**Ảnh có chứa quảng trường

Mô tả được tạo tự động­­ĐẠI HỌC QUỐC GIA TP HCM**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙞🕮🙝



**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

Môn học: HỆ HỖ TRỢ QUYẾT ĐỊNH

ĐỀ TÀI : DỰ ĐOÁN KHẢ NĂNG CHẤM DỨT HỢP ĐỒNG CỦA NHÂN VIÊN

Giáo viên hướng dẫn: **TS. Cao Thị Nhạn**

**ThS. Vũ Minh Sang**

Nhóm thực hiện: **Nhóm 7**

Thành Viên:

Hồ Trọng Khang 19521661

Nguyễn Cao Khoa 19521694

**TP Hồ Chí Minh tháng 5 năm 2022**

**ĐẠI HỌC QUỐC GIA TP HCM**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙞🕮🙝



**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

Môn học: HỆ HỖ TRỢ QUYẾT ĐỊNH

ĐỀ TÀI : DỰ ĐOÁN KHẢ NĂNG CHẤM DỨT HỢP ĐỒNG CỦA NHÂN VIÊN

Giáo viên hướng dẫn: **TS. Cao Thị Nhạn**

**ThS. Vũ Minh Sang**

Nhóm thực hiện: **Nhóm 7**

Thành Viên:

Hồ Trọng Khang 19521661

Nguyễn Cao Khoa 19521694

**TP Hồ Chí Minh tháng 5 năm 2022**

**MỤC LỤC**

[**LỜI CẢM ƠN** 5](#_Toc102671615)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 6](#_Toc102671616)

[**CHƯƠNG I : TỔNG QUAN ĐỀ TÀI** 7](#_Toc102671617)

[**1.** **Đặt vấn đề** 7](#_Toc102671618)

[**2.** **Mục tiêu** 8](#_Toc102671619)

[**3.** **Công cụ sử dụng** 8](#_Toc102671620)

[**CHƯƠNG II: TIỀN XỬ LÍ DỮ LIỆU** 8](#_Toc102671621)

[**2.1 Mô tả dữ liệu** 8](#_Toc102671622)

[*2.1.1 Nguồn dữ liệu* 8](#_Toc102671623)

[*2.1.2 Số trường dữ liệu* 8](#_Toc102671624)

[*2.1.3 Số thuộc tính và giá trị thuộc tính* 8](#_Toc102671625)

[*2.1.4 Thống kê các giá trị thuộc tính* 9](#_Toc102671626)

[*Bảng thống kê thuộc tính*: 15](#_Toc102671627)

[*2.1.5 Số phân lớp* 18](#_Toc102671628)

[**2.2 Tiền xử lý dữ liệu** 18](#_Toc102671629)

[*2.2.1 Nhập dữ liệu và thực hiện biến đổi dữ liệu* 18](#_Toc102671630)

[*2.2.2 Thu gọn dữ liệu* 21](#_Toc102671631)

[*2.2.3 Đánh giá bộ dữ liệu* 22](#_Toc102671632)

[**CHƯƠNG III: THUẬT TOÁN VÀ THỰC NGHIỆM** 23](#_Toc102671633)

[**3.1** **Thuật toán sử dụng** 23](#_Toc102671634)

[*3.1.1* *Cây quyết định* 23](#_Toc102671635)

[*3.1.2* *Random Forest* 25](#_Toc102671636)

[*3.1.3* *Naiv Bayes* 25](#_Toc102671637)

[*3.1.3.1 Định lý Bayes* 25](#_Toc102671638)

[*3.1.3.2 Thuật toán phân lớp Naive Bayes* 27](#_Toc102671639)

[*3.1.4* *Logistic Regression* 28](#_Toc102671640)

[*3.1.5 K-nearest Neighbors* 29](#_Toc102671641)

[**3.2** **Thực nghiệm trên Jupyter Notebook** 29](#_Toc102671642)

[**3.2.1** **Cây quyết định ID3** 33](#_Toc102671643)

[**3.2.2** **Cây quyết định CART** 33](#_Toc102671644)

[**3.2.3** **Thuật toán Random Forest** 33](#_Toc102671645)

[**3.2.4** **Thuật toán Naive Bayes** 35](#_Toc102671646)

[**3.2.5** **Thuật toán Logistic Regression** 35](#_Toc102671647)

[**3.2.6** **Thuật toán K-Nearest Neighbors** 36](#_Toc102671648)

[**3.3 So sánh, đánh giá** 37](#_Toc102671649)

# **LỜI CẢM ƠN**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, đầu tiên nhóm chúng em xin gởi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới cô Cao Thị Nhạn – giảng viên lí thuyết và Thầy Vũ Minh Sang - giảng viên thực hành môn Khai thác dữ liệu đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn nhóm trong suốt quá trình làm đồ án môn học. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đồ án. Nếu không có những lời hướng dẫn, dạy bảo của Cô, Thầy thì nhóm chúng em nghĩ đồ án này của nhóm rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn Cô, Thầy. Ngoài ra, để đồ án được hoàn thành thì không thể nào cám ơn những người đã làm ra đó, cám ơn các bạn các thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ.

Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới để ứng dụng vào thực hiện đề tài **“Dự đoán khả năng chấp dứt hợp đồng làm việc của nhân viên”**. Tuy nhiên, do kiến thức và kinh nghiệm còn hạn hẹp, nhóm tác giả sẽ khó tránh khỏi những thiếu sót. Chính vì vậy, nhóm rất mong nhận được những sự góp ý từ phía Cô, Thầy nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn Cô, Thầy !

Nhóm sinh viên thực hiện

# **NHẬN XÉT CỦA GIẢNG VIÊN**

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

# **CHƯƠNG I : TỔNG QUAN ĐỀ TÀI**

1. **Đặt vấn đề**

Trong thời đại ngày nay, khi mà công nghệ 4.0 ngày càng lên ngôi, việc thao tác giao dịch tiền tệ ngày càng cần sự nhanh chóng, tiện lợi và an toàn nhất cho các bên. Cùng với sự đi lên đó, cũng là lúc mà sự phát triển của những ngân hàng và thẻ tín dụng được đẩy lên cao. Bên cạnh đó, các ngân hàng cũng đóng góp những vai trò cực kì quan trọng trong việc phát triển nền kinh tế của nước ta, như là :

* Cung cấp nhu cầu vay vốn cho sự phát triển [**kinh tế**](http://phantichtaichinh.com/hoc-phan-tich-bao-cao-tai-chinh-o-dau-tot/)
* Tạo nền: Tạo vốn cho sự phát triển kinh tế
* Biến tiết kiệm thành đầu tư
* Nâng cao hiệu quả kinh tế
* Rút ngắn tốc độ lưu thông hàng hóa và tiền tệ
* Góp phần làm cho sản xuất kinh doanh diễn ra liên tục không bị đứt quãng cung cấp vốn đầu tư và các công cụ lưu thông tín dụng [**biểu thuế xuất nhập khẩu 2020**](https://xuatnhapkhauleanh.edu.vn/bieu-thue-xuat-nhap-khau-theo-quy-dinh-moi-nhat.html)
* Tham gia kiểm soát các hoạt động kinh tế
* Tham gia vào sự ổn định của thị trường tài chính và thị trường chứng khoán. Tạo điều kiện phát triển nên những thị trường này thông qua việc chiết khấu giải quyết khả năng lưu thông nhanh của chứng khoán
* Cung cấp thông tin, tư vấn và dịch vụ đầu tư

Tuy nhiên, với sự phát triển cùng những vai trò quan trọng như vậy, nhưng các ngân hàng cũng có những nỗi lo riêng của họ. Mà 1 trong số đó, đáng lưu tâm nhất mà các giám đốc của các ngân hàng thẻ tín dụng phải đối mặt đó là tiêu hao nguồn khách hàng. Đối mặt với vấn đề này, nhóm chúng em đã đưa ra 1 giải pháp ứng dụng công nghệ thông tin và cụ thể là trong lĩnh vực Khai Thác Dữ liệu được lấy từ nguồn dữ liệu Data BankChurn, từ đó có thể tạo ra máy móc để đưa ra dự đoán về những khách hàng rời bỏ sử dụng dịch vụ tín dụng của ngân hàng, qua đó giúp những người đứng đầu có thêm cái nhìn về dịch vụ, về sự tiêu hoa nguồn khách hàng, từ đó có những quyết định cũng như là định hướng cho tương lai.

1. **Mục tiêu**

* Xây dựng hệ thống dữ liệu về ngôn ngữ tự nhiên, sử dụng máy học để huấn luyện máy móc có thể đưa ra những thông tin , những dự đoán có độ tin cậy cao phục vụ con người.
* Dự đoán khách hàng rời bỏ sử dụng thẻ tín dụng tiêu dùng, giúp hạn chế những rủi ro, tối ưu hóa công việc cho ngân hàng.

1. **Công cụ sử dụng**

Trong quá trình thực hiện, nhóm đã sử dụng một số phần mềm phục vụ cho việc tìm  hiểu và xây dựng đề tài:

* Phần thu thập và phân tích thông tin sử dụng thư viện  và ngôn ngữ lập trình python
* Dữ liệu: [Bank Churn Data Exploration And Churn Prediction | Kaggle](https://www.kaggle.com/code/thomaskonstantin/bank-churn-data-exploration-and-churn-prediction/data)

Tất cả các phần mềm trên được nhóm cài đặt và sử dụng trên Hệ điều hành Microsoft Windows 10. Việc tương thích các phần mềm trên với các hệ điều hành khác không nằm trong phạm vi nghiên cứu của đề tài này.

# **CHƯƠNG II: TIỀN XỬ LÍ DỮ LIỆU**

## **2.1 Mô tả dữ liệu**

### *2.1.1 Nguồn dữ liệu*

Tác giả: DREDLAW

### *2.1.2 Số trường dữ liệu*

Tổng số dòng dữ liệu: 49700

### *2.1.3 Số thuộc tính và giá trị thuộc tính*

Tổng số thuộc tính: 18

Đặc điểm tập dữ liệu: Đa biến

Đặc điểm số thuộc tính: ký tự, số thực, số nguyên

Giá trị bị mất: không có

### *2.1.4 Thống kê các giá trị thuộc tính*

Kí hiệu: # - số , -ký tự

Nguồn: [Predict employment termination | Kaggle](https://www.kaggle.com/code/dredlaw/predict-employment-termination/data)

Graphical user interface, chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated

### *Bảng thống kê thuộc tính*:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Thuộc tính | Ý nghĩa thuộc tính | Loại thuộc tính | Giá trị của thuộc tính | Giá trị trung bình | Giá trị trung vị | Mốt |
| 1 | EmployeeID | Mã số của khác hàng |  | Từ 1319-8336 | 4860 | 5031 |  |
| 2 | Recorddate\_key | Ngày ghi lại |  |  |  |  |  |
| 3 | Birthday\_key | Ngày sinh của khách hàng |  |  |  |  |  |
| 4 | Orighiredate\_key | Ngày thuê chính thức |  |  |  |  |  |
| 5 | Terminationdate\_key | Ngày bị sa thải |  |  |  |  |  |
| 6 | Age | Tuổi | Ordinal | 19 – 65 tuổi | 42.1 | 42 | 27 |
| 7 | Length\_of\_service | Thời gian phục vụ |  | 0 – 26 năm | 10.4 | 10 | 14 |
| 8 | City\_name | Tên công ty |  | 40 công ty |  |  | Vancouver |
| 9 | Department\_name | Tên phòng ban |  | 21 phòng ban |  |  | Meats |
| 10 | Job\_title | Chức danh |  | 47 chức danh |  |  | Meat Cutter |
| 11 | Store\_name | Tên cửa hàng |  | Từ 1 - 46 | 27.3 | 28 | 42 |
| 12 | Gender\_short | Giới tính viết tắt |  | F - M |  |  | F |
| 13 | Gender\_full | Giới tính |  | Female-Male |  |  | Female |
| 14 | Termreason\_desc | Lý do kỳ hạn | Nominal | Layoff, Not Applicable, Retirement, Resignaton |  |  | Not Applicable |
| 15 | Termtype\_desc | Thể thức kì hạn | Nominal | Involuntary, Not applicable, voluntary |  |  | Not applicable |
| 16 | Status\_Year |  |  | 2006 - 2015 | 2011 | 2011 | 2013 |
| 17 | Status | Trạng thái | Nominal | Active, Terminated |  |  | Active |
| 18 | Business\_Unit | Đơn vị kinh doanh | Nominal | Stores, Headoffice |  |  | Stores |

### *2.1.5 Số phân lớp*

Thuộc tính phân lớp: Business\_Unit, Status, Termtype\_desc, Termreason\_desc, gender\_full, gender\_short, job\_title, department\_name

## **2.2 Tiền xử lý dữ liệu**

Mục đích:

* Biến đổi dữ liệu
* Thu giảm dữ liệu
* Nhận xét dữ liệu

#### 1. Import thư viện

Graphical user interface, text, application, email

Description automatically generated

#### 2. Nhập dữ liệu

Graphical user interface

Description automatically generated with medium confidence

#### 3. Kiểm tra kiểu dữ liệu

Graphical user interface, application

Description automatically generated

#### Đánh giá tổng quan về dữ liệu:

* Dữ liệu là sự kết hợp của các giá trị chuỗi và số nguyên.
* Các cột recorddate\_key, birthdate\_key, orighiredate\_key, terminationdate\_key nên đổi từ kiểu chuỗi thành kiểu date time.
* Tuổi có thể tính bằng cách sử dụng ngày ghi(recorddate\_key) và ngày sinh(birthdate\_key). Vì vậy có thể xóa cột age.
* Thời lượng dịch vụ(length\_of\_service) có thể được tìm thấy bằng cách sử dụng ngày ghi(recorddate\_key) và ngày thuê ban đầu(orighiredate\_key). Vì vậy có thể bỏ cột length\_of\_service.
* Ngày chấm dứt sử dụng sẽ đặt mặc định là 1/1/1900 nếu như nhân viên vẫn còn trong hợp đồng.
* Tên cửa hàng(store\_name) hiển thị dưới dạng số, mặc dù nó là một tính năng phân loại dữ liệu(nominal). Bản thân tên cửa hàng không có khả năng là nguyên nhân dẫn đến việc thôi việc.
* Giới tính có 2 cột là gender\_short và gender\_full. Có thể loại bỏ 1 trong 2 cột.
* Các cột ngày chấm dứt, lý do chấm dứt và loại chấm dứt là kết quả của việc chấm dứt hợp đồng của nhân viên vì vậy không nên sử dụng cho việc đánh giá mô hình.
* Cột status\_year lặp lại năm trong ngày ghi(recorddate\_key).
* Cột STATUS là cột dữ liệu để dự đoán vì vậy phải được chuyển đổi từ chuỗi sang số.

#### Mô tả thông tin các dữ liệu kiểu số

Table

Description automatically generated

#### Mô tả thông tin các dữ liệu kiểu chuỗi

Graphical user interface, text

Description automatically generated

#### Đổi giá trị cột BUSINESS\_UNIT từ chuỗi sang dạng boolean



#### Tiến hành phân loại cho 3 cột city\_name, department\_name, job\_title

##### Cột job\_title

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

##### Cột department\_name

Table

Description automatically generated with medium confidence

A picture containing scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

##### Cột city\_names

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

#### Xử lý dữ liệu cho cột Status

# **CHƯƠNG III: THUẬT TOÁN VÀ THỰC NGHIỆM**

* 1. **Thuật toán sử dụng**
     1. *Cây quyết định*

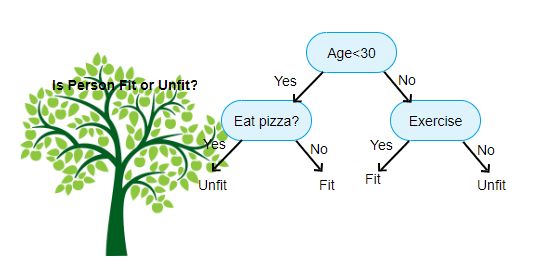
Cây quyết định (Decision Tree) là cấu trúc cây sao cho:

+ Mỗi nút trong ứng với một phép kiểm tra trên một thuộc tính

+ Mỗi nhánh biểu diễn kết quả phép kiểm tra

+ Các nút lá biểu diễn các lớp hay các phân bố lớp

+ Nút cao nhất trong cây là nút gốc. Hình dạng cây quyết định:



*Hình 3.1 Hình dạng cây quyết định*

Chiến lược cơ bản:

+  Bắt đầu từ nút đơn biểu diễn tất cả các mẫu

+  Nếu các mẫu thuộc về cùng một lớp, nút trở thành nút lá và được gán nhãn bằng lớp đó

+  Ngược lại, dùng độ đo thuộc tính để chọn thuộc tính sẽ phân tách tốt nhất các mẫu vào các lớp

+  Một nhánh được tạo cho từng giá trị của thuộc tính được chọn và các mẫu được phân hoạch theo.

+  Dùng đệ quy cùng một quá trình để tạo cây quyết định

+  Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng

* Tất cả các mẫu cho một nút cho trước đều thuộc về cùng một lớp.
* Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
* Không còn mẫu nào tại nút

**ID3** là một thuật toán được sử dụng trong cây quyết định. Thuật toán này sử dụng độ lợi thông tin (Information Gain) để tiến hành xây dựng cây quyết định**. Thuộc tính Information Gain lớn nhất sẽ được chọn làm nút gốc.**

+ Độ lợi thông tin (Information Gain):

A picture containing text, watch, clock

Description automatically generated

+ Lượng thông tin cần để phân lớp một phần tử trong S dựa trên thuộc tính A: InfoA(S)Icon

Description automatically generated with medium confidence

* Information gain chính là độ sai biệt giữa giá trị thông tin Info(S) ban đầu (trước phân hoạch) và giá trị thông tin mới InfoA(S)(sau phân hoạch với A).



+ Độ bất định (Entropy):

* Entropy đo lường lượng thông tin trong một thuộc tính của một bộ giá trị thu thập được của tập mẫu.
* Entropy được dùng để xác định nút nào được tách kế tiếp trong thuật toán.
* Entropy càng cao thì khả năng cải thiện phân lớp càng cao.
* Công thức:

Text

Description automatically generated

𝑛 là số giá trị khác nhau của thuộc tính A đang xét

𝐴i là số mẫu tương ứng với mỗi giá trị 𝑖 của thuộc tính A

𝑓S(𝐴i) là tỷ lệ của số mẫu có thuộc tính 𝐴𝑖 với S

**CART:** Khác với ID3 dùng công thức Information Cain thì thuật toán Cart sử dụng công thức Gini. **Thuộc tính có giá trị Gini nhỏ nhất sẽ được làm nút gốc**.

+ Chỉ số Gini của tập S

A picture containing text, clock, clipart

Description automatically generated

P(j|S) là tần xuất của j trong S

+ Gini của thuộc tính

A picture containing text

Description automatically generated

Trong đó: ni là số mẫu trong nốt I, n là số mẫu trong nốt A

* + 1. *Random Forest*

Thuật toán Random Forest là một thuật toán học máy có thể sử dụng để giải cả bài toán phân loại(classification) và hồi quy(regression). Nó làm việc bằng cách xây dựng một tập hợp các cây quyết định trong quá trình training, sau đó kết hợp kết quả trả về của mỗi cây để đưa ra quyết định dự đoán cuối cùng.

* + 1. *Naiv Bayes*

*3.1.3.1 Định lý Bayes*

Định lý Bayes (Bayes’ Theorem) là một định lý toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.

Định lý này đặt theo tên nhà toán học Thomas Bayes, người Anh sống ở thế kỷ 18. Đây là một trong những công cụ vô cùng hữu ích, người bạn thân của các Data Scientist, những người làm trong ngành khoa học dữ liệu.

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.

Ta có thể phát biểu lại bằng công thức như sau:

* Nếu A và B là 2 sự kiện độc lập, ta có xác suất để xảy ra A và B đồng thời là:

P(A,B) = P(A) P(B)

Trong đó:

* P(A)P(A) là xác suất xảy ra A riêng biệt.
* P(B)P(B) là xác suất xảy ra B riêng biệt.
* Nếu A và B là 2 sự kiện liên quan đến nhau, và xác suất xảy ra sự kiện B lớn hơn 0, ta có thể định nghĩa xác suất xảy ra A khi biết B xảy ra như sau:Text

  Description automatically generated

* Định lý Bayes dựa trên định nghĩa về xác suất có điều kiện ở trên, được phát biểu dưới dạng công thức như sau:A picture containing text

  Description automatically generated

Kí hiệu ¬A là không A (hay bù A). Ta có P(A)+P(¬A) = 1.

Từ đó: P(B) =P(B,A) + P(B,¬A) = P(B∣A)P(A) + P(B∣¬A)P(¬A)

Định lý Bayes được viết dưới dạng biến thể như sau: A picture containing shape

Description automatically generated

*3.1.3.2 Thuật toán phân lớp Naive Bayes*

**Naive Bayes Classification** (NBC) là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes. Thuật toán này thuộc nhóm thuật toán học có giám sát.

* Mỗi mẫu dữ liệu được biểu diễn bằng X=(x1, x2,..., xn) với các thuộc tính A1, A2,..., An
* Các lớp C1, C2, ..., Cm. Cho trước mẫu chưa biết X
* Phân lớp Naive Bayes sẽ xác định X thuộc lớp Ci nếu và chỉ nếu: Logo, company name

  Description automatically generated with medium confidence

* Theo định lý BayesA picture containing text

  Description automatically generated

* Do P(X) là hằng số cho tất cả các lớp, chỉ cần cực đại P(X|Ci) x P(Ci). Nếu chưa biết P(Ci) cần giả định P(C1)=P(C2)=...= P(Cm) và chúng ta sẽ cực đại P(X|Ci). Ngược lại, ta cực đại P(X|Ci) x P(Ci)
* Tuy nhiên vấn đề tính P(X|Ci) là bất khả thi!

* **Thừa nhận Naive:** giả định sự độc lập thuộc tính

Diagram

Description automatically generated

Có thể tính xấp xỉ P(x1|Ci), ..., P(xn|Ci) từ các mẫu huấn luyện

+ Nếu Ak là thuộc tính định tính thì P(xk|Ci) = sik/si với sik là số mẫu huấn luyện của Ci có trị xk cho Ak và si là số các mẫu thuộc về lớp Ci

+ Nếu Ak là liên tục thì nó được giả định có phân bố Gaussian:

Text

Description automatically generated

* + 1. *Logistic Regression*

Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu  ra rời rạc (discrete target variable) y ứng với một véc-tơ đầu vào **x**. Việc này tương đương với chuyện phân loại các đầu vào **x** vào các nhóm y tương ứng.

Phân tích hồi qui logistic là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập (biến số hoặc biến phân loại) với biến phụ thuộc là biến nhị phân. Trong hồi qui tuyến tính đơn, biến độc lập x và phụ thuộc y là biến số liên tục liên hệ qua phương trình:

y = a + bx + e

Trong đó: a là alpha, b là beta, e là epxilon

Trong hồi qui logistic, biến phụ thuộc y chỉ có 2 trạng thái 1 và 0. Muốn đổi ra biến số liên tục người ta tính xác suất của 2 trạng thái này. Nếu gọi p là xác suất để một biến cố xảy ra, thì 1-p là xác suất để biến cố không xảy ra. Phương trình hồi qui logistic phát biểu:A picture containing text, watch, clock

Description automatically generated

Ta tính được xác xuất:Chart

Description automatically generated

*3.1.5 K-nearest Neighbors*

* K-nearest Neighbors là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai thác dữ liệu. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học (nên KNN được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới.
* Lớp (nhãn) của một đối tượng dữ liệu mới có thể dự đoán từ các lớp (nhãn) của k hàng xóm gần nó nhất.
* Cho trước tập dữ liệu huấn luyện D với các lớp, phân lớp X vào các lớp dựa vào k phần tử tương tự với X nhất (quy tắc bỏ phiếu: majority vote).
* Các bước thực hiện:

+  Ta có D là tập các điểm dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.

+  Đo khoảng cách (Euclidian, Manhattan, Minkowski, Minkowski hoặc Trọng số) từ dữ liệu mới A đến tất cả các dữ liệu khác đã được phân loại trong D.

+  Chọn K (K là tham số mà bạn định nghĩa) khoảng cách nhỏ nhất.

+  Kiểm tra danh sách các lớp có khoảng cách ngắn nhất và đếm số lượng của mỗi lớp xuất hiện.

+  Lấy đúng lớp (lớp xuất hiện nhiều lần nhất).

+  Lớp của dữ liệu mới là lớp đã nhận được ở bước 5.

* Phụ thuộc:

+  Độ đo khoảng cách để xác định sự tương tự.

+  Trị k, số phần tử láng giềng => k<=|D|1/2

* Độ đo Euclidean

A picture containing text, watch, gauge

Description automatically generated

## **Thực nghiệm trên Jupyter Notebook**

Vẽ biểu đồ, đếm và xem tỉ lệ nhãn để có cái nhìn tổng quát về khả năng rời bỏ sử dụng dịch vụ thẻ tín dụng của khách hàng

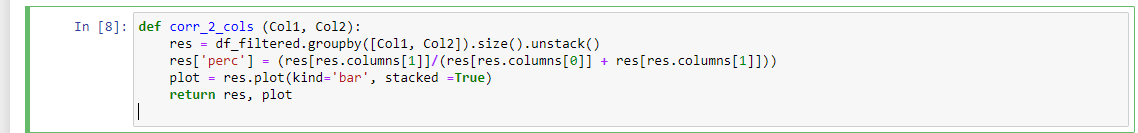
Xây dựng thuộc tính quyết định, với thuộc tính quyết định là **Attrition\_Flag**.

A picture containing chart

Description automatically generated

Xét mối tương quan của từng thuộc tính với thuộc tính quyết định

*Hàm tính Correlation:*



Perc (phần trăm) = (Số giá trị A của thuộc tính đang xét mang nhãn ‘yes’ của thuộc tính quyết định) / (Tổng số giá trị A của thuộc tính đang xét)

Xét thuộc tính Education\_Level

Chart

Description automatically generated

Xét thuộc tính Martial\_Status

Chart

Description automatically generated

Xét thuộc tính Income\_Category

Chart

Description automatically generated

Xét thuộc tính Card\_Category

Graphical user interface

Description automatically generated

Xét thuộc tính Total\_Relationship\_Count

Graphical user interface

Description automatically generated

* + 1. **Cây quyết định ID3**

**Bước 1:** Xây dựng mô hình và lưu lại thời gian chạy (0.0438861)

Graphical user interface, text, application

Description automatically generated

**Bước 2:** Xuất báo cáo để xem độ chính xác và độ phủ của mô hình đã xây dựng:

* Độ chính xác của mô hình là 94%
* Độ nhạy mô hình là 80%
* Độ đặc hiệu mô hình là 96%

Graphical user interface, text, application

Description automatically generated

**Bước 3:** Tính toán ma trận nhầm lẫn và biểu diễn bằng heatmap:

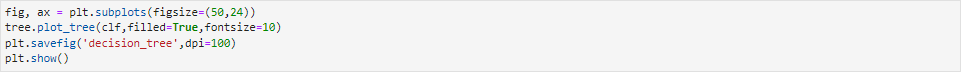
Graphical user interface, text, application

Description automatically generated

Chart, treemap chart

Description automatically generated

**Bước 4:** Xây dựng cây ID3 và lưu lại file ‘decision\_tree.png’ với chất lượng tốt:



Chart, scatter chart

Description automatically generated

**Cây quyết định CART**

**Bước 1:** Xây dựng mô hình và lưu lại thời gian chạy (0.0368033)

A picture containing chart

Description automatically generated

**Bước 2:** Xuất báo cáo để xem độ chính xác và độ phủ của mô hình đã xây dựng:

* Độ chính xác của mô hình là 93%
* Độ nhạy mô hình là 75%
* Độ đặc hiệu mô hình là 96%

Graphical user interface, text, application, email

Description automatically generated

**Bước 3:** Tính toán ma trận nhầm lẫn và biểu diễn bằng heatmap:

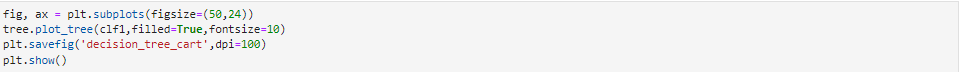
Graphical user interface, application

Description automatically generated

Chart, treemap chart

Description automatically generated

**Bước 4:** Xây dựng cây CART và lưu lại file ‘decision\_tree\_cart.png’ với chất lượng tốt:



Chart, scatter chart

Description automatically generated

* + 1. **Thuật toán Random Forest**

**Bước 1**: Xây dựng mô hình và lưu thời gian chạy

A screenshot of a computer

Description automatically generated with medium confidence

Thời gian chạy là : 0.7156193256378174s

**Bước 2** : Xuất báo cáo để xem độ chính xác và độ phủ của mô hình đã xây dựng:

Text

Description automatically generated

Table

Description automatically generated

* Độ chính xác của mô hình là 96%
* Độ nhạy mô hình là 85%
* Độ đặc hiệu mô hình là 98%

**Bước 3** : Tính toán ma trận nhầm lẫn và biểu diễn bằng Heatmap

Chart, treemap chart

Description automatically generated

* + 1. **Thuật toán Naive Bayes**
    2. **Thuật toán Logistic Regression**

Xây dựng mô hình và xem thời gian chạy:

Bước 1: Kiểm thử và cài đặt cho thuật toán Logistic Regression

* + 1. **Thuật toán K-Nearest Neighbors**

Xây dựng mô hình và xem thời gian chạy:

**Bước 1**: Kiểm thử và cài đặt cho thuật toán K-nearest Neighbors

Graphical user interface, text, application

Description automatically generated

**Bước 2**: Xây dựng mô hình và lưu lại thời gian chạy

Thời gian chạy : 2.7344095706939697s

Text

Description automatically generated

**Bước 3**: Xuất báo cáo để xem độ chính xác và độ phủ của mô hình đã xây dựng:

Table

Description automatically generated with medium confidence

* Độ chính xác của mô hình là 84%
* Độ nhạy mô hình là 10%
* Độ đặc hiệu mô hình là 91%

**Bước 4** : Tính toán ma trận nhầm lẫn và biểu diễn bằng Heatmap

Graphical user interface, application

Description automatically generated

## **3.3 So sánh, đánh giá**