**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: KHAI PHÁ DỮ LIỆU SỬ DỤNG NAIVE BAYES**

**ĐỂ DỰ ĐOÁN KHẢ NĂNG SỐNG SÓT CỦA MỘT NGƯỜI**

**TRÊN TÀU TITANIC**

**Giảng viên hướng dẫn: VŨ VĂN ĐỊNH**

**Sinh viên thực hiện: NGÔ HỮU HƯỞNG**

**PHẠM HỒNG THÁI**

**Lớp: D14CNPM1**

**Chuyên Ngành: CÔNG NGHỆ PHẦN MỀM**

**Hà Nội, ngày 16 tháng 6 năm 2022**

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện

|  |  |  |
| --- | --- | --- |
| **Họ và Tên** | **Chữ ký** | **Ghi chú** |
| NGÔ HỮU HƯỞNG |  |  |
| PHẠM HỒNG THÁI |  |  |

Giảng viên chấm

|  |  |  |
| --- | --- | --- |
| **Họ và Tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1 |  |  |
| Giảng viên chấm 2 |  |  |

**MỤC LỤC**

[**LỜI MỞ ĐẦU** 6](#_Toc72785149)

[**CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU** 7](#_Toc72785150)

[1.1. Khái niệm về khai phá dữ liệu 7](#_Toc72785151)

[1.1.1 Khai phá dữ liệu là gì? 7](#_Toc72785152)

[1.1.2 Các bước trong khai phá dữ liệu 7](#_Toc72785153)

[1.1.3 Ứng dụng 8](#_Toc72785155)

[1.2. Các phương pháp khai phá dữ liệu 9](#_Toc72785156)

[1.2.1 Phân lớp dữ liệu 9](#_Toc72785157)

[1.2.3 Phân cụm 9](#_Toc72785158)

[1.2.3 Luật kết hợp 10](#_Toc72785159)

[1.2.4 Hồi quy tuyến tính 10](#_Toc72785160)

[1.3. Bài toán phân lớp 10](#_Toc72785161)

[1.3.1 Phân lớp nhị phân 11](#_Toc72785162)

[1.3.2 Phân lớp nhiều lớp 11](#_Toc72785163)

[1.3.3 Quá trình phân lớp dữ liệu 11](#_Toc72785164)

[**CHƯƠNG 2 THUẬT TOÁN NAIVE BAYES TRONG BÀI TOÁN PHÂN LỚP** 13](#_Toc72785165)

[2.1 Tổng quan về thuật toán Naive Bayes 13](#_Toc72785166)

[2.1.1 Giới thiệu thuật toán 13](#_Toc72785167)

[2.1.2 Định luật Bayes 13](#_Toc72785168)

[2.2. Ví dụ thực tế 15](#_Toc72785170)

[2.3 Ứng dụng của thuật toán 17](#_Toc72785172)

[2.4 Ưu nhược điểm của thuật toán 18](#_Toc72785173)

[2.4.1 Ưu điểm 18](#_Toc72785174)

[2.4.2 Nhược điểm 18](#_Toc72785175)

[**CHƯƠNG 3 ỨNG DỤNG THUẬT TOÁN NAIVE BAYES CHO ĐỀ TÀI** 19](#_Toc72785176)

[3.1 Bài toán dự đoán khả năng sống sót của một người trên tàu titanic bằng thuật toán Naive Bayes 19](#_Toc72785177)

[3.1.1 Phát biểu bài toán 19](#_Toc72785178)

[3.1.2 Yêu cầu bài toán 19](#_Toc72785179)

[3.1.3 Xây dựng bộ dữ liệu 19](#_Toc72785180)

[3.1.4 Áp dụng thuật toán vào bài toán 20](#_Toc72785181)

[3.1.5 Cài đặt với thư viện sklearn và python 20](#_Toc72785182)

[3.1.6 Kết quả 23](#_Toc72785185)

[**KẾT LUẬN** 24](#_Toc72785187)

[**TÀI LIỆU THAM KHẢO** 25](#_Toc72785188)

**DANH MỤC HÌNH ẢNH**

[Hình 1.1 Các bước của Khai phá dữ liệu 11](#_Toc104881569)

[Hình 2.2 Ví dụ thực tế 18](#_Toc104881586)

[Hình 3.1 Dữ liệu về người trên tàu 23](#_Toc104881596)

[Hình 3.2 Mã nguồn 1 24](#_Toc104881599)

[Hình 3.3 Mã nguồn 2 25](#_Toc104881600)

[Hình 3.4 Kết quả chạy chương trình 26](#_Toc104881602)

# **LỜI MỞ ĐẦU**

Trong thời đại ngày nay, yếu tố quyết định thành công trong mọi lĩnh vực luôn gắn liền với việc nắm bắt, thống kê và khai thác thông tin hiệu quả. Dữ liệu ngày càng lớn nên việc tìm ra những thông tin tiềm ẩn trong chúng càng khó khăn hơn. Khai phá tri thức là một lĩnh vực nghiên cứu mới, mở ra một thời kỳ trong việc tìm ra thông tin hữu ích. Nhiệm vụ cơ bản của lĩnh vực này là khai phá tri thức trong cơ sở dữ liệu, khai phá dữ liệu trong cơ sở dữ liệu không phải là một hệ thống phân tích tự động mà là một quá trình tương tác thường xuyên giữa con người với cơ sở dữ liệu được sự trợ giúp của nhiều phương pháp và công cụ tin học.

Ứng dụng phân cụm dữ liệu trong khai phá dữ liệu nhằm xây dựng hệ thống chuẩn đoán là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữ liệu thu nhập được về số lượng của sản phẩm và phân loại hình, đề tài đề xuất ứng dụng mô hình phân lớp dữ liệu bằng thuật toán Naive Bayes để tìm ra quy luật tiềm ẩn trong dữ liệu.

Nhóm chúng em xin được gửi lời cảm ơn của mình tới Thầy Vũ Văn Định, người đã trực tiếp hướng dẫn, chỉ bảo tận tình, cung cấp tài liệu để nhóm chúng em hoàn thành bản báo cáo này.

Trong suốt quá trình nghiên cứu, mặc dù đã hết sức cố gắng nhưng chắc chắn bài báo cáo không tránh khỏi những thiếu sót, chúng em rất mong thầy cô góp ý để báo cáo được hoàn chỉnh hơn. Chúng em xin chân thành cảm ơn.

# **CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

## 1.1. Khái niệm về khai phá dữ liệu

### 1.1.1 Khai phá dữ liệu là gì?

Data mining – khai phá dữ liệu là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Các MCU khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai.

Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

* Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
* Tính toán dự đoán kết quả
* Tạo thông tin phản hồi để phân tích
* Tập trung vào cơ sở dữ liệu lớn hơn.
* Phân cụm dữ liệu trực quan

### 1.1.2 Các bước trong khai phá dữ liệu

* **Bước 1**: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.
* **Bước 2**: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.
* **Bước 3**: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.
* **Bước 4**: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.
* **Bước 5**: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.
* **Bước 6**: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.
* **Bước 7**: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.

### *Hình 1.1 Các bước của Khai phá dữ liệu*

### 1.1.3 Ứng dụng

* Phân tích thị trường và chứng khoán
* Phát hiện gian lận
* Quản lý rủi ro và phân tích doanh nghiệp
* Phân tích giá trị trọn đời của khách hàng
* Khám phá thêm 10 ứng dụng khai phá dữ liệu

## 1.2. Các phương pháp khai phá dữ liệu

Khai phá dữ liệu là một chuyên ngành rất rộng và có rất nhiều hướng nghiên cứu bài toán khác nhau. Tuy nhiên, chúng được tiếp cận theo các hướng chính sau:

### 1.2.1 Phân lớp dữ liệu

Phân lớp dữ liệu là việc xây dựng một mô hình mà có thể phân các đối tượng thành những lớp để dự đoán giá trị bị mất tại một số thuộc tính của dữ liệu hay tiêu đoán giá trị của dữ liệu sẽ xuất hiện trong tương lai.

Quá trình gồm 2 bước:

* Bước học (giai đoạn huấn luyện): xây dựng bộ phân loại (classifier) bằng việc phân tích/ học tập huấn luyện.
* Bước phân loại (classification): phân loại dữ liệu/đối tượng mới nếu mức độ chính xác của bộ phận phân loại được đánh giá là có thể chấp nhận được.

Các giải thuật phân loại dữ liệu:

* Phân loại dữ liệu với cây quyết định (decision tree)
* Phân loại dữ liệu với mạng Bayesian
* Phân loại dữ liệu với mạng neural
* Phân loại dữ liệu với k phần tử gần nhất (k-nearest neighbor)
* Phân loại dữ liệu với SVM…

### 1.2.3 Phân cụm

Phân cụm là việc nhóm một tập các đối tượng có cùng đặc điểm giống nhau hay gần giống nhau vào cùng một nhóm.

Các đối tượng trong cùng một nhóm tương tự với nhau hơn so với đối tượng ở cụm khác. Phương pháp phân cụm hỗ trợ giai đoạn tiền xử lý dữ liệu, mô tả sự phân bố dữ liệu/đối tượng…

Các phương pháp phân cụm phổ biến:

* Phân hoạch(partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
* Phân cấp(hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
* Dựa trên cấp độ (density-based): dựa trên connectivity and density functions
* Dựa trên lưới (grid-based): dựa trên multiple-level granularity structure
* Dựa trên mô hình (model-based): một mô hình giả thuyết được tạo ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất……

### 1.2.3 Luật kết hợp

Là quá trình khám phá các tập giá trị thuộc tính xuất hiện phổ biến trong các đối tượng dữ liệu. Từ tập phổ biến có thể tạo ra các luật kết hợp giữa các giá trị thuộc tính trong tập các đối tượng.

### 1.2.4 Hồi quy tuyến tính

Phương pháp hồi quy được sử dụng để đưa ra các dự báo dựa trên các dữ liệu đang tồn tại bằng các áp dụng các công thức. Một hàm sẽ được học ra từ bộ dữ liệu hiện có bằng cách sử dụng các kỹ thuật hồi quy và tuyến tính từ việc thống kê. Sau đó, dữ liệu mới sẽ căn cứ vào hàm này để đưa ra những dự đoán.

Trong bài báo cáo này bọn em chọn phương pháp phân lớp bởi vì nó là một thuật toán tuy đơn giản nhưng lại khá hiệu quả và được sử dụng rộng rãi.

## 1.3. Bài toán phân lớp

Bài toán phân lớp (classification) và bài toán phân cụm (cluster) là hai bài toán lớn trong lĩnh vực Machine Learnig (ML). Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.

Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phần lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào.

Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.

### 1.3.1 Phân lớp nhị phân

Bài toán phân lớp nhị phân là bài toán gắn nhãn dữ liệu cho đối tượng vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (feature) của bộ phân lớp

### 1.3.2 Phân lớp nhiều lớp

Bài toán phân lớp đa lớp là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn hai. Như vậy với từng dữ liệu chúng ta phải xem xét và phân lớp chúng vào những lớp khác nhau chứ không phải là hai lớp như bài toán phân lớp nhị phân. Và thực chất bài toán phân lớp nhị phân là một bài toán đặt biệt của phân lớp đa lớp.

### 1.3.3 Quá trình phân lớp dữ liệu

**Bước 1: Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)**

Công đoạn này được xem là công đoạn quan trọng trong các bài toán về Machine Learning. Vì đây là input cho việc học dể tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặt trưng tốt (good feature) của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu (noise). Uớc lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số chiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu (reduce demension).

Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập datasets thành hai phần, một phần phục vụ cho training (training datasets) và phần còn lại phục vụ cho mục đích testing trên mô hình (testing dataset). Có hai cách thường sử dụng trong cross-validation là splitting và k-fold.

**Bước 2: Xây dựng mô hình phân lớp (classifier model)**

Mục đích của mô hình huấn luyện là tìm ra hàm f(x) và thông qua hàm f tìm được để chúng ta gán nhãn cho dữ liệu. Bước này thường được gọi là học hay training.

*f(x) = y*

Trong đó:

* x là các feature hay input đầu vào của dữ liệu
* y là nhãn lớp hay output đầu ra.

Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát (supervised learning) như KNN, Neural Network, SVM, Decision Tree, Navie Bayers

**Bước 3: Kiểm tra dữ liệu với mô hình (make prediction)**

Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này chúng ta sẽ đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

**Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất**

Bước cuối cùng chúng ta sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu traning thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn của chúng ta thì phải thay đổi các tham số (turning parameter) của các thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta.

# **CHƯƠNG 2 THUẬT TOÁN NAIVE BAYES TRONG BÀI TOÁN PHÂN LỚP**

## 2.1 Tổng quan về thuật toán Naive Bayes

### 2.1.1 Giới thiệu thuật toán

Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê. Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.

Ví dụ như ta có thể ứng dụng vào việc thiết kế một ứng dụng nghe nhạc có thể phán đoán được sở thích của nghe nhạc của người dùng dựa trên các hành vi như nhấn nút “thích” bài hát, “nghe đi nghe” lại nhiều lần các bài hát, “bỏ qua” các bài hát không thích …. Dựa trên tập dữ liệu đó ta có thể áp dụng NBC để tính toán ra các phong cách nhạc mà người dùng thích nhất, từ đó chúng ta có thể đưa ra các “gợi ý” nghe nhạc gần đúng nhất cho người dùng từ việc học hỏi từ những thói quen đó.

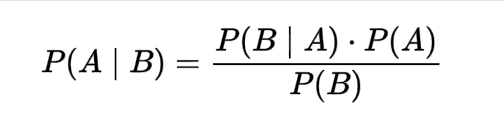
### 2.1.2 Định luật Bayes

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó. Theo định lí Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

* Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là P(A) và đọc là xác suất của A. Đây được gọi là xác suất biên duyên hay xác suất tiên nghiệm, nó là “tiên nghiệm” theo nghĩa rằng nó không quan tâm đến bất kỳ thông tin nào về B.
* Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là P(B) và đọc là “xác suất của B”. Đại lượng này còn gọi là hằng số chuẩn hóa (normalising constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.
* Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là P(B|A) và đọc là “xác suất của B nếu có A”. Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra. Chú ý không nhầm lẫn giữa khả năng xảy ra B khi biết A và xác suất xảy ra A khi biết B.

Tóm lại định lý Bayes sẽ giúp ta tính ra xác suất xảy ra của một giả thuyết bằng cách thu thập các bằng chứng nhất quán hoặc không nhất quán với một giả thuyết nào đó. Khi các bằng chứng tích lũy, mức độ tin tưởng vào một giả thuyết thay đổi. Khi có đủ bằng chứng, mức độ tin tưởng này thường trở nên rất cao hoặc rất thấp, tức là xác xuất sảy ra giả thuyết sẽ thay đổi thì các bằng chứng liên quan đến nó thay đổi.

Công thức của định luật Bayes được phát biểu như sau:

*Hình 2.1 Công thức của định luật*

### *Trong đó:*

– P(A|B) là xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.

– P(B|A) là xác suất xảy ra B khi biết A xảy ra

– P(A) là xác suất sảy ra của riêng A mà không quan tâm đến B.

– P(B) là xác suất xảy ra của riêng B mà không quan tâm đến A.

Ở trên ta có thể thấy xác suất sảy ra của giả thuyết A phụ thuộc và xác suất của giả thuyết B, nhưng trong thực tế xác suất A có thể phụ thuộc vào xác suất của nhiều các giác thuyết khác có thể là B1, B2, B3 … Bn.

## 2.2. Ví dụ thực tế

Trong một vụ thu hoạch ở một đồn điền trang trại các người làm đã thu hoạch được hơn 1000 trái cây các loại được phân loại thành 3 nhóm trái cây chính là “Chuối (banana)”, “Cam (orange)” và “các loại trái cây khác (other fruit)” và được phân l oại thành các kiểu như loại trái cây “dài (long), “không dài (not long), “ngọt (sweet)”, “không ngọt (not sweet)”, “màu vàng (yellow)”, “không phải màu vàng (not yellow)”:

### *Hình 2.2 Ví dụ thực tế*

Bây giờ bài toàn đặt ra là tính ra tỷ lệ một quả chuối có thuộc tính là “màu vàng, dài, và ngọt” với tỷ lệ quả cảm và các loại hoa quả khác có cũng có thuộc tính là “màu vàng, dài, và ngọt”.

Áp dụng định lý Bayes ta sẽ có 3 công thức tính cho 3 loại trái cây như sau:

*Tỷ lệ quả chuối với thuộc tính “vàng, dài và ngọt”:*

P(Long|Banana) = 400/500 = 0.8

P(Sweet|Banana) = 350/500= 0.7

P(Yellow|Banana) =450/500= 0.9

P(Banana) = 500/1000 = 0.5

P(Long) = 500/1000 = 0.5

P(Sweet) = 650/1000 = 0.65

P(Yellow) = 800/1000= 0.8

P(Banana|Long,Sweet,Yellow) = (0.8 \* 0.7 \* 0.9 \* 0.5) / (0.5 \* 0.65 \* 0.8) = 0.97

Tức là tỷ lệ chuối với thuộc tính “vàng, dài và ngọt” là 97%

*Tương tự ta cũng có thể tính ra tỷ lệ quả cam với thuộc tính “vàng dài và ngọt” với công thức sau:*

Do tỷ lệ P(Long|Orange) = 0/500 = 0 cho nên P(Orange|Long, Sweet, Yellow) = 0 tức là tỷ lệ quả cam với thuộc tính “vàng dài và ngọt” là 0%.

*Cũng thế ta ốp công thức Bayes để tính các trái cây còn lại với thuộc tính “vàng dài và ngọt” với công thức sau:*

P(Long|Other Fruit) = 100/200= 0.5

P(Sweet|Other Fruit) = 150/200 = 0.75

P(Yellow|Other Fruit) =50/200= 0.25

P(Other Fruit) = 200/1000 = 0.2

P(Banana|Long,Sweet,Yellow) = 0.5 \* 0.75 \* 0.25 \* 0.2 / (0.5 \* 0.65 \* 0.8) = 0.072

Tức là tỷ lệ các trái cây khác có thuộc tính “vàng dài và ngọt” chỉ là khoảng 7,2%

Vậy suy ra với trái cây với ba thuộc tính là “Vàng, dài và ngọt” thì có khả năng cao nhất đó là quả chuối.

Chúng ta có thể ứng dụng Naive Bayes Classification để tính tỷ lệ xác suất với rất nhiều các dạng bài toán khác nhau, với dữ liệu càng nhiều thì độ chính xác của thuật toán sẽ càng cao, và khi dữ liệu thay đổi thì kết quả cũng thay đổi theo.

## 2.3 Ứng dụng của thuật toán

* *Real time Prediction:* NBC chạy khá nhanh nên nó thích hợp áp dụng ứng dụng nhiều vào các ứng dụng chạy thời gian thực, như hệ thống cảnh báo, các hệ thống trading …
* *Multi class Prediction:* Nhờ vào định lý Bayes mở rộng ta có thể ứng dụng vào các loại ứng dụng đa dự đoán, tức là ứng dụng có thể dự đoán nhiều giả thuyết mục tiêu.
* *Text classification/ Spam Filtering/ Sentiment Analysis:* NBC cũng rất thích hợp cho các hệ thống phân loại văn bản hay ngôn ngữ tự nhiên vì tính chính xác của nó lớn hơn các thuật toán khác. Ngoài ra các hệ thống chống thư rác cũng rất ưu chuộng thuật toán này. Và các hệ thống phân tích tâm lý thị trường cũng áp dụng NBC để tiến hành phân tích tâm lý người dùng ưu chuộng hay không ưu chuộng các loại sản phẩm nào từ việc phân tích các thói quen và hành động của khách hàng.
* *Recommendation System:* Naive Bayes Classifier và Collaborative Filtering được sử dụng rất nhiều để xây dựng cả hệ thống gợi ý, ví dụ như xuất hiện các quảng cáo mà người dùng đang quan tâm nhiều nhất từ việc học hỏi thói quen sử dụng internet của người dùng, hoặc như ví dụ đầu bài viết đưa ra gợi ý các bài hát tiếp theo mà có vẻ người dùng sẽ thích trong một ứng dụng nghe nhạc…

## 2.4 Ưu nhược điểm của thuật toán

### 2.4.1 Ưu điểm

* Giả định độc lập: hoạt động tốt cho nhiều bài toán/miền sữ liệu và ứng dụng.
* Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán như phân lớp văn bản, lọc spam,..
* Cho phép kết hợp tri thức tiền nghiệm (prior knowledge) và dữ liệu quan sát được (obserwed data).
* Tốt khi có sự chệnh lệch số lượng giữa các lớp phân loại.
* Huấn luyện mô hình (ước lượng tham số) dễ và nhanh.

### 2.4.2 Nhược điểm

* Giả định độc lập (ưu điểm cũng chính 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.
* Vấn đề zero (đã nêu cách giải quyết ở phía trên)
* Mô hình không được huẩn luyện bằng phượng 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.

# **CHƯƠNG 3. ỨNG DỤNG THUẬT TOÁN NAIVE BAYES CHO ĐỀ TÀI**

## 3.1 Bài toán dự đoán khả năng sống sót của một người trên tàu titanic bằng thuật toán Naive Bayes

### 3.1.1 Phát biểu bài toán

Bài toán dự đoán khả năng sống sót của một người trên tàu titanic đưa ra tất cả các thông tin của người trên tàu titanic để dự đoán

- Giá trị input: Thông tin của người trên tàu

- Giá trị output: Kết quả người đó sống hay chết

### 3.1.2 Yêu cầu bài toán

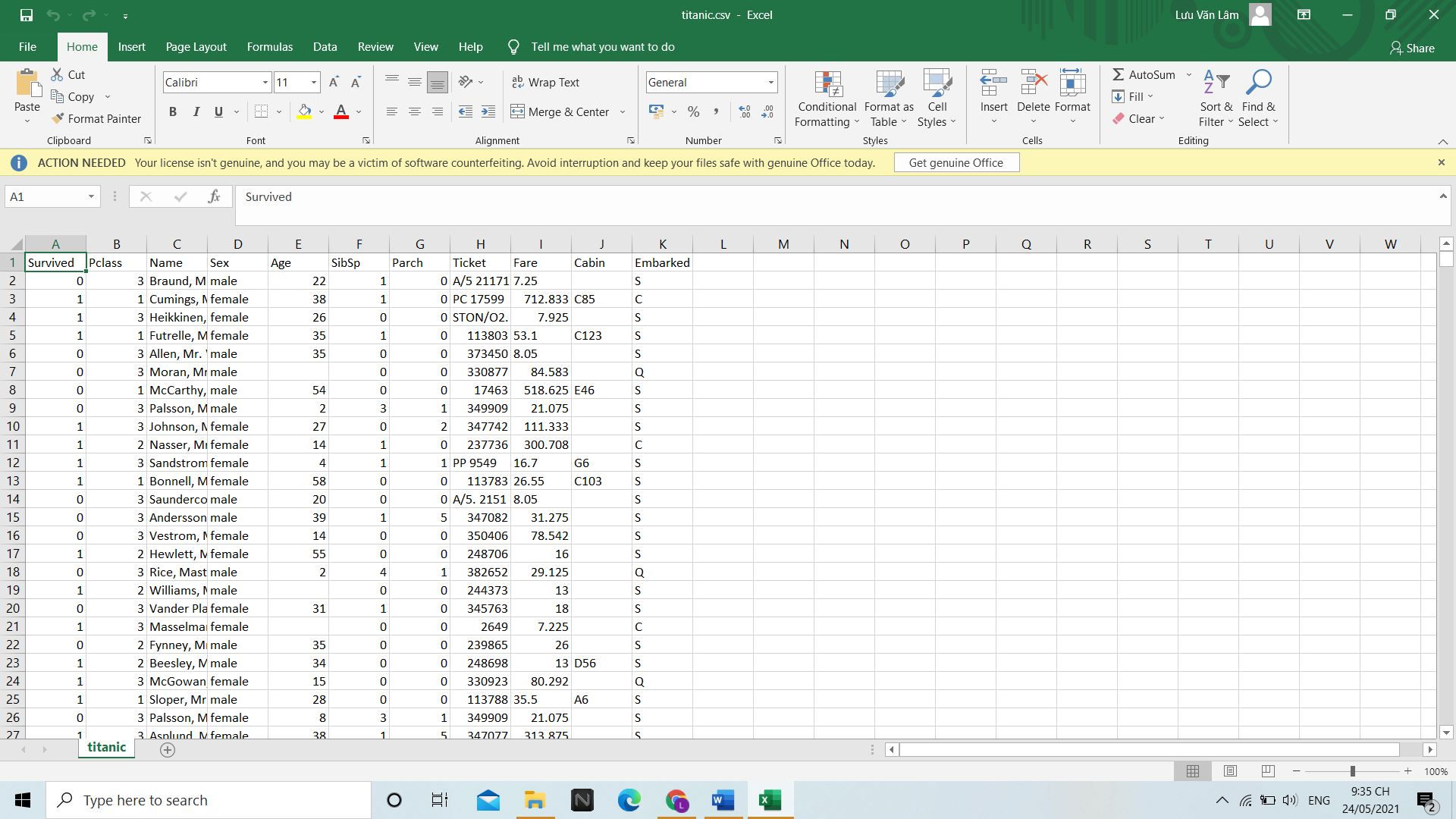
- Lấy dữ liệu về người trên tàu

- Trích chọn đặc trưng từ tập dữ liệu lấy được

- Tiến hành huấn luyện tập dữ liệu

- Dự đoán khả năng sống sót

### 3.1.3 Xây dựng bộ dữ liệu



### *Hình 3.1: Dữ liệu về người trên tàu*

### 3.1.4 Áp dụng thuật toán vào bài toán

Sử dụng thuật toán Naive Bayes để dự đoán khả năng sống sót

Dữ liệu chuẩn đoán

* Sex
* Pclass
* Age
* Fare

=>> Khả năng sống sót? 0 =>chết , 1 =>sống.

Dùng thuật toán Naive Bayes để để dự đoán khả năng sống sót của một người trên tàu.

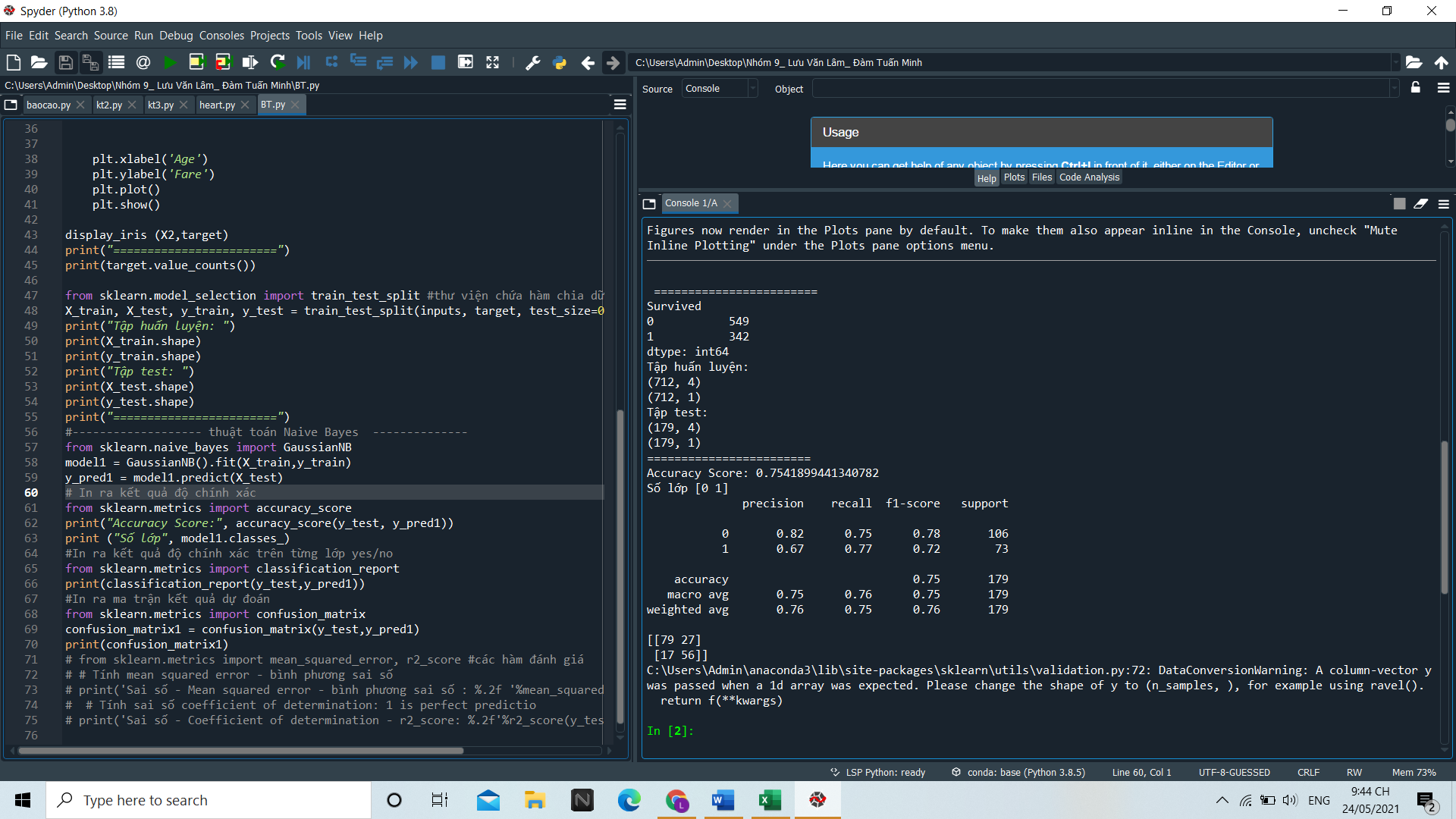
### 3.1.5 Cài đặt với thư viện sklearn và python

### *Hình 3.2 Mã nguồn* *1*

### *Hình 3.3 Mã nguồn 2*

### 

### 3.1.6 Kết quả



### Hình 3.4 Kết quả chạy chương trình

# **KẾT LUẬN**

Như vậy, đã minh họa cho các bạn thấy cách thức thực hiện một thuật toán Naive Bayes đơn giản nhất ứng dụng trong thực tế để dự đoán khả năng sống sót của một người trên tàu titanic.Tuy trong thực tế, chúng ta phải xét rất nhiều biến khác nhau tác động lên kết quả, và phân tích một khối lượng lớn dữ liệu, sử dụng nhiều phương pháp khác nhau để có được kết quả chính xác nhưng Naive Bayes vẫn được coi là một trong những cách tiếp cận hiệu quả.

**Những kết quả đạt được:**

* Sự hiểu biết về thuật toán Naive Bayes cơ bản tương đối tốt
* Hiểu biết thêm về ngôn ngữ lập trình Python
* Tìm hiểu được kiến thức về khai phá dữ liệu và biết được ứng dụng của khai phá dữ liệu với học máy

**Hạn chế:**

* Mô hình học vẫn còn chưa hiệu quả, tỉ lệ đúng vẫn còn ở mức trung bình
* Kết quả đưa ra có thể có sự thay đổi
* Vì là code thuần nên code sẽ kém tối ưu, kém hiệu quả

**TÀI LIỆU THAM KHẢO**

[1] https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM

[[2]](https://gist.github.com/curran/a08a1080b88344b0c8a7) <https://scikit-learn.org/>

[3] <https://www.kaggle.com/>

[4] <https://machinelearningcoban.com/2017/01/08/knn/>

[5] Bài giảng Khai Phá dữ liệu – Thầy Vũ Văn Định

[6] https://gialuan1991.wordpress.com/2016/04/15/2-naive-bayes/

[7] https://vi.wikipedia.org/wiki/Khai\_ph%C3%A1\_d%E1%BB%AF\_li%E1%BB%87u#:~:text=Khai%20ph%C3%A1%20d%E1%BB%AF%20li%E1%BB%87u%20(data,c%E1%BB%A7a%20khoa%20h%E1%BB%8Dc%20m%C3%A1y%20t%C3%ADnh

[8] http://hoctructuyen123.net/tong-quan-ve-thuat-toan-phan-lop-naive-bayes-classification