

# **Học phần: Học máy**

**ĐỀ 9: XÂY DỰNG MÔ HÌNH NAIVE  
BAYES ĐÁNH GIÁ ĐIỂM TÍN DỤNG**

**Nhóm 11**



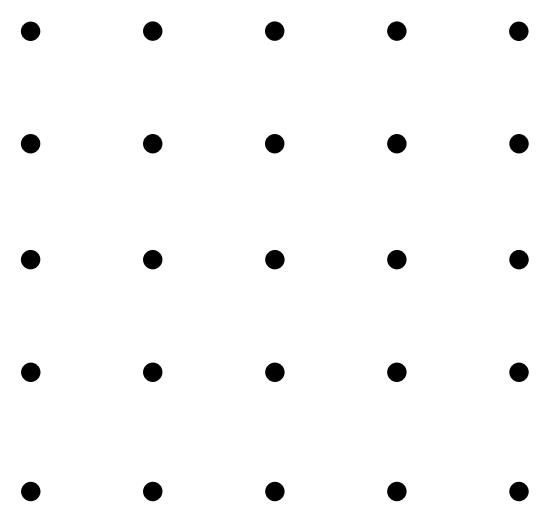
# **Thành viên nhóm**

Nguyễn Văn Chuyên - MSV:1671020047

Nguyễn Đức Hoàn - MSV:1671020124

Mai Đức Hòa - MSV:1671020121

Trịnh Kế Quang - MSV:1671020260



# Tổng quan về mô hình Naive Bayes

MÔ HÌNH NAIVE BAYES LÀ MỘT LOẠI PHÂN LOẠI  
PHƯƠNG PHÁP TRONG MÁY HỌC DỰA TRÊN ĐỊNH  
NGHĨA BAYES VÀ GIẢ ĐỊNH RẰNG CÁC (TÍNH NĂNG)  
CỤ THỂ CỦA DỮ LIỆU ĐỘC LẬP KẾT HỢP VỚI NHAU.

# Chương 1: Giới thiệu

## 1. Nguyên lý hoạt động

Naive Bayes dựa trên định lý Bayes, mô tả mối quan hệ giữa xác suất của một sự kiện với các yếu tố liên quan.

Công thức cơ bản của định lý Bayes là:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Trong đó:

- $P(C|X)$ : Xác suất của lớp  $C$  khi biết đặc trưng  $X$ .
- $P(X|C)$ : Xác suất của đặc trưng  $X$  khi biết lớp  $C$ .
- $P(C)$ : Xác suất của lớp  $C$ .
- $P(X)$ : Xác suất của đặc trưng  $X$ .

# Chương 1: Giới thiệu

- CÁC LOẠI NAIVE BAYES:
- Có một số biến thể của thuật toán Naive Bayes, bao gồm:
  - Gaussian Naive Bayes: Dùng cho dữ liệu liên tục, giả định rằng các đặc trưng tuân theo phân phối Gaussian.
  - Multinomial Naive Bayes: Thích hợp cho dữ liệu phân loại, thường được sử dụng trong phân loại văn bản.
  - Bernoulli Naive Bayes: Sử dụng cho dữ liệu nhị phân, nơi các đặc trưng có thể có giá trị 0 hoặc 1.

# Chương 1: Giới thiệu

## 1.2. Giới thiệu về bài NAIVE BAYES trong đánh giá điểm tín dụng:

Bài toán đánh giá điểm tín dụng là một ứng dụng quan trọng trong lĩnh vực tài chính, giúp các tổ chức ngân hàng và tín dụng xác định khả năng trả nợ của cá nhân hoặc doanh nghiệp. Thuật toán Naive Bayes có thể được áp dụng để phân loại và dự đoán rủi ro tín dụng dựa trên các đặc trưng của người vay, v.v.

GAUSSIAN  
**NAIVE BAYES**  
CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \times P(\text{class})}{P(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon

# Chương 1: Giới thiệu

## 1.2.1 Các đặc trưng (features) trong dữ liệu:

Trong bài toán đánh giá điểm tín dụng, dữ liệu đầu vào bao gồm nhiều đặc trưng (features). Các đặc trưng này được mô tả như sau:

- **Annual Income :** Thu nhập hàng năm.
- **Num of Loan:** Số lượng khoản vay đang có.
- **Num of Delayed Payment:** Số lần thanh toán trễ.
- **Changed Credit Limit** Thay đổi giới hạn tín dụng.
- **Num Credit Inquiries:** Số lần tra cứu tín dụng.
- **Outstanding Debt:** Nợ hiện tại.
- **Amount Invested Monthly:** Số tiền đầu tư hàng tháng.
- **Monthly Balance:** Số dư hàng tháng.
- **Credit History Age:** Tuổi tín dụng (chuyển đổi từ năm/tháng sang ngày).
- **Credit Mix:** Tỷ lệ tín dụng tiêu dùng khác nhau.
- **Payment Behavior:** Hành vi thanh toán.
- **Payment of Min Amount:** Thanh toán số tiền tối thiểu.

# Chương 1: Giới thiệu

## 1.2.2 Bước để xây dựng và triển khai một mô hình học máy Naive Bayes

**1. Mục tiêu của bài toán:** Mục tiêu chính là xây dựng một mô hình phân loại để dự đoán điểm tín dụng của khách hàng. Điểm tín dụng thường thể hiện khả năng trả nợ của một cá nhân, ảnh hưởng đến quyết định cho vay của ngân hàng.

### 2. Tiền xử lý dữ liệu

- a) Xử lý thiếu dữ liệu: Sử dụng SimpleImputer để thay thế giá trị thiếu bằng giá trị trung bình của từng cột.
- b) Mã hóa biến phân loại: Chuyển đổi các biến phân loại thành dạng số để có thể sử dụng trong mô hình học máy.
- c) Làm sạch dữ liệu: Loại bỏ các giá trị không hợp lệ hoặc không cần thiết trong tập dữ liệu.

### 3. Phân chia dữ liệu

**Chia tập dữ liệu:** Dữ liệu được chia thành tập huấn luyện và tập kiểm tra (80/20) để đánh giá độ chính xác của mô hình

### 4. Xây dựng mô hình

**Mô hình Naive Bayes:** Sử dụng thuật toán Gaussian Naive Bayes để xây dựng mô hình phân loại. Mô hình này dựa trên giả định rằng các đặc trưng là độc lập với nhau.

### 5. Đánh giá mô hình

**Báo cáo phân loại:** Cung cấp các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu và F1-score để đánh giá hiệu suất mô hình.

**Ma trận nhầm lẫn:** Cho phép phân tích số lượng dự đoán đúng và sai của từng lớp điểm tín dụng.

# Chương 1: Giới thiệu

## 1.2.3 Ưu điểm NAIVE BAYES trong đánh giá điểm tín dụng

- **Đơn giản và dễ hiểu:** Cấu trúc của thuật toán rất dễ hiểu và triển khai. Điều này giúp người dùng nhanh chóng nắm bắt và áp dụng trong thực tế.
- **Tính toán nhanh chóng :** Naive Bayes có khả năng xử lý các tập dữ liệu lớn một cách nhanh chóng. Điều này rất quan trọng trong môi trường ngân hàng, nơi cần đưa ra quyết định kịp thời.
- **Hiệu quả với ít dữ liệu :** Thuật toán này có thể hoạt động tốt ngay cả khi có ít dữ liệu huấn luyện. Điều này hữu ích trong các trường hợp mới, khi dữ liệu về khách hàng chưa đầy đủ.
- **Khả năng phân loại tốt:** Naive Bayes thường cho kết quả phân loại chính xác, đặc biệt trong các bài toán phân loại văn bản và đánh giá khả năng trả nợ.

# Chương 1: Giới thiệu

## 1.2.3 Ưu điểm NAIVE BAYES trong đánh giá điểm tín dụng

- **Không yêu cầu dữ liệu đặc biệt:** Thuật toán không yêu cầu các giả định phức tạp về phân phối của dữ liệu, điều này giúp dễ dàng áp dụng cho nhiều loại dữ liệu khác nhau.
- **Khả năng giải thích:** Kết quả của Naive Bayes có thể dễ dàng giải thích, giúp các nhà quản lý và chuyên gia tài chính hiểu rõ hơn về lý do tại sao một khách hàng được phân loại vào nhóm rủi ro cao hay thấp.
- **Phát hiện gian lận :** Naive Bayes có thể được sử dụng để phát hiện các hành vi gian lận trong quá trình cho vay, nhờ vào khả năng phân tích các đặc trưng khác nhau của giao dịch.
- 
- **Thích hợp cho dữ liệu nhị phân và định lượng.**

# Chương 1: Giới thiệu

## 1.2.4 Nhược điểm của NAIVE BAYES trong đánh giá điểm tín dụng

- **Giả Định Độc Lập:** Naive Bayes giả định rằng các đặc trưng là độc lập với nhau trong từng lớp. Trong thực tế, nhiều đặc trưng có thể có mối quan hệ phụ thuộc, điều này có thể làm giảm độ chính xác của mô hình.
- **Kém Hiệu Quả với Dữ Liệu Có Tương Quan:** Khi các đặc trưng có mối liên hệ mạnh mẽ với nhau, hiệu suất của Naive Bayes có thể giảm. Điều này đặc biệt quan trọng trong lĩnh vực tín dụng, nơi các yếu tố như thu nhập và lịch sử tín dụng có thể liên quan chặt chẽ.
- **Nhạy cảm với dữ liệu thiếu:** Naive Bayes có thể gặp khó khăn khi xử lý dữ liệu thiếu hoặc không đầy đủ. Nếu một đặc trưng nào đó không xuất hiện trong tập huấn luyện, mô hình có thể không thể dự đoán chính xác cho các trường hợp mới.

# Chương 1: Giới thiệu

## 1.2.4 Nhược điểm của NAIVE BAYES trong đánh giá điểm tín dụng

- **Không Thích Hợp cho Dữ Liệu Không Phân Phối:** Nếu dữ liệu không tuân theo phân phối Gaussian (đối với Gaussian Naive Bayes) hoặc không phù hợp với các giả định khác của các loại Naive Bayes, kết quả có thể không chính xác.
- **Khó Khăn Trong Việc Xử Lý Các Tình Huống Hiếm:** Naive Bayes có thể gặp khó khăn trong việc xử lý các trường hợp hiếm gặp, chẳng hạn như khách hàng có lịch sử tín dụng xấu nhưng vẫn có khả năng trả nợ tốt.
- **Thiếu Độ Chính Xác Trong Một Số Tình Huống:** Trong một số tình huống phức tạp, Naive Bayes có thể không đạt được độ chính xác mong muốn, đặc biệt khi so sánh với các mô hình phức tạp hơn như cây quyết định hoặc mạng nơ-ron.

# Chương 2: Phân tích

## Tiền xử lý dữ liệu

Thiếu dữ liệu là một vấn đề phổ biến trong phân tích dữ liệu và học máy. Xử lý thiếu dữ liệu đúng cách là rất quan trọng để đảm bảo độ chính xác và độ tin cậy của mô hình. Dưới đây là phân tích chi tiết về các phương pháp và kỹ thuật xử lý thiếu dữ liệu.

### 1. Tại Sao Thiếu Dữ Liệu Xuất Hiện?

Lỗi trong thu thập dữ liệu: Nhập liệu sai, thiết bị hỏng.

Dữ liệu không có sẵn: Một số thông tin có thể không được cung cấp bởi người dùng.

Dữ liệu không liên quan: Một số trường không áp dụng cho tất cả các đối tượng (ví dụ: tuổi hư).

### 2. Tác Động của Thiếu Dữ Liệu

Giảm hiệu suất mô hình: Thiếu dữ liệu có thể dẫn đến mô hình không chính xác.

Thiên lệch trong kết quả: Nếu không xử lý đúng, có thể dẫn đến các kết luận sai lầm.

# Chương 2: Phân tích

## 3. Các Phương Pháp Xử Lý Thiếu Dữ Liệu

### 3.1. Loại Bỏ Dữ Liệu

**Loại bỏ hàng:** Xóa các hàng có giá trị thiếu. Thích hợp khi tỷ lệ thiếu dữ liệu thấp.

**Loại bỏ cột:** Xóa cột có quá nhiều giá trị thiếu. Thích hợp khi cột không quan trọng đối với phân tích.

### 3.2. Điền Giá Trị Thiếu

**Điền bằng giá trị trung bình:** Sử dụng giá trị trung bình (hoặc trung vị) của cột để thay thế giá trị thiếu. Phù hợp cho dữ liệu số.

# Chương 2: Phân tích

**Điền bằng giá trị trung bình:** Sử dụng giá trị trung bình (hoặc trung vị) của cột để thay thế giá trị thiếu. Phù hợp cho dữ liệu số.

```
from sklearn.impute import SimpleImputer  
imputer = SimpleImputer(strategy='mean')  
data[column] = imputer.fit_transform(data[[column]])
```

**Điền bằng giá trị phổ biến nhất:** Sử dụng giá trị thường gặp nhất cho các biến phân loại.

```
imputer = SimpleImputer(strategy='most_frequent')
```

**Điền bằng giá trị cố định:** Thay thế giá trị thiếu bằng một giá trị cố định, ví dụ 0 hoặc unknown.

# Chương 2: Phân tích

## 3.3. Sử Dụng Mô Hình Dự Đoán

**Dự đoán giá trị thiểu:** Sử dụng mô hình học máy để dự đoán giá trị thiểu dựa trên các đặc trưng khác. Ví dụ, hồi quy có thể được sử dụng để dự đoán giá trị của một cột từ các cột khác.

# Chương 2: Phân tích

## 4. Lợi Ích của Việc Xử Lý Thiếu Dữ Liệu

Cải thiện độ chính xác của mô hình: Xử lý thiếu dữ liệu giúp mô hình hoạt động tốt hơn.

Giảm thiểu thiên lệch: Đảm bảo rằng các kết quả phân tích không bị ảnh hưởng bởi giá trị thiếu.

Tăng cường tính khả thi: Các mô hình có thể áp dụng cho nhiều trường hợp khác nhau.

# Chương 2: Phân tích

## 5. Lưu Ý Khi Xử Lý Thiếu Dữ Liệu

**Phân tích nguyên nhân:** Trước khi quyết định xử lý, cần phân tích nguyên nhân gây ra thiếu dữ liệu để chọn phương pháp phù hợp.

**Tỷ lệ thiếu dữ liệu:** Nếu tỷ lệ thiếu quá cao, có thể cần xem xét lại toàn bộ dữ liệu hoặc thu thập thêm dữ liệu.

**Kiểm tra sau khi xử lý:** Sau khi xử lý, cần kiểm tra lại dữ liệu để đảm bảo rằng các giá trị thay thế không gây ra vấn đề mới.

# Chương 2: Phân tích

## Phân Tích Phân Chia Dữ Liệu

### 1. Mục Đích của Phân Chia Dữ Liệu

**Đánh Giá Mô Hình:** Giúp kiểm tra khả năng tổng quát của mô hình trên dữ liệu chưa thấy, từ đó đánh giá hiệu suất thực sự.

**Ngăn Ngừa Overfitting:** Bằng cách giữ lại một phần dữ liệu cho kiểm tra, giúp mô hình không chỉ học thuộc lòng mà còn học cách tổng quát.

# Chương 2: Phân tích

## 2. Phương Pháp Phân Chia Dữ Liệu

### Phân Chia Ngẫu Nhiên:

**Train-Test Split:** Dữ liệu được chia thành hai tập: **tập huấn luyện (thường 70-80%)** và **tập kiểm tra (20-30%)**

```
x, y = data.iloc[:, :-1], data.iloc[:, -1]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)
x_train.shape, y_train.shape
x_test.shape, y_test.shape
```

### Phân Chia K-Fold:

Dữ liệu được chia thành K phần. Mô hình được huấn luyện K lần, mỗi lần sử dụng một phần để kiểm tra và phần còn lại để huấn luyện. Giúp đánh giá mô hình một cách toàn diện hơn.

### Stratified Split:

Giữ nguyên tỷ lệ của các lớp trong dữ liệu. Phương pháp này rất quan trọng cho các bài toán phân loại không cân bằng.

# Chương 2: Phân tích

## 3. Lợi Ích của Phân Chia Dữ Liệu

**Đánh Giá Chính Xác:** Giúp có cái nhìn rõ hơn về hiệu suất mô hình trong thực tế.

**Tối Ưu Hóa Mô Hình:** Cho phép thử nghiệm với nhiều cấu hình và tham số mà không làm giảm hiệu suất.

## 4. Lưu Ý Khi Phân Chia Dữ Liệu

**Tính Ngẫu Nhiên:** Đảm bảo rằng việc phân chia là ngẫu nhiên để tránh thiên lệch.

**Kích Thước Tập Kiểm Tra:** Tập kiểm tra cần đủ lớn để có thể đưa ra kết luận chính xác về hiệu suất của mô hình.

**Đặc Tính Dữ Liệu:** Cần xem xét đặc tính của dữ liệu khi phân chia, đặc biệt với dữ liệu không cân bằng.

# Chương 2: Phân tích

## Xây dựng mô hình

### 1. Giới Thiệu về Naive Bayes

**Định Nghĩa:** Naive Bayes là một thuật toán phân loại dựa trên định lý Bayes, giả định rằng các đặc trưng độc lập với nhau.

**Ưu Điểm:**

**Đơn giản và dễ triển khai.**

**Tính toán nhanh, hiệu quả với dữ liệu lớn.**

**Hoạt động tốt với dữ liệu phân loại không cân bằng.**

# Chương 2: Phân tích

## 2. Quy Trình Xây Dựng Mô Hình

**Chuẩn Bị Dữ Liệu:**

**Tiền xử lý dữ liệu:** Làm sạch, xử lý giá trị thiểu, mã hóa biến phân loại.

**Phân Chia Dữ Liệu:**

**Chia thành tập huấn luyện và tập kiểm tra.**

**Huấn Luyện Mô Hình:**

**Sử dụng tập huấn luyện để xây dựng mô hình Naive Bayes.**

**Sử dụng các phiên bản khác nhau như Gaussian Naive Bayes cho dữ liệu  
liên tục hoặc Multinomial Naive Bayes cho dữ liệu rời rạc.**

```
from sklearn.naive_bayes import GaussianNB  
model = GaussianNB()  
model.fit(x_train, y_train)
```

# Chương 2: Phân tích

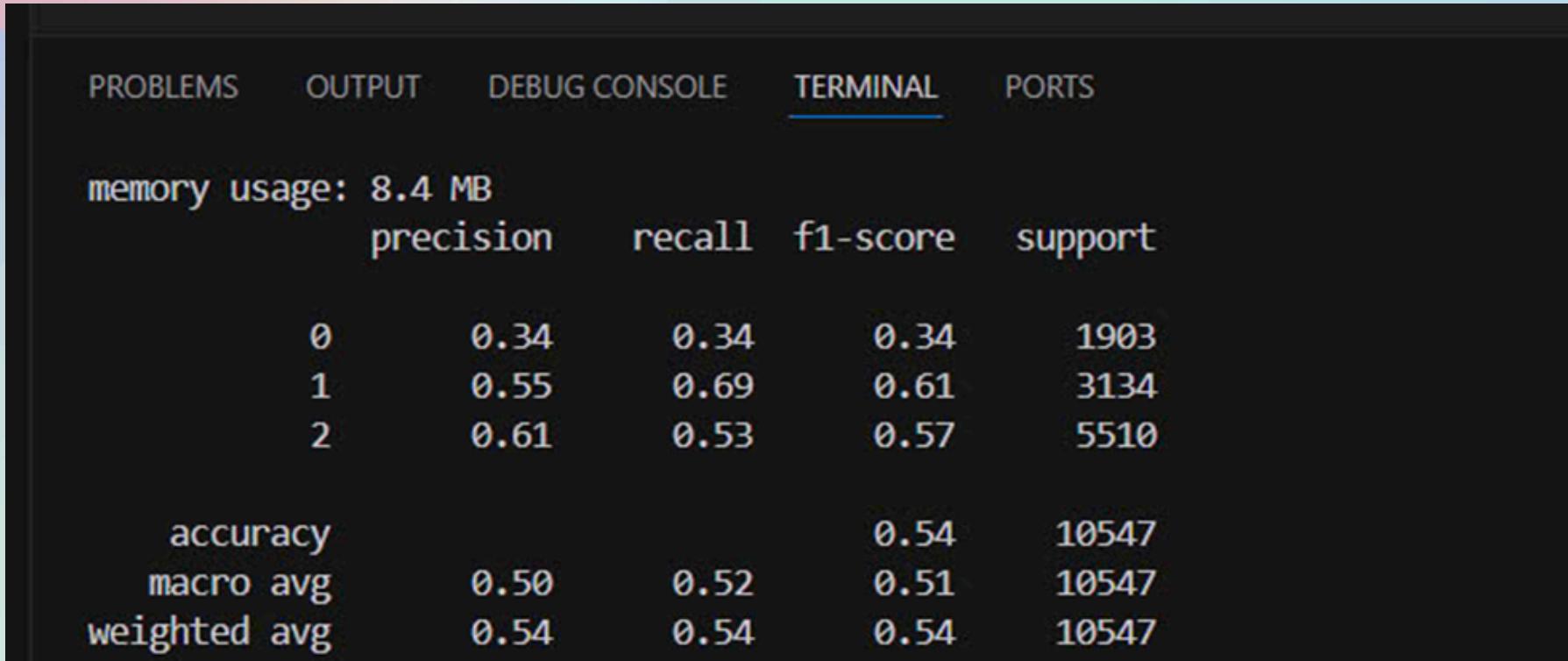
## Dự Đoán:

Áp dụng mô hình vào tập kiểm tra để dự đoán nhãn.

## Đánh Giá Mô Hình:

Sử dụng các chỉ số như độ chính xác, precision, recall, và F1-score để đánh giá hiệu suất.

```
from sklearn.metrics import classification_report  
print(classification_report(y_test, y_pred))
```



The screenshot shows a terminal window with the following classification report output:

	precision	recall	f1-score	support
0	0.34	0.34	0.34	1903
1	0.55	0.69	0.61	3134
2	0.61	0.53	0.57	5510
accuracy			0.54	10547
macro avg	0.50	0.52	0.51	10547
weighted avg	0.54	0.54	0.54	10547

## Lợi Ích

Hiệu suất tốt với các tập dữ liệu lớn.

Đơn giản trong việc giải thích và triển khai.

## Lưu Ý

Giả định độc lập giữa các đặc trưng có thể không luôn đúng, nhưng thuật toán vẫn hoạt động tốt trong thực tế.

Cần kiểm tra kỹ lưỡng để đảm bảo mô hình phù hợp với dữ liệu.

# Kết Luận

**Bài toán đánh giá điểm tín dụng không chỉ đơn thuần là một bài toán phân loại, mà còn là một phần quan trọng trong việc xây dựng hệ thống tài chính bền vững. Việc áp dụng công nghệ học máy vào bài toán này không chỉ nâng cao độ chính xác trong việc dự đoán khả năng trả nợ mà còn giúp các tổ chức tài chính tối ưu hóa quy trình cho vay, từ đó tăng cường sự công bằng và giảm thiểu rủi ro.**