**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

-------- 🕮 --------



**BÀI TẬP LỚN MÔN KHAI PHÁ DỮ LIỆU**

***Đề tài:***

**PHÂN LOẠI DỮ LIỆU VỚI MẠNG BAYESIAN**

**(NAÏVE BAYES)**

***Giảng viên hướng dẫn:*  Nguyễn Quốc Tuấn**

***Nhóm sinh viên thực hiện:***

|  |  |
| --- | --- |
| Họ và Tên | Mã sinh viên |
| Khuất Văn Chung |  |
| Trần Quốc Nam |  |
| Nguyễn Dức Dương |  |
| Nguyễn Huy Sơn |  |
| Trịnh Thị Hồng | 181200752 |

**Lớp: Công nghệ thông tin 2- K59**

[LỜI MỞ ĐẦU 3](#_Toc85989129)

[PHẦN 1: TỔNG QUAN VỀ PHÂN LỚP DỮ LIỆU. 4](#_Toc85989130)

[**1.1 Phân lớp dữ liệu là gì?** 4](#_Toc85989131)

[**1.2 Cách tiếp cận chung để phân loại dữ liệu.** 4](#_Toc85989132)

[PHẦN 2: PHƯƠNG PHÁP PHÂN LỚP NAÏVE BAYES (BAYES CLASSIFICATION METHODS). 7](#_Toc85989133)

[**2.1 Giới thiệu.** 7](#_Toc85989134)

[**2.2 Định lý Bayes.** 7](#_Toc85989135)

[**2.3 Phân lớp Naïve Bayes.** 8](#_Toc85989136)

[**2.3.1 Cách hoạt động của phương pháp phân lớp Naïve Bayes.** 8](#_Toc85989137)

[**2.3.2 Ưu điểm và nhược điểm của phương pháp phân lớp Naïve Bayes.** 10](#_Toc85989138)

[**2.4 Một vài ví dụ.** 10](#_Toc85989139)

[PHẦN 3: TỔNG KẾT. 15](#_Toc85989140)

[**3.1 Kết luận.** 15](#_Toc85989141)

[**3.2 Tài liệu tham khảo.** 15](#_Toc85989142)

# **LỜI MỞ ĐẦU**

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin  
trong nhiều lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng  
nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích luỹ  
nhiều lên. Đi cùng với việc này là việc bùng nổ dữ liệu (big data) là một vấn đề hiện nay. Đây có thể được coi là vùng đất mới cho các nhà phát triển công nghệ. Tương lai không xa nữa thì việc khai thác dữ liệu sẽ là điều không thể thiếu trong công việc của từng nhà phát triển. Vì vậy việc khai thác những tinh hoa từ bộ dữ liệu khổng lồ sẽ là miếng bánh mà được các nhà phát triển nhắm tới. Dữ liệu có thể được khai thác theo nhiều cách khác nhau, theo nhiều mục đích khác nhau để phục vụ cho nhiều vấn đề thiết yếu của xã hội, người dùng.. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có  
nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ  
liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai  
thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát  
triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai  
phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Trong suốt quá trình tìm hiểu và thực hiện, ngoài những cố gắng của bản thân, chúng em sẽ không thể hoàn thành tốt được nếu không có sự giúp đỡ, hướng dẫn của thầy Nguyễn Quốc Tuấn. Chúng em xin gửi lời cảm ơn sâu sắc nhất đến thầy.

Trong quá trình tìm hiểu và thực hiện, vì thời gian có hạn nên sẽ không tránh được những sai sót. Chúng em rất mong nhận được những ý kiến đóng góp của thầy để đề tài của chúng em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

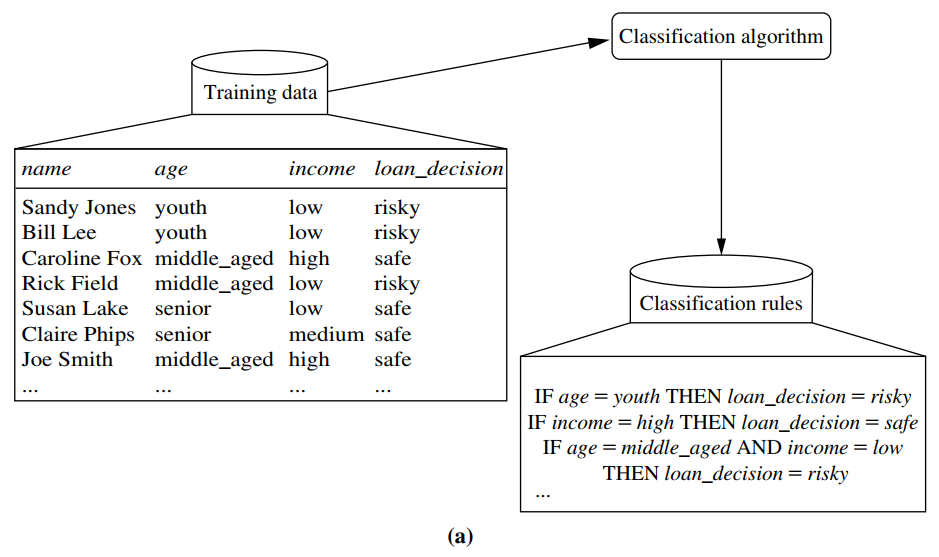
# **PHẦN 1: TỔNG QUAN VỀ PHÂN LỚP DỮ LIỆU.**

## **1.1 Phân lớp dữ liệu là gì?**

Phân lớp dữ liệu (classification) là một trong những hướng nghiên cứu chính của khai phá dữ liệu. Thực tế đặt ra nhu cầu là từ một cơ sở dữ liệu với nhiều thông tin ẩn con người có thể trích rút ra các quyết định nghiệp vụ thông minh.Công nghệ này đã, đang và sẽ có nhiều ứng dụng trong các lĩnh vực thương mại, ngân hàng, y tế,giáo dục…Trong các mô hình phân lớp đã được đề xuất, cây quyết định được  coi là công cụ mạnh, phổ biến và đặc biệt thích hợp với các ứng dụng khai phá dữ  liệu. Thuật toán phân lớp là nhân tố trung tâm trong một mô hình phân lớp

## **1.2 Cách tiếp cận chung để phân loại dữ liệu.**

Phân loại dữ liệu là một quá trình hai bước, bao gồm bước học tập (nơi mô hình phân loại được xây dựng) và bước phân loại (nơi mô hình được sử dụng để dự đoán nhãn lớp cho dữ liệu nhất định).



*H1.1(a) Bước học tập, giai đoạn huấn luyện (Learning).*

Hình H1.1(a): Dữ liệu đào tạo được phân tích bằng một thuât toán phân loại. Ở đây thuộc tính nhãn lớp là quyết định cho vay, và mô hình hay bộ phân loại đã học được biểu diễn dưới dạng các quy tắc phân loại.

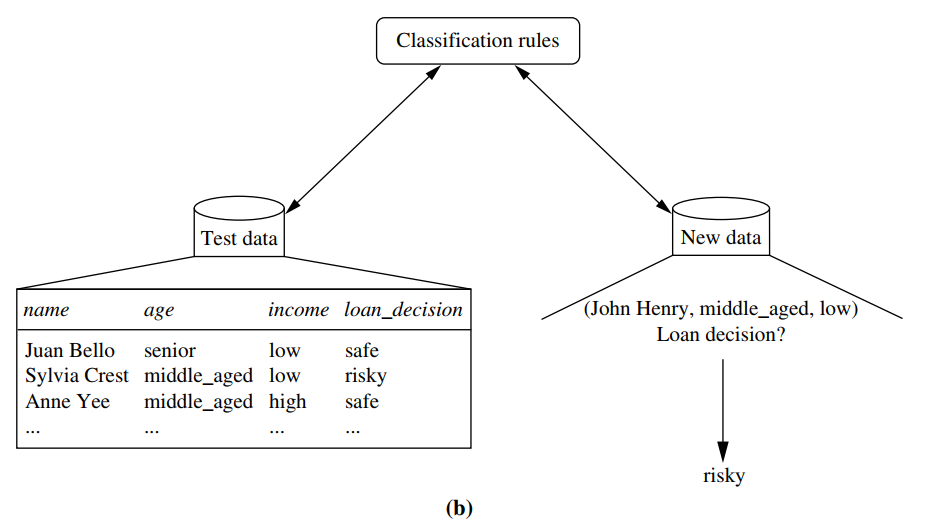
* **Bước học (Giai đoạn huấn luyện)**: Xây dựng bộ phân lớp (classifier) bằng việc phân tích, học tập huấn luyện.

Trong bước đầu tiên, bộ phân loại được xây dựng mô tả một tập hợp các lớp dữ liệu hoặc khái niệm được xác định trước, trong đó một thuật toán phân loại xây dựng bộ phân loại bằng cách phân tích hoặc “học từ” một tập huấn luyện được tạo thành từ các bộ dữ liệu cơ sở dữ liệu và các nhãn lớp được liên kết của chúng.

Một bộ dữ liệu được biểu diễn bằng một vector thuộc tính chiều, mô tả phép đo được thực hiện trên bộ thuộc tính cơ sở dữ liệu tương ứng . Mỗi bộ được giả định là thuộc về một lớp được xác định trước bỏi một thuộc tính cơ sở dữ liệu khác gọi là lớp thuộc tính nhãn . Thuộc tính nhãn lớp có giá trị rời rạc và không có thứ tự. Nó mang tính phân loại (hoặc danh nghĩa) trong đó mỗi giá trị cá nhân tạo thành bộ đào tạo được gọi là bộ đào tạo và được lấy mẫu ngẫu nhiên từ cơ sở dữ liệu có thể được gọi là mẫu.

Vì nhãn lớp của mỗi bộ đào tạo được cung cấp, bước này còn được gọi là học có giám sát (tức là quá trình học của bộ phân loại được “giám sát” trong đó nó được thông báo cho lớp nào mà mỗi bộ đào tạo thuộc về). Nó trái ngược với học không giám sát (hoặc phân cụm), trong đó nhãn lớp của mỗi bộ đào tạo không được biết và số lượng hoặc nhóm lớp sẽ học có thể không được biết trước. Ví dụ: nếu chúng tôi không có sẵn dữ liệu quyết định khoản vay cho tập huấn luyện, chúng tôi có thể sử dụng phân cụm để cố gắng xác định “các nhóm như các bộ giá trị”, có thể tương ứng với các nhóm rủi ro trong dữ liệu đơn đăng ký khoản vay.

Bước đầu tiên của quá trình phân loại này cũng có thể được xem như việc học một ánh xạ hoặc hàm, , có thể dự đoán nhãn lớp liên quan của một bộ dữ liệu nhất định . Theo quan điểm này, chúng tôi muốn học một ánh xạ hoặc hàm phân tách các lớp dữ liệu. Thông thường, ánh xạ này được biểu diễn dưới dạng các quy tắc phân loại, cây quyết định hoặc công thức toán học. Trong ví dụ dưới đây, việc lập bản đồ được trình bày dưới dạng các quy tắc phân loại để xác định các đơn xin vay là an toàn hay rủi ro (Hình H1.1(a)). Các quy tắc có thể được sử dụng để phân loại các bộ dữ liệu trong tương lai, cũng như cung cấp thông tin chi tiết sâu hơn về nội dung dữ liệu. Chúng cũng cung cấp một biểu diễn dữ liệu nén.



*H1.1(b) Bước phân lớp (Classification).*

Hình H1.1(b): Dữ liệu thử nghiệm được sử dụng để ước tính độ chính xác của quy tắc phân loại. Nếu độ chính xác được coi là chấp nhận được các quy tắc có thể được áp dụng để đưa vào phân loại các bộ dữ liệu mới.

* **Bước phân lớp (Classification)**: Phân lớp dữ liệu hay đối tượng nếu độ chính xác của bộ phân lớp được đánh giá là có thể chấp nhận được (acceptable).

Trong bước thứ hai (Hình H1.1(b)), mô hình được sử dụng để phân loại. Đầu tiên, độ chính xác dự đoán của trình phân loại được ước tính. Nếu chúng ta sử dụng tập huấn luyện để đo độ chính xác của trình phân loại, thì ước tính này có thể sẽ lạc quan, bởi vì trình phân loại có xu hướng trang bị quá nhiều dữ liệu (tức là trong quá trình học, nó có thể kết hợp một số điểm bất thường cụ thể của dữ liệu huấn luyện không có trong tập dữ liệu chung tổng thể). Do đó, một tập hợp thử nghiệm được sử dụng, bao gồm các bộ thử nghiệm và các nhãn lớp liên quan của chúng. Chúng độc lập với các bộ đào tạo, có nghĩa là chúng không được sử dụng để xây dựng bộ phân loại.

Độ chính xác của bộ phân loại trên một tập hợp thử nghiệm nhất định là phần trăm các bộ thử nghiệm được phân loại chính xác bởi bộ phân loại. Nhãn lớp được liên kết của mỗi bộ kiểm tra được so sánh với dự đoán lớp của bộ phân loại đã học cho bộ đó.

# **PHẦN 2: PHƯƠNG PHÁP PHÂN LỚP NAÏVE BAYES (BAYES CLASSIFICATION METHODS).**

## **2.1 Giới thiệu.**

Định lý Bayes được đặt theo tên của Thomas Bayes, một giáo sĩ người Anh không theo đạo, người đã sớm nghiên cứu về lý thuyết xác suất và quyết định trong thế kỷ 18. Khác với các phương pháp như cây quyết định hay là khai thác luật kết hợp (giải thuật FP-Growth, giải thuật Apriori) đều dựa trực tiếp vào dữ liệu để xây dựng mô hình phù hợp với dữ liệu đó, Naïve Bayesian Classification (NBC) là phương pháp phân lớp dữ liệu dựa trên mô hình xác suất, nhằm mục đích muốn xác định xem xác suất (hay dự đoán xác suất) mà một mẫu mới là thành viên thuộc lớp là bao nhiêu. Để dự đoán được xác suất đó cần dựa trên một tập dữ liệu có sẵn (hay tập dữ liệu huấn luyện) và dựa trên định lý Bayes.



Reverend Thomas Bayes (1702 - 1761)

## **2.2 Định lý Bayes.**

* Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên khi biết một sự kiện liên quan đã xảy ra. Xác suất này được kí hiệu là đại lượng này được gọi là xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của hoặc phụ thuộc vào giá trị đó.
* Định lý Bayes hữu ích ở chỗ nó cung cấp một cách tính toán xác suất hậu nghiệm:

Trong đó:

* X, H là hai biến cố
* P(H) là xác suất tiên nghiệm của biến cố H
* P(X) là xác suất tiên nghiệm của biến cố X
* P(H|X) là xác suất hậu nghiệm của H khi đã biết X
* P(X|H) là xác suất hậu nghiệm của X khi đã biết H
* Ví dụ: Giả sử tập dữ liệu khách hàng được mô tả bởi thuộc tính tuổi và thu nhập.
* Khách hàng 35 tuổi, thu nhập 40.000$.
* Giả thiết khách hàng mua máy.
* là xác suất khách hàng sẽ mua máy tính mà không cần quan tâm tới điều kiện tuổi và thu nhập.
* là xác suất khách hàng có tuổi là 35 và thu nhập là 40.000$.
* là xác suất người dùng có tuổi là 35 và thu nhập là 40.000$ mua máy tính.
* là xác suất người dùng mua máy tính có tuổi là 35 và thu nhập là 40.000$.
* => Trong ví dụ trên, theo định lý Bayes ta xác định được xác suất của người dùng có độ tuổi 35 với thu nhập 40.000$ mua máy tính là: tích của xác suất người dùng mua máy tính có tuổi là 35 và thu nhập là 40.000$ với xác suất khách hàng sẽ mua máy tính mà không cần quan tâm tới điều kiện tuổi và thu nhập chia cho xác suất khách hàng có tuổi là 35 và thu nhập là 40.000$.

## **2.3 Phân lớp Naïve Bayes.**

### **2.3.1 Cách hoạt động của phương pháp phân lớp Naïve Bayes.**

Bộ phân lớp **Naïve Bayes**, hoặc bộ phân lớp **Simple Bayesian** hoạt động như sau:

1. Gọi là tập hợp các bộ giá trị huấn luyện và nhãn lớp tương ứng. Mỗi bộ được biểu diễn bằng một vector thuộc tính chiều, , mô tả phép đo thực hiện trên bộ từ thuộc tính tương ứng là
2. Giả sử có lớp . Cho một bộ giá trị , bộ phân lớp sẽ dự đoán rằng thuộc lớp có xác suất hậu nghiệm cao nhất. Cụ thể bộ phân lớp Bayes sẽ dự đoán thuộc vào lớp khi và chỉ khi:

Do đó, chúng ta lấy giá trị lớn nhất của . Lớp mà là lớn nhất gọi là giả thuyết posteriori cực đại. Theo định lý Bayes:

1. Để tìm xác suất lớn nhất, ta nhận thấy các giá trị là không đổi với tất cả các lớp, ta chỉ cần tìm giá trị lớp nhất của . Nếu xác suất tiên nghiệm của lớp không được biết (hay không được biết) thì người ta thường giả định rằng các lớp có khả năng như nhau, nghĩa là , do đó ta cần tính giá trị lớn nhất của . Nếu không chúng ta tính giá trị lớn nhất của . Lưu ý rằng các xác suất tiên nghiệm của lớp có thể được ước tính bởi trong đó là số lượng bộ huấn luyện của lớp trong .
2. Với bộ dữ liệu có nhiều thuộc tính, việc tính toán sẽ rất tốn kém. Để giảm bớt tính toán trong việc đánh giá , giả định duy nhất về tính độc lập có điều kiện của lớp được đưa ra. Điều này giả định rằng các giá trị của thuộc tính là độc lập có điều kiện với nhau, với nhãn lớp của bộ (nghĩa là không có mối quan hệ phụ thuộc nào giữa các thuộc tính). Ta có:

Chúng ta có thể dễ dàng ước tính các xác suất từ các bộ giá trị đào tạo. ở đây là giá trị của thuộc tính tương ứng cho mỗi bộ dữ liệu . Đối với mỗi thuộc tính , chúng ta sẽ xem xét liệu thuộc tính đó là thuộc tính phân loại hay thuộc tính liên tục.

* là thuộc tính phân loại, thì là số bộ giá trị của lớp trong mà có giá trị đối với thuộc tính .
* là thuộc tính liên tục, một thuộc tính có giá trị liên tục thường được giả định có phân phối chuẩn Gauss với giá trị trung bình và độ lệch chuẩn , được định nghĩa theo hàm mật độ xác suất như sau:

vậy nên:

5. Để dự đoán nhãn lớp của , được đánh giá cho mỗi lớp . Bộ phân loại dự đoán rằng nhãn lớp của bộ dữ liệu là lớp khi và chỉ khi:

for

Nói cách khác, nhãn lớp được dự đoán là lớp mà là lớn nhất.

### **2.3.2 Ưu điểm và nhược điểm của phương pháp phân lớp Naïve Bayes.**

**Ưu điểm:**

1. Dự đoán một lớp tập dữ liệu thử nghiệm dể dàng và nhanh chóng.
2. Bộ phân loại Naïve Bayes hoạt động tốt hơn so với các mô hình giả định độc lập.
3. Giả định độc lập: Hoạt động tốt cho nhiều bài toán / miền dữ liệu và ứng dụng. Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam, …
4. Yêu cầu ít dữ liệu đào tạo.

**Nhược điểm:**

1. Giả định độc lập (vừa là ưu điểm vừa là nhược điểm): Hầu hết các trường hợp thực tế trong đó có các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau.
2. Nếu một thử nghiệm trong tập dữ liệu thử nghiệm có một danh mục không xuất hiện trong quá trình huấn luyện thì sẽ gán cho nó xác suất “không” và sẽ không thể đưa ra dự đoán. Đây là vấn đề được gọi là tần số Zero.
3. Mô hình không được huấn luyện bằng phương pháp tối ưu mạnh và chặt chẽ. Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ, không tính đến sự tương tác giữa các ước lượng này.

## **2.4 Một vài ví dụ.**

**Ví dụ 1.**

Phân các bênh nhân thành 2 lớp ung thư hoặc không ưng thư. Giả sử xác suất để một người bị ung thư là 0.008 tức là P (ung thư) = 0.008; và P (không ung thư) = 0.992. Xác suất để bệnh nhân ung thư có kết quả xét nghiệm dương dương tính là 0.98 và xác suất để bệnh nhân không ung thư có kết quả dương tính là 0.03 tức là P(+/ungthu) = 0.98, P(+/khongungthu) = 0.03. Bây giờ giả sử một bệnh nhân có kết quả xét nghiệm dương tính. Ta có:

P (+/ungthu) P(ungthu) = 0.98 \* 0.008 = 0.0078

P (+/khongungthu) P(khongungthu) = 0.03 \* 0.992 = 0.0298

* P (+/khongungthu) P(khongungthu) > P (+/ungthu) P(ungthu)

**Như vậy, ta xét đoán rằng: bệnh nhân là không ung thư**

**Ví dụ 2:**

Đề bài: Hãy dự đoán nhãn lớp của bộ dữ liệu dưới đây bằng cách sử dụng phân loại Bayes. Dữ liệu huấn luyện được cung cấp như bảng phía dưới. Các bộ dữ liệu được mô tả bằng các thuộc tính age, income, student, credit\_rating, thuộc tính nhãn lớp buys\_computer có hai giá trị riêng biệt (cụ thể là yes hoặc no). Bộ dữ liệu cần phân loại là

AllElectronics Customer Database:

Table

Description automatically generated

Ta có:

* Có 2 lớp ,
* Ta cần tìm:

hay

* Tính :
* Tính :

Vì đều là các thuộc tính phân loại nên ta có:

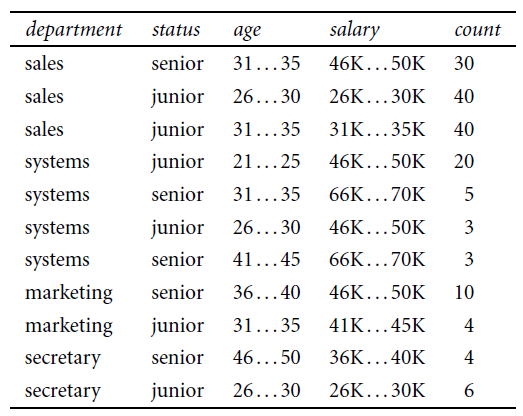
* Tính :

Từ kết quả tính toán có giá trị lớn nhất, do đó thuật toán Bayes dự đoán người khách hàng này sẽ mua máy tính.

**Ví dụ 3.**

Đề bài: Cho một bộ dữ liệu có các giá trị “system”, “26…30”, và “46K-50K” tương ứng với các thuộc tính “department”, “age”, “salary”. Dùng phương pháp Bayesian phân lớp cho bộ dữ liệu theo thuộc tính “status”?

Cho cơ sở dữ liệu:



Ta có:

* Có 2 lớp ,
* Tính :
* Tính :

Vì {department, age, salary} đều là các thuộc tính phân loại nên ta có:

* Tính :

Ta thấy kết quả tính toán có giá trị lớn nhất. Do đó dự đoán bộ dữ liệu gồm các giá trị {department=”system”, age=”26…30”, salary=”46K…50K”} thuộc lớp status = “Junior”.

**PHẦN 3: TỔNG KẾT.**

**3.1 Kết luận.**

Mô hình Naive Bayes là mô hình phân lớp đơn giản dễ cài đặt, có tốc độ xử lý nhanh. Tuy nhiên có nhược điểm lớn là yêu cầu các đặc trưng đầu vào phải độc lập, mà điều này khó xảy ra trong thực tế làm giảm chất lượng của mô hình. Thuật toán này thường được sử dụng trong phân lớp văn bản, lọc spam(spam filtering) , recommendation systems, ...

**3.2 Tài liệu tham khảo.**

1. Data Mining: [1] 2011 Data Mining - Concepts and Techniques (3rd Ed).pdf
2. Thuật toán phân lớp Naïve Bayes: <https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM>