**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: HỌC MÁY**

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

**Giáo viên hướng dẫn: GV. TRẦN HỒNG VIỆT**

**Sinh viên thực hiện: Nhóm 11**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stt** | **Mã sv** | **Họ và tên** | **Lớp** |
| **1** | **1671020047** | **Nguyễn Văn Chuyện** | **CNTT 16-01** |
| 2 | **1671020121** | **Mai Đức Hòa** | **CNTT 16-01** |
| 3 | **1671020124** | **Nguyễn Đức Hoàn** | **CNTT 16-01** |
| 4 | **1671020260** | **Trịnh Kế Quang** | **CNTT 16-01** |

**Hà Nội, năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: HỌC MÁY**

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Mã sinh viên** | **Họ và Tên** | **Ngày Sinh** | **Điểm** | |
| **Bằng Số** | **Bằng Chữ** |
| **1** | **1671020047** | **Nguyễn Văn Chuyện** | **28/08/2004** |  |  |
| **2** | **1671020121** | **Mai Đức Hòa** | **08/04/2004** |  |  |
| **3** | **1671020124** | **Nguyễn Đức Hoàn** | **02/05/2004** |  |  |
| **5** | **1671020260** | **Trịnh Kế Quang** | **01/03/2004** |  |  |

**CÁN BỘ CHẤM THI 1 CÁN BỘ CHẤM THI 2**

**Hà Nội, năm 2024**

# MỤC LỤC

[MỞ ĐẦU 5](#_Toc179561559)

[CHƯƠNG 1 . GIỚI THIỆU 6](#_Toc179561560)

[1.1 Giới thiệu về thuật toán Naive Bayes trong Machine Learning: 6](#_Toc179561561)

[1.2 Giới thiệu về bài toán Naive Bayes trong đánh giá điểm tín dụng: 6](#_Toc179561562)

[1.2.1 Các đặc trưng (features) trong dữ liệu: 8](#_Toc179561563)

[1.2.2 Bước để xây dựng và triển khai một mô hình học máy Naive Bayes: 8](#_Toc179561564)

[1.2.3 Ưu điểm của Naive Bayes trong đánh giá tín dụng điểm 9](#_Toc179561565)

[1.2.4 Nhược điểm của Naive Bayes trong đánh giá tín dụng điểm 10](#_Toc179561566)

[CHƯƠNG 2 . PHÂN TÍCH 12](#_Toc179561567)

[2.1 Tiền xử lý dữ liệu: 12](#_Toc179561568)

[2.2 Phân chia dữ liệu: 13](#_Toc179561569)

2.3 Xây dựng mô hình.....................................................................................................15

[2.4 Đánh giá mô hình:.....................................................................................................16](#_Toc179561571)

Kết Luận..................................................................................................................................18

DANH MỤC TÀI LIỆU THAM KHẢO................................................................................20

# MỞ ĐẦU

Trong bối cảnh công nghệ thông tin phát triển mạnh mẽ, học máy (Machine Learning) đã trở thành một lĩnh vực then chốt, có khả năng cách mạng hóa nhiều ngành nghề và lĩnh vực. Học máy, một nhánh của trí tuệ nhân tạo, cho phép máy tính tự động học hỏi và cải thiện hiệu suất từ dữ liệu mà không cần phải được lập trình cụ thể cho từng nhiệm vụ. Bài toán đánh giá điểm tín dụng (credit scoring) là một vấn đề quan trọng trong lĩnh vực tài chính, nhằm xác định khả năng trả nợ của cá nhân hoặc doanh nghiệp

Thuật toán Naive Bayes là một phương pháp phân loại đơn giản nhưng rất hiệu quả, dựa trên định lý Bayes. Nó được gọi là "naive" (ngây thơ) vì nó giả định rằng các đặc trưng (features) trong dữ liệu đầu vào là độc lập với nhau. Trong bài toán đánh giá điểm tín dụng, thuật toán Naive Bayes thường được sử dụng để phân loại và dự đoán khả năng thanh toán tín dụng của một cá nhân hoặc doanh nghiệp.

Học phần này sẽ cung cấp cho học viên kiến thức cơ bản về thuật toán NAIVE BAYES, các bước triển khai NAIVE BAYES để xây dựng mô hình đánh giá và dự đoán khả năng thanh toán tín dụng của một cá nhân hoặc một doanh nghiệp. Qua đó, học viên không chỉ nắm vững lý thuyết mà còn giúp học viên hiểu về xác suất, khả năng phân loại, kỹ năng lập trình tư duy phản biện và ứng dụng thực tiễn.

# . GIỚI THIỆU

## Giới thiệu về thuật toán Naive Bayes trong Machine Learning:

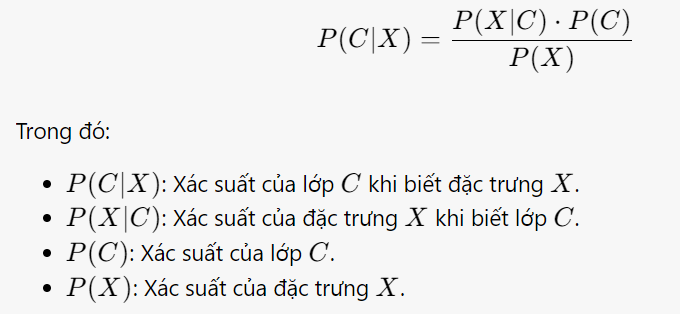
**Naive Bayes**

Thuật toán **Naive Bayes** là một trong những kỹ thuật phân loại phổ biến trong lĩnh vực machine learning, đặc biệt trong phân loại văn bản và phân tích dữ liệu

**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:

****

**Các loại Naive Bayes:**

Có một số biến thể của thuật toán Naive Bayes, bao gồm:

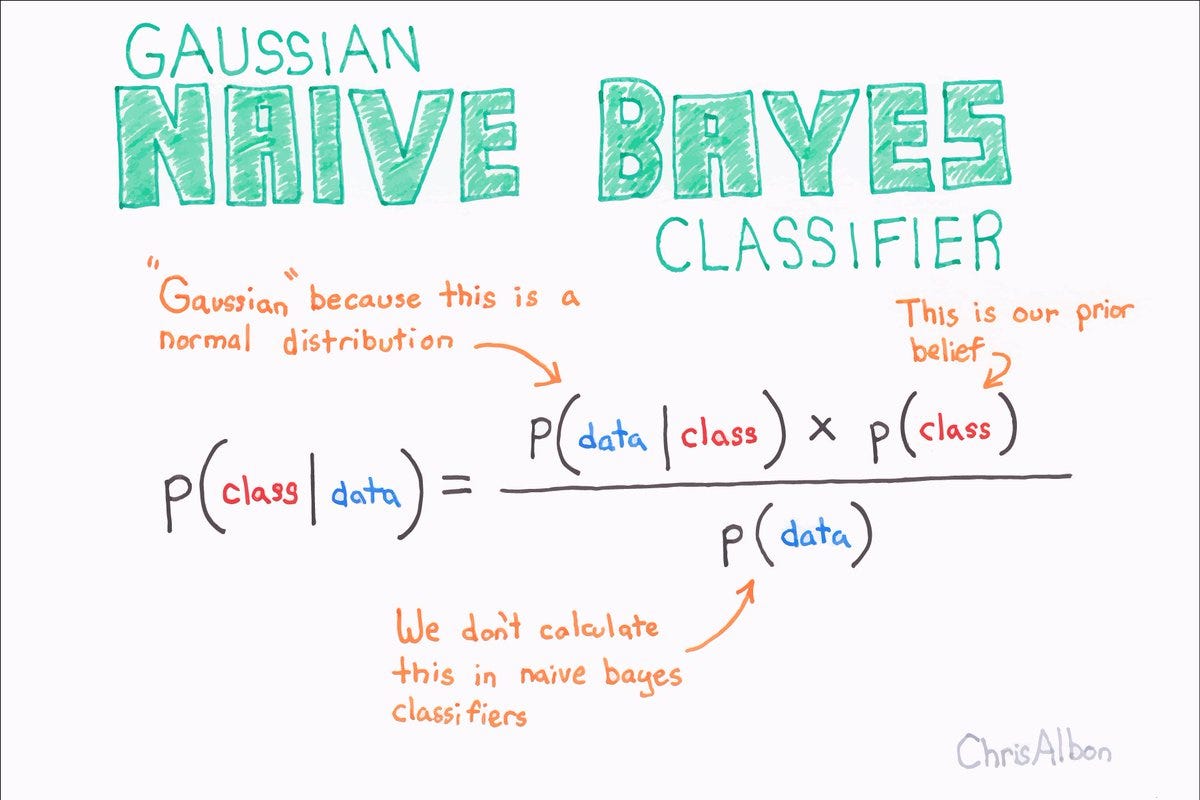
**Gaussian Naive Bayes:** Dùng cho dữ liệu liên tục, giả định rằng các đặc trưng tuân theo phân phối Gaussian.

**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.

## 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.



Trong bài toán đánh giá điểm tín dụng, ngôn ngữ lập trình **Python** được sử dụng rộng rãi nhờ vào sự linh hoạt, dễ đọc và cộng đồng hỗ trợ mạnh mẽ. Python cung cấp một hệ sinh thái phong phú các thư viện giúp phát triển và triển khai các mô hình máy học một cách hiệu quả.

Các thư viện chính được sử dụng trong bài toán này bao gồm:

1. **Pandas**: Thư viện này là công cụ chính để quản lývà thao tác dữ liệu. Với pandas, người dùng có thể dễ dàng đọc, ghi và xử lý dữ liệu từ các tệp CSV, Excel hoặc SQL. Cấu trúc dữ liệu DataFrame cho phép tổ chức và phân tích dữ liệu một cách trực quan.
2. **NumPy**: Đây là thư viện cốt lõi cho tính toán số học trong Python. NumPy hỗ trợ việc xử lý các mảng (array) đa chiều và cung cấp các hàm toán học mạnh mẽ cho việc thực hiện các phép toán phức tạp một cách nhanh chóng và hiệu quả.
3. **SimpleImputer**: là một lớp trong thư viện scikit-learn được sử dụng để xử lý các giá trị thiếu (missing values) trong tập dữ liệu. Việc xử lý giá trị thiếu là rất quan trọng trong tiền xử lý dữ liệu để đảm bảo rằng các mô hình học máy hoạt động hiệu quả.
4. **LabelEncoder**: là một lớp trong thư viện scikit-learn được sử dụng để chuyển đổi các nhãn (labels) phân loại thành các số nguyên tương ứng. Điều này rất hữu ích trong việc tiền xử lý dữ liệu, đặc biệt là khi làm việc với các mô hình học máy yêu cầu đầu vào là dữ liệu số.
5. **GaussianNB**: Là một lớp trong thư viện scikit-learn sử dụng cho thuật toán Naive Bayes, đặc biệt phù hợp cho các bài toán phân loại với dữ liệu có các đặc trưng liên tục mà giả định rằng các đặc trưng này tuân theo phân phối Gaussian (phân phối chuẩn).
6. **Classification\_report**: Là một hàm trong thư viện scikit-learn được sử dụng để đánh giá hiệu suất của một mô hình phân loại. Nó cung cấp một cái nhìn tổng quan về các chỉ số chính như độ chính xác, độ nhạy (recall), độ chính xác (precision), và điểm F1 (F1 score) cho từng lớp trong bài toán phân loại.
7. **Matplotlib:** Là một thư viện mạnh mẽ và phổ biến trong Python được sử dụng để tạo các biểu đồ và đồ thị 2D. Nó cung cấp nhiều công cụ và tùy chọn cho việc trực quan hóa dữ liệu, giúp người dùng dễ dàng tạo ra các biểu đồ từ đơn giản đến phức tạp.
8. **Seaborn:** là một thư viện trực quan hóa dữ liệu mạnh mẽ dựa trên Matplotlib, cung cấp giao diện đơn giản để tạo ra các biểu đồ đẹp mắt và thông tin. Seaborn được thiết kế để làm việc tốt với các cấu trúc dữ liệu như DataFrame của Pandas, giúp việc trực quan hóa dữ liệu trở nên dễ dàng hơn.

### 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.

### 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**
3. **Xử lý thiếu dữ liệu**: Sử dụng SimpleImputer để thay thế giá trị thiếu bằng giá trị trung bình của từng cột.
4. **Mã hóa biến phân loại**: Chuyển đổi các biến phân loại thành dạng số để có thể sử dụng trong mô hình học máy.
5. **Làm sạch dữ liệu**: Loại bỏ các giá trị không hợp lệ hoặc không cần thiết trong tập dữ liệu.
6. **Phân chia dữ liệu**
7. **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.
8. **Xây dựng mô hình**
9. **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.
10. **Đánh giá mô hình**
11. **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.
12. **Ma trận nhầm lẫn**: Cho phép phân tích số lượng dự đoán đúng và sai của từng lớp điểm tín dụng.

### 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 tróng** : 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.
* **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** :

### 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
* **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.

Tóm lại:

Thuật toán **Naive Bayes** là một công cụ mạnh mẽ và dễ sử dụng trong lĩnh vực đánh giá điểm tín dụng. Với khả năng phân loại nhanh chóng và hiệu quả, nó giúp các tổ chức tài chính đưa ra quyết định kịp thời về khả năng trả nợ của khách hàng.

**Ưu điểm** nổi bật của **Naive Bayes** bao gồm tính đơn giản, khả năng xử lý dữ liệu lớn, và hiệu quả ngay cả với ít dữ liệu huấn luyện. Tuy nhiên, thuật toán này cũng có những **nhược điểm**, đặc biệt là giả định độc lập giữa các đặc trưng, điều này có thể dẫn đến sai lệch trong dự đoán khi các đặc trưng có mối liên hệ chặt chẽ.

Do đó, trong thực tiễn, **Naive Bayes** cần được sử dụng một cách cân nhắc, có thể kết hợp với các phương pháp khác hoặc áp dụng các kỹ thuật tiền xử lý để tối ưu hóa hiệu suất. Khi được triển khai đúng cách, **Naive Bayes** có thể đóng góp tích cực vào việc đánh giá rủi ro tín dụng, giúp các tổ chức tài chính giảm thiểu rủi ro và nâng cao hiệu quả trong hoạt động cho vay.

# . 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 hưu).

**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.

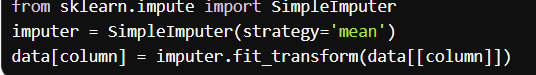
**3. Các Phương Pháp Xử Lý Thiếu Dữ Liệu**

**3.1. Loại Bỏ Dữ Liệu**

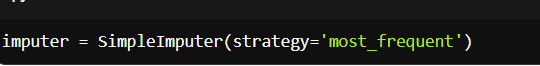
Loại bỏ hàng: Xóa các hàng có giá trị thiếu. Thích hợp khi tỷ lệ thiếu dữ liệu thấp.

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. 

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



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

**3.3. Sử Dụng Mô Hình Dự Đoán**

Dự đoán giá trị thiếu: Sử dụng mô hình học máy để dự đoán giá trị thiếu 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.

**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.

**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.

**Kết Luận**

Xử lý thiếu dữ liệu là một bước quan trọng trong quy trình tiền xử lý dữ liệu. Sử dụng phương pháp phù hợp sẽ giúp nâng cao chất lượng dữ liệu và cải thiện hiệu suất của mô hình học máy. Việc lựa chọn phương pháp cần dựa trên tính chất của dữ liệu và bối cảnh cụ thể của bài toán.

## Phân chia dữ liệu

**Phân Tích Phân Chia Dữ Liệu**

Phân chia dữ liệu là một bước quan trọng trong quy trình xây dựng mô hình học máy. Dưới đây là phân tích chi tiết về mục đích, phương pháp và lưu ý khi thực hiện phân chia dữ liệu.

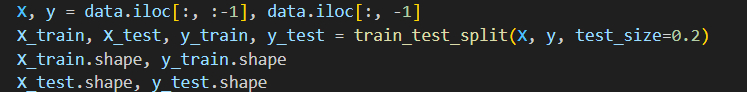
**1. Mục Đích của Phân Chia Dữ Liệu**

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.

**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%). 

**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.

**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.

**Kết Luận**

Phân chia dữ liệu là bước thiết yếu trong quy trình học máy, giúp đảm bảo rằng mô hình hoạt động tốt trên dữ liệu chưa thấy. Việc lựa chọn phương pháp phân chia phù hợp sẽ cải thiện khả năng tổng quát của mô hình và tăng độ tin cậy trong các dự đoán.

## 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.

**2. Quy Trình Xây Dựng Mô Hình**

**Chuẩn Bị Dữ Liệu:**

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

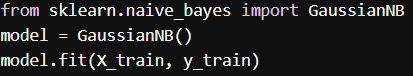
**Phân Chia Dữ Liệu:**

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.



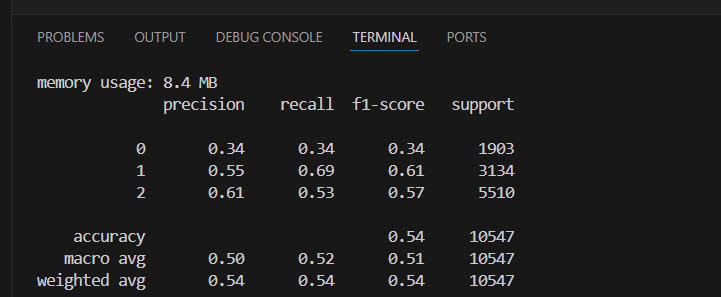
**Dự Đoán:**

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.





**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**

Naive Bayes là một thuật toán mạnh mẽ cho các bài toán phân loại, đặc biệt khi dữ liệu lớn và không cân bằng. Việc xây dựng mô hình với Naive Bayes nhanh chóng và hiệu quả, giúp tạo ra các dự đoán chính xác trong nhiều ứng dụng khác nhau.

# KẾT LUẬN

Bài toán đánh giá điểm tín dụng là một trong những ứng dụng quan trọng nhất của học máy trong lĩnh vực tài chính. Dưới đây là các điểm chính cần lưu ý:

1. Mục Đích và Ý Nghĩa

Quản lý rủi ro: Giúp các tổ chức tài chính xác định khả năng trả nợ của khách hàng, từ đó giảm thiểu rủi ro tín dụng.

Ra quyết định thông minh: Cung cấp thông tin hỗ trợ cho các quyết định cho vay, giúp tối ưu hóa quy trình cấp tín dụng.

2. Dữ Liệu và Đặc Trưng

Đặc trưng đa dạng: Dữ liệu có thể bao gồm thông tin cá nhân, thông tin tài chính và lịch sử tín dụng của khách hàng. Việc chọn lọc và xử lý dữ liệu là rất quan trọng để đảm bảo tính chính xác của mô hình.

Giá trị thiếu và tiền xử lý: Cần xử lý các giá trị thiếu và thực hiện các bước tiền xử lý để cải thiện chất lượng dữ liệu.

3. Mô Hình và Kỹ Thuật

Sử dụng thuật toán học máy: Các mô hình như Naive Bayes, hồi quy logistic, cây quyết định và mạng nơ-ron đều có thể được áp dụng. Việc lựa chọn mô hình phụ thuộc vào tính chất của dữ liệu và yêu cầu cụ thể của bài toán.

Đánh giá hiệu suất: Sử dụng các chỉ số như độ chính xác, precision, recall và F1 score để đánh giá hiệu suất của mô hình.

4. Triển Khai và Ứng Dụng

Triển khai trong thực tế: Mô hình cần được triển khai vào hệ thống thực tế để hỗ trợ quy trình ra quyết định. Điều này có thể bao gồm phát triển API hoặc tích hợp vào phần mềm hiện có.

Theo dõi và bảo trì: Cần liên tục theo dõi hiệu suất của mô hình và cập nhật khi cần thiết với dữ liệu mới để duy trì độ chính xác.

5. Thách Thức và Cơ Hội

Thách thức: Dữ liệu không đầy đủ, thiên lệch trong dữ liệu, và sự phức tạp trong việc giải thích các mô hình có thể gây khó khăn trong việc xây dựng một hệ thống đánh giá tín dụng hiệu quả.

Cơ hội: Sử dụng công nghệ học máy có thể cải thiện đáng kể quy trình đánh giá tín dụng, mở ra cơ hội cho các sản phẩm tài chính mới và tăng cường sự công bằng trong việc tiếp cận tín dụng.

**Tóm Lại**

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.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Thuật toán Naive Bayes <https://fitdnu.net/mod/resource/view.php?id=1619>
2. Tài liệu Naive Bayes https://fitdnu.net/mod/resource/view.php?id=1620
3. https://www.kaggle.com/code/shivangamsoni/credit-score-classofication-naive-bayes