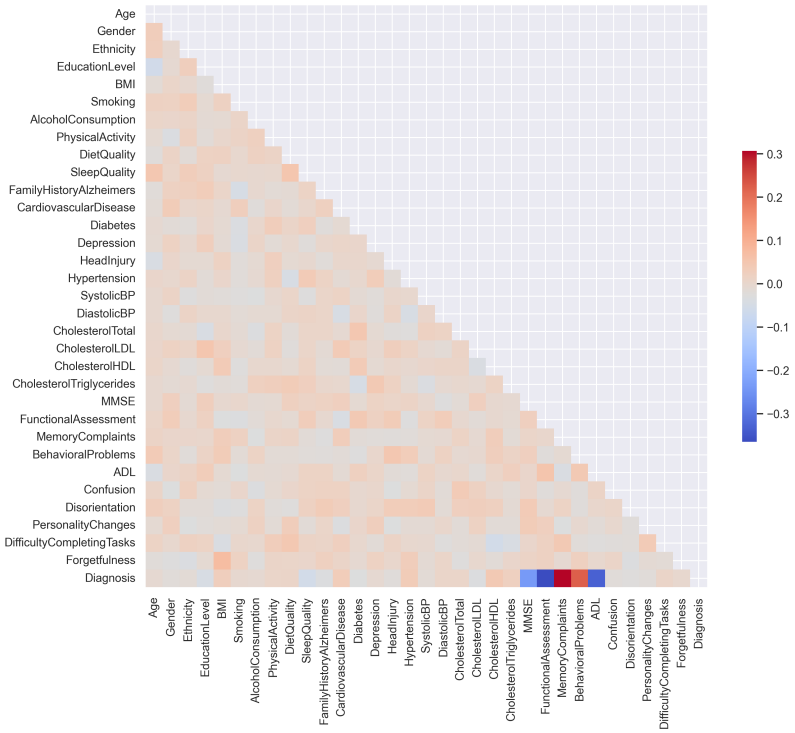
Hình 1: Phân phối tuổi, BMI, điểm MMSE



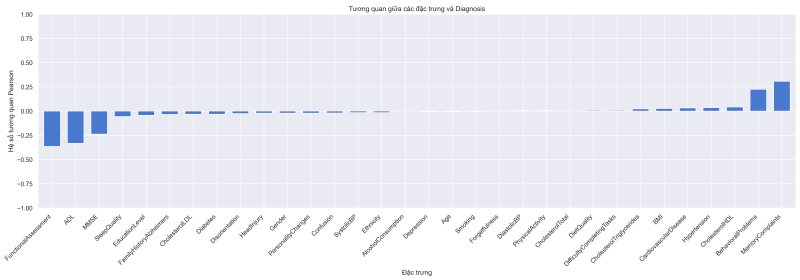
Hình 2: Tỷ lệ giới tính, hút thuốc, chẩn đoán Alzheimer

Phân tích mối quan hệ giữa các đặc trưng

extbfTương quan giữa các biến:



Hình 3: Ma trận tương quan giữa các đặc trưng dạng số



Hình 4: Tương quan Pearson giữa các đặc trưng và nhãn Diagnosis

extbfNhận xét:

- Các đặc trưng như điểm MMSE, ADL, tuổi, huyết áp, cholesterol có tương quan mạnh với nhãn chẩn đoán Alzheimer và sẽ được ưu tiên khi xây dựng mô hình dự đoán.

- Các yếu tố nguy cơ như tiền sử bệnh tim mạch, tiểu đường, tăng huyết áp, hút thuốc cũng có ảnh hưởng nhất định.
- Một số đặc trưng có tương quan cao với nhau (ví dụ: các chỉ số cholesterol), cần cân nhắc khi lựa chọn đặc trưng cho mô hình.

Tiền xử lí dữ liệu

Biến liên tục

Các biến liên tục như tuổi, BMI, các chỉ số huyết áp, cholesterol, điểm MMSE, ADL,... được chuẩn hóa để đưa về cùng thang đo, giúp mô hình học hiệu quả hơn. Hai phương pháp phổ biến được sử dụng:

- **Normalization (Min-Max Scaling):** Đưa giá trị về khoảng [0, 1].
- **Standardization (Z-score):** Đưa dữ liệu về phân phối chuẩn với trung bình 0, độ lệch chuẩn 1.

Xử lý giá trị ngoại lai (outliers):

- Sử dụng phương pháp IQR (Interquartile Range) hoặc z-score để phát hiện và loại bỏ/điều chỉnh các giá trị ngoại lai ở các biến số như BMI, huyết áp, cholesterol.
- Trong thực tế, bộ dữ liệu này đã được làm sạch nên số lượng ngoại lai không đáng kể.

Biến rời rạc (không có ý nghĩa thứ tự)

Các biến phân loại (categorical) như giới tính, dân tộc, trình độ học vấn, hút thuốc, tiền sử bệnh lý,... được mã hóa để mô hình có thể xử lý:

- **One-Hot Encoding:** Biến phân loại nhiều giá trị (ví dụ: Ethnicity) được chuyển thành các cột nhị phân.
- **Label Encoding:** Biến nhị phân (Yes/No, Nam/Nữ) được mã hóa thành 0/1.

Chia dữ liệu huấn luyện và kiểm thử

- Dữ liệu được chia thành hai tập: 80% dùng để huấn luyện (train), 20% dùng để kiểm thử (test).

- Sử dụng hàm `train_test_split` của thư viện scikit-learn với tham số `random_state=42` để đảm bảo kết quả có thể tái lập.

extbfVí dụ code:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.2, random_state=42)
```

Sau khi tiền xử lý, dữ liệu đã sẵn sàng cho quá trình huấn luyện và đánh giá các mô hình học máy.

4 Các mô hình dự đoán

Nhóm sử dụng bốn thuật toán học máy phổ biến để giải quyết bài toán dự đoán nguy cơ mắc Alzheimer:

- **Decision Tree (Cây quyết định)**
- **K-Nearest Neighbors (KNN)**
- **Logistic Regression (Hồi quy Logistic)**
- **Support Vector Machine (SVM)**

Mô tả các mô hình

- **Decision Tree:** Mô hình phân loại dựa trên cấu trúc cây, dễ giải thích, cho phép xác định các đặc trưng quan trọng. Tham số chính: `max_depth` (độ sâu tối đa của cây).
- **KNN:** Phân loại dựa trên số lượng láng giềng gần nhất. Tham số chính: `n_neighbors` (số láng giềng).
- **Logistic Regression:** Mô hình tuyến tính cho bài toán phân loại nhị phân. Tham số chính: `C` (hệ số điều chỉnh regularization).
- **SVM:** Tìm siêu phẳng phân tách tối ưu giữa hai lớp. Tham số chính: `C` (điều chỉnh regularization), `gamma` (tham số kernel).

Quy trình huấn luyện và đánh giá

1. Chia dữ liệu thành tập huấn luyện (80%) và kiểm thử (20%).
2. Sử dụng **GridSearchCV** để tìm bộ tham số tối ưu cho từng mô hình với 5-fold cross-validation.

3. Đánh giá mô hình trên tập kiểm thử bằng các chỉ số: accuracy, precision, recall, F1-score.

4. So sánh kết quả giữa các mô hình để chọn ra phương pháp phù hợp nhất.

Kết quả chi tiết và so sánh hiệu quả các mô hình sẽ được trình bày ở phần tiếp theo.

extbfVí dụ code huấn luyện và đánh giá:

```
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report

param_grid = {'max_depth': [3, 5, 7, 12, None]}
grid = GridSearchCV(DecisionTreeClassifier(), param_grid, cv=5,
    scoring='accuracy')
grid.fit(X_train, y_train)
y_pred = grid.best_estimator_.predict(X_test)
print(classification_report(y_test, y_pred))
```

5 Kết quả và so sánh các mô hình

Tất cả các chỉ số dưới đây đều tính trên tập kiểm thử (20% dữ liệu), sau khi tiền xử lý và tối ưu tham số bằng GridSearchCV trên tập huấn luyện. Các chỉ số chính gồm: accuracy (độ chính xác), recall (độ nhạy), specificity (độ đặc hiệu), với công thức:

- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$
- **Recall** = $\frac{TP}{TP+FN}$
- **Specificity** = $\frac{TN}{TN+FP}$

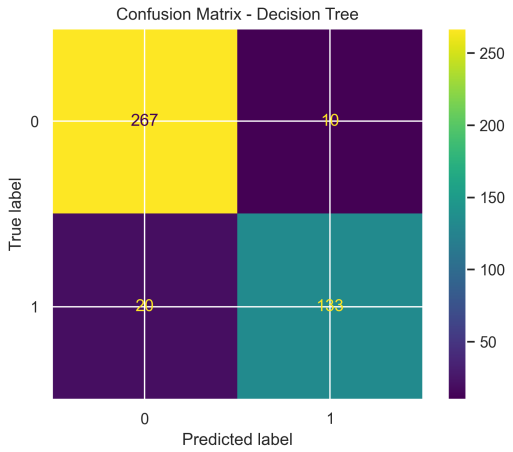
TP: dự đoán đúng người mắc, TN: đúng người không mắc, FP: dự đoán nhầm mắc, FN: dự đoán nhầm không mắc.

5.1 Decision Tree

Mô hình Cây quyết định đạt:

- TP = 133, TN = 267, FP = 10, FN = 20
- Accuracy = $\frac{133+267}{133+267+10+20} = \frac{400}{430} = 0.930$
- Recall = $\frac{133}{133+20} = \frac{133}{153} = 0.87$
- Specificity = $\frac{267}{267+10} = \frac{267}{277} = 0.96$

Ma trận nhầm lẫn dưới đây cho thấy số lượng dự đoán đúng/sai của mô hình trên tập kiểm thử (TP: đúng người mắc, TN: đúng người không mắc, FP: nhầm mắc, FN: nhầm không mắc).



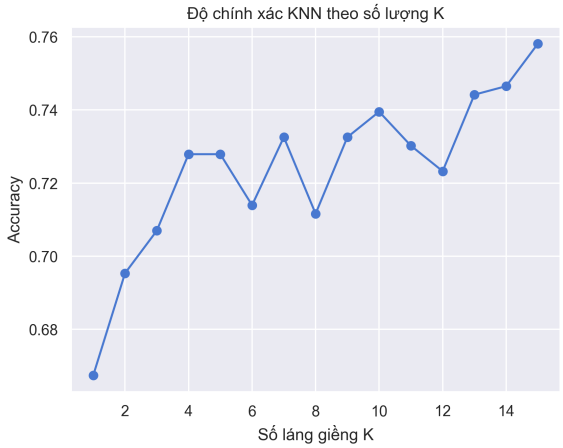
Hình 5: Ma trận nhầm lẫn Decision Tree

5.2 K-Nearest Neighbors (KNN)

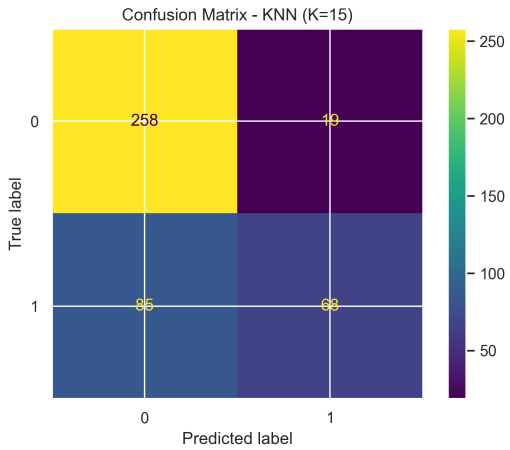
KNN cho thấy accuracy thay đổi theo số lượng láng giềng K. Khi K=15, mô hình đạt:

- TP = 68, TN = 258, FP = 19, FN = 85
- Accuracy = $\frac{68+258}{68+258+19+85} = \frac{326}{430} = 0.758$
- Recall = $\frac{68}{68+85} = \frac{68}{153} = 0.44$
- Specificity = $\frac{258}{258+19} = \frac{258}{277} = 0.93$

Biểu đồ dưới đây minh họa sự thay đổi accuracy theo từng giá trị K. Ma trận nhầm lẫn cho thấy KNN ưu tiên dự đoán đúng người không mắc bệnh (specificity cao), nhưng khả năng nhận diện đúng người mắc bệnh còn hạn chế (recall thấp).



Hình 6: Độ chính xác KNN theo số lượng láng giềng K



Hình 7: Ma trận nhầm lẫn KNN (K=15)

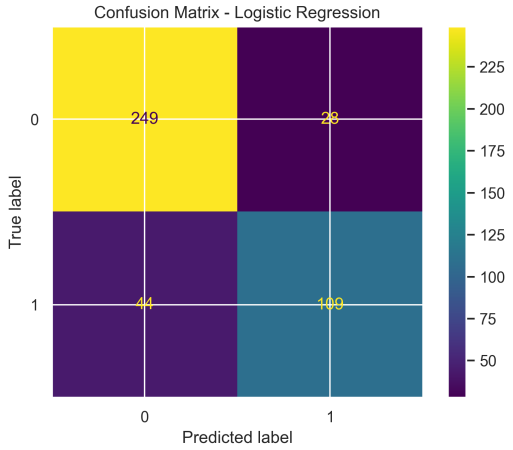
5.3 Logistic Regression

Logistic Regression đạt:

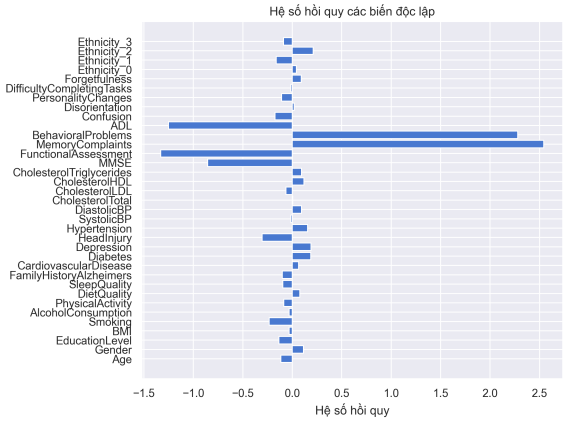
- TP = 109, TN = 249, FP = 28, FN = 44
- Accuracy = $\frac{109+249}{109+249+28+44} = \frac{358}{430} = 0.833$

- Recall $= \frac{109}{109+44} = \frac{109}{153} = 0.71$
- Specificity $= \frac{249}{249+28} = \frac{249}{277} = 0.90$

Ma trận nhầm lẫn cho thấy mô hình cân bằng giữa hai nhóm. Biểu đồ hệ số hồi quy giúp giải thích tác động của từng biến: hệ số dương làm tăng xác suất mắc bệnh, hệ số âm làm giảm xác suất. Các yếu tố lâm sàng và khả năng thực hiện hoạt động thường ngày ảnh hưởng mạnh nhất đến dự đoán.



Hình 8: Ma trận nhầm lẫn Logistic Regression



Hình 9: Hệ số hồi quy các biến trong Logistic Regression

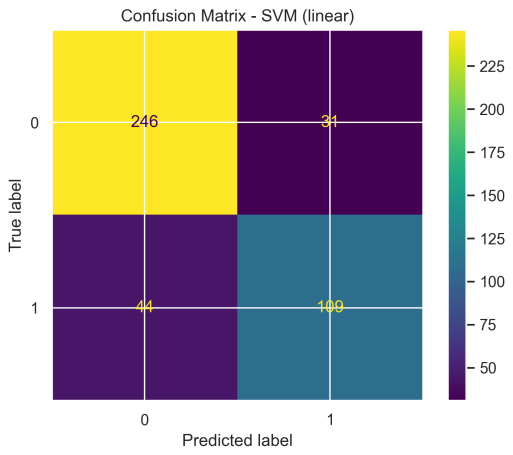
5.4 Support Vector Machine (SVM)

SVM với kernel RBF và tuyến tính đạt:

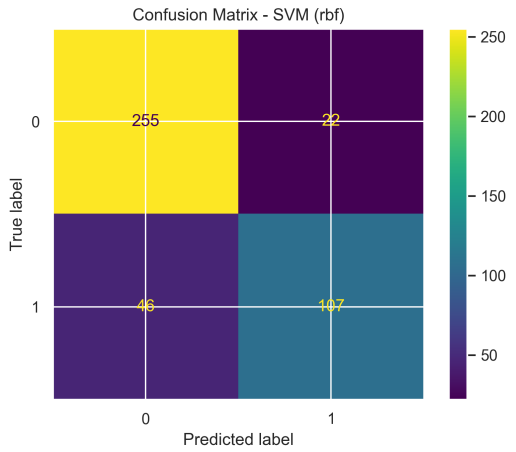
- **Linear kernel:** TP = 109, TN = 246, FP = 31, FN = 44
  Accuracy = $\frac{109+246}{109+246+31+44} = \frac{355}{430} = 0.826$
  Recall = $\frac{109}{109+44} = \frac{109}{153} = 0.71$
  Specificity = $\frac{246}{246+31} = \frac{246}{277} = 0.89$

- **RBF kernel:** TP = 107, TN = 255, FP = 22, FN = 46
  Accuracy = $\frac{107+255}{107+255+22+46} = \frac{362}{430} = 0.842$
  Recall = $\frac{107}{107+46} = \frac{107}{153} = 0.70$
  Specificity = $\frac{255}{255+22} = \frac{255}{277} = 0.92$

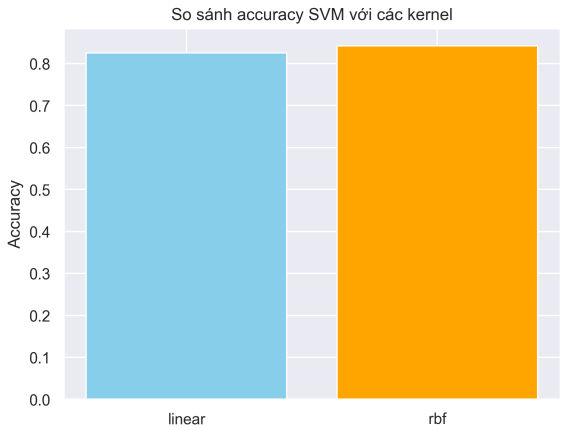
Ma trận nhầm lẫn dưới đây cho thấy sự khác biệt giữa các kernel. Biểu đồ cột minh họa sự khác biệt accuracy giữa các kernel.



Hình 10: Ma trận nhầm lẫn SVM (Linear)



Hình 11: Ma trận nhầm lẫn SVM (RBF)



Hình 12: So sánh accuracy SVM với các kernel

5.5 Tóm tắt kết quả và so sánh

Bảng dưới đây tổng hợp các chỉ số chính trên tập kiểm thử:

extbfMô hình	**Accuracy**	**Recall**	**Specificity**
Decision Tree	0.930	0.87	0.96
KNN (K=15)	0.758	0.44	0.93
Logistic Regression	0.833	0.71	0.90
SVM (Linear)	0.826	0.71	0.89
SVM (RBF)	0.842	0.70	0.92

Bảng 2: Bảng tổng hợp kết quả các mô hình trên tập kiểm thử

extbfKết luận: Decision Tree vượt trội về accuracy, recall và specificity, phù hợp nhất cho bài toán này. SVM (RBF) và Logistic Regression cũng là lựa chọn tốt khi cần cân bằng giữa các chỉ số hoặc giải thích mô hình. KNN chỉ phù hợp nếu ưu tiên specificity, nhưng recall thấp nên hạn chế dùng khi cần phát hiện bệnh nhân mắc Alzheimer.

6 Code và Bộ dữ liệu

7 Tài liệu tham khảo

[1] Marcel F D'Eon (2023), The overcrowded curriculum is alarming. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC10500406/