**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. Nguyễn Thanh Bình**

**ThS. Nguyễn Hồ Duy Trí**

Nhóm thực hiện: **Nhóm 18**

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. Nguyễn Thanh Bình**

**ThS. Nguyễn Hồ Duy Trí**

Nhóm thực hiện: **Nhóm 18**

Thành Viên:

Hồ Trọng Khang 19521661

Nguyễn Cao Khoa 19521694

**TP Hồ Chí Minh tháng 5 năm 2022**

**MỤC LỤC**

[**LỜI CẢM ƠN** 4](#_Toc104455555)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 5](#_Toc104455556)

[**CHƯƠNG I : TỔNG QUAN ĐỀ TÀI** 6](#_Toc104455557)

[**1.** **Đặt vấn đề** 6](#_Toc104455558)

[**2.** **Mục tiêu** 7](#_Toc104455559)

[**3.** **Công cụ sử dụng** 7](#_Toc104455560)

[**CHƯƠNG II: TIỀN XỬ LÍ DỮ LIỆU** 8](#_Toc104455561)

[2.1 Mô tả dữ liệu 8](#_Toc104455562)

[*2.1.1 Nguồn dữ liệu* 8](#_Toc104455563)

[*2.1.2 Số trường dữ liệu* 8](#_Toc104455564)

[*2.1.3 Số thuộc tính và giá trị thuộc tính* 8](#_Toc104455565)

[*2.1.4 Thống kê các giá trị thuộc tính* 8](#_Toc104455566)

[*Bảng thống kê thuộc tính*: 12](#_Toc104455567)

[*2.1.5 Số phân lớp* 14](#_Toc104455568)

[2.2 Tiền xử lý dữ liệu 14](#_Toc104455569)

[**CHƯƠNG III: THUẬT TOÁN VÀ THỰC NGHIỆM** 33](#_Toc104455570)

[1. Thuật toán sử dụng 33](#_Toc104455571)

[1.1 Decision Tree 33](#_Toc104455572)

[1.2 Random Forest 35](#_Toc104455573)

[1.3 Naive Bayes 35](#_Toc104455574)

[1.4 K-nearest Neighbors 38](#_Toc104455575)

[2. Thực nghiệm trên Jupyter Notebook 39](#_Toc104455576)

[1. Tách cột thuộc tính quyết định ra 1 cột riêng 39](#_Toc104455577)

[2. Tách dữ liệu train và test (dữ liệu train chiếm 70%, test chiếm 30%) 39](#_Toc104455578)

[3. Thuật toán K-Nearest Neighbors (KNN) 39](#_Toc104455579)

[4. Thuật toán Random Forest 42](#_Toc104455580)

[5. Thuật toán Decision Trees 44](#_Toc104455581)

[6. Thuật toán Naive Bayes 49](#_Toc104455582)

[7. So sánh, đánh giá 50](#_Toc104455583)

[***Link tham khảo*** 52](#_Toc104455584)

# **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 thầy Nguyễn Thanh Bình – giảng viên lí thuyết và Thầy Nguyễn Hồ Duy Trí - giảng viên thực hành môn Hệ hỗ trợ quyết định 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 hai 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 hai 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 hai 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 một tổ chức, công ty hay là doanh nghiệp thì trong quá trình hoạt động, họ không thể giữ vững mãi được các nguồn lao động, các nhân viên của mình. Hiện nay đất nước ta đang trong giai đoạn công nghiệp hóa, hiện đại hóa, đổi mới mô hình, cơ cấu nền kinh tế, nâng cao năng suất lao động, sức cạnh tranh của doanh nghiệp, các tổ chức; đóng góp to lớn nâng cao tiềm lực kinh tế đất nước. Trong đó người lao động đóng vai trò quan trọng và là nòng cốt để đưa công cuộc đổi mới đi vào chiều sâu, đẩy mạnh công nghiệp. Cho nên việc có cho mình một nguồn lao động vững chắc là điều vô cùng cần thiết bởi vì việc chấm dứt hợp đồng là điều không thể tránh khỏi và do hậu quả của việc chấm dứt hợp đồng lao động có liên quan đến vấn đề về kinh tế và xã hội vì mỗi con người dù ở mọi vị trí, mọi đơn vị khác nhau đều là những chủ thể quan trọng trong việc sáng tạo và phát huy những lợi thế của tổ chức. Nếu tổ chức phát triển thì sẽ giúp cho nền kinh tế tốt hơn. Cho nên khi có người chấm dứt hợp đồng ở một vị trí nào đó thì cần có một người khác để thay thế vào vị trí đó, đặc biệt là những vị trí quan trọng trong tổ chức thì việc tìm ra người phù hợp trong thời gian ngắn là rất khó khăn.
* 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 Employment Termination, từ đó có thể tạo ra máy móc để đưa ra dự đoán về những nhân viên có thể sẽ chấm dứt hợp đồng, qua đó giúp những người đứng đầu có thêm cái nhìn về những vấn đề trong công ty như là môi trường làm việc, vị trí của mỗi nhân viên, về sự rời đi của nhân viên, từ đó có những quyết định, định hướng trong tương lại cũng như là có sự chuẩn bị sớm hơn về nhân lực của mình.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/)

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 khả năng nhân viên rời bỏ công ty, giúp công ty nhận ra vấn đề và khắc phục sớm để phát triển cho tương lai.

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: [Predict employment termination | Kaggle](https://www.kaggle.com/code/dredlaw/predict-employment-termination/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 khách 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 thành phố nhân viên làm việc |  | 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

*Hình 1. Các thư viện cần thiết*

#### 2. Nhập dữ liệu

Graphical user interface

Description automatically generated with medium confidence

*Hình 2. Sử dụng thư viện Pandas để nhập dữ liệu dạng csv*

#### 3. Kiểm tra kiểu dữ liệu

Graphical user interface, application

Description automatically generated

*Hình 3. Kiểm tra kiểu dữ liệu cho toàn bộ thuộc tính*

#### 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

*Hình 4. Thông tin dữ liệu kiểu số*

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

Graphical user interface, text

Description automatically generated

*Hình 5. Thông tin dữ liệu kiểu chuỗi*

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



*Hình 6. Đổi kiểu dữ liệu sang số cho cột BUSINESS\_UNIT*

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

##### Cột job\_title

Text

Description automatically generated with medium confidence

*Hình 7.1. Phân loại cho cột Job\_title*

Text, letter

Description automatically generated

*Hình 7.2. Phân loại cho cột Job\_title*

Graphical user interface, text, application

Description automatically generated

*Hình 7.3. Phân loại cho cột Job\_title*

Text

Description automatically generated

*Hình 7.4. Phân loại cho cột Job\_title*

##### Cột department\_name

Table

Description automatically generated with medium confidence

*Hình 8.1. Phân loại cho cột department\_name*

A picture containing scatter chart

Description automatically generated

*Hình 8.2. Phân loại cho cột department\_name*

Graphical user interface, text, application

Description automatically generated

*Hình 8.3. Phân loại cho cột department\_name*

Text

Description automatically generated

*Hình 8.4. Phân loại cho cột department\_name*

##### Cột city\_names

Graphical user interface, text

Description automatically generated

*Hình 9.1 Phân loại cho cột city\_names*

Text

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

*Hình 9.2 Phân loại cho cột city\_names*

Text

Description automatically generated with medium confidence

*Hình 9.3 Phân loại cho cột city\_names*

Text

Description automatically generated

*Hình 9.4 Phân loại cho cột city\_names*

Graphical user interface, text, application

Description automatically generated

*Hình 9.5 Phân loại cho cột city\_names*

Text

Description automatically generated

*Hình 9.6 Phân loại cho cột city\_names*

Graphical user interface, text, application

Description automatically generated

*Hình 9.7 Phân loại cho cột city\_names*

#### Xử lý dữ liệu cho cột Status

Graphical user interface, text, application

Description automatically generated

*Hình 10. Xử lý dữ liệu cho cột Status*

#### Sử dụng Matplot để đánh giá về vấn đề ở lại hay rời công ty của nhân viên dựa vào độ tuổi (Age) và thời gian làm việc (Length\_of\_service)

Chart

Description automatically generated

*Hình 11. Biểu đồ Matplot*

Nhận xét biểu đồ: Nhân viên có thể rời công ty sau khi làm việc trong bất kì khoảng thời gian nào từ 0 đến 25 năm và mọi lứa tuổi từ 20 đến 60

#### Đánh giá kỹ hơn về vấn đề độ tuổi và thời gian làm lên việc chấm dứt hợp đồng của nhân viên

Text, Word

Description automatically generated

Chart, scatter chart

Description automatically generated

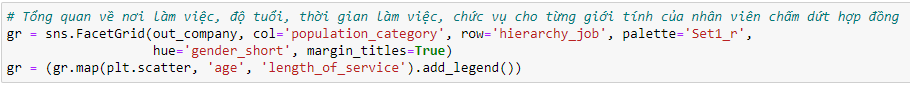
*Hình 12. Biểu đồ Jointplot*

**Tổng quan về biểu đồ:**

Có 5 cột đỉnh về độ tuổi, 4 cột đỉnh về thời gian làm việc

* Đỉnh tuổi lớn nhất: Độ tuổi > 61 trùng với đỉnh cột thời gian làm việc khoảng 25 năm. Đây là những người sắp nghỉ hưu.
* Đỉnh tuổi thứ hai: Độ tuổi khoảng 58 – 61 tuổi trùng với cột thời gian làm việc khoảng 19 – 21 năm. Đây cũng có thể là những người sắp nghỉ hưu.
* Đỉnh tuổi thứ ba: độ tuổi khoảng 19-24 tuổi trùng với đỉnh cột thời gian làm việc khoảng 0 năm. Đây có thể là những người đang thử việc để tìm kiếm thứ họ mong muốn.
* Đỉnh tuổi thứ tư: độ tuổi khoảng 25-27 tuổi trùng với cột thời gian làm việc khoảng 1-3 năm. Đây có thể là những người cảm thấy không phù hợp với công việc hoặc môi trường làm việc.
* Đỉnh tuổi thứ năm: độ tuổi khoảng 28-31 tuổi trùng với đỉnh cột thời gian làm việc là 8 năm. Đây có thể là những người đã cảm thấy mệt mỏi với công việc của họ và muốn thay đổi nghề nghiệp. Cũng có thể là những người có những lí do gia đình buộc họ phải thay đổi.

#### Tổng quan về nơi làm việc, độ tuổi, thời gian làm việc, chức vụ cho từng giới tính của nhân viên chấm dứt hợp đồng



Chart, scatter chart

Description automatically generated

*Hình 13 Biểu đồ lưới (Grid)*

#### Tổng quan về nơi làm việc, độ tuổi, thời gian làm việc, chức vụ cho từng giới tính của nhân viên còn hợp đồng

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing device

Description automatically generated

*Hình 14. Biểu đồ lưới (Grid)*

**Nhận xét về 2 biểu đồ:**

* Không có nhân viên nào nghỉ việc khi họ đang làm ở chức vụ Giám đốc điều hành
* Giám đốc điều hành và thành viên hội đồng quản trị chỉ làm việc trong các thành phố
* Dường như không có sự khác biệt lớn trong việc chấm dứt việc làm giữa nam và nữ
* Các nhà quản lí và thành viên hội đồng quản trị chỉ ngừng việc sau ít nhất 14 năm phục vụ. Điều này có nghĩa là họ có khả năng đã được đề bạt nội bộ cho các vị trí đó

#### Tổng quan về nơi làm việc, độ tuổi, thời gian làm việc. Lý do chấm dứt hợp đồng cho việc nhân viên đó có tự nguyện không khi chấm dứt hợp đồng.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Hình 15. Biểu đồ lưới (Grid)*

**Nhận xét về biểu đồ:**

* Việc sa thải xảy ra cho lứa tuổi và tất cả các thời gian dịch vụ ở vùng xa xôi hẻo lánh và nông thôn
* Từ chức là không phổ biến ở các vùng xa xôi.
* Đúng như dự đoán, sa thải là không tự nguyện, trong khi việc từ chức và nghỉ hưu là tự nguyện.

#### Thống kê số lượng nhân viên chấm dứt hợp đồng mỗi năm

Chart, bar chart

Description automatically generated

*Hình 16. Thống kê số nhân viên rời công ty mỗi năm*

**Nhận xét về biểu đồ**

* Năm 2014 có số lượng chấm dứt việc làm cao bất thường.
* Ngoài năm 2014 ra thì cột cao thứ hai là năm 2008, do năm 2008 xảy ra cuộc khủng hoàn kinh tế toàn cầu (GFC)
* Năm 2015 nhiều tương tự như cột 2008, nhưng thấp hơn nhiều so với đỉnh 2014.

#### Tiến hành xóa các cột không cần thiết

Text, letter

Description automatically generated

*Hình 17. Xóa các cột không cần thiết*

Table

Description automatically generated

*Hình 18. Sau khi đã tiến hành các cột không cần thiết*

#### Sử dụng phương thức get\_dummies để chuyển đổi cột giới tính (Gender\_short), bộ phận phục vụ (department\_service) từ chuỗi sang số

Text, Word

Description automatically generated

*Hình 19. Chuyển đồ kiểu dữ liệu cho 2 cột Gender\_short và department\_service*

Table

Description automatically generated

*Hình 20. Sau khi đã xử lý dữ liệu 2 cột gender\_short và department\_service*

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

## Thuật toán sử dụng

### Decision Tree

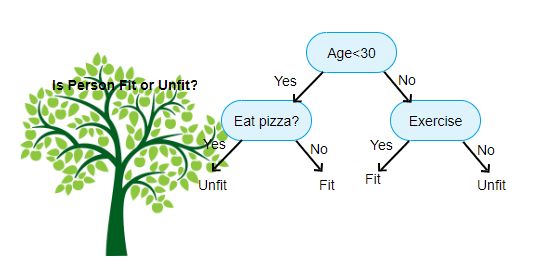
Cây quyết định (Decision Tree) là cấu trúc cây sao cho:

+ 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

### 1.2 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.

### 1.3 Naive Bayes

#### 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

#### 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

### 1.4 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 khỏi công ty của nhân viên

Xây dựng thuộc tính quyết định, với thuộc tính quyết định là **STATUS**

### Tách cột thuộc tính quyết định ra 1 cột riêng

Text

Description automatically generated with low confidence

*Hình 20. Tách cột thuộc tính quyết định*

### Tách dữ liệu train và test (dữ liệu train chiếm 70%, test chiếm 30%)

Text

Description automatically generated with medium confidence

*Hình 21. Tách dữ liệu ra thành 2 bộ train và test*

### Thuật toán K-Nearest Neighbors (KNN)

Text, letter

Description automatically generated

*Hình 22.1 Thuật toán KNN*

Table

Description automatically generated

*Hình 22.2 Kết quả của mô hình thuật toán KNN*

* Thời gian chạy của thuật toán KNN: 0.6912s
* Độ chính xác: 98.677%

Graphical user interface, text

Description automatically generated

*Hình 22.3 Thuật toán KNN*

Chart, treemap chart

Description automatically generated

*Hình 22.4 Ma trận nhầm lẫn của mô hình thuật toán KNN*

* Thông qua ma trận nhầm lẫn của mô hình thuật toán KNN, ta biết được
* Độ nhạy (Precision) của mô hình thuật toán: 98.707%
* Độ đặc hiệu (Recall) của mô hình thuật toán: 99.840%
* F1-Score: 99.270%

### Thuật toán Random Forest

A picture containing text

Description automatically generated

*Hình 23.1 Thuật toán Random Forest và kết quả*

* Độ chính xác của mô hình thuật toán Random Forest: 98.630%
* Thời gian chạy của thuật toán Random Forest: 1.6293s

Text

Description automatically generated

*Hình 23.2 Thuật toán KNN*

Chart

Description automatically generated

*Hình 23.3 Ma trận nhầm lẫn của mô hình thuật toán Random Forest*

* Thông qua ma trận nhầm lẫn của mô hình thuật toán Random Forest, ta biết được
* Độ nhạy (Precision) của mô hình thuật toán: 98.76%
* Độ đặc hiệu (Recall) của mô hình thuật toán: 99.840%
* F1-Score: 99.297%

### Thuật toán Decision Trees

#### Cây ID3

Table

Description automatically generated

*Hình 24.1 Thuật toán Decision Tree(ID3)*

* Độ chính xác của thuật toán: 98.570%
* Thời gian chạy của thuật toán: 0.0449s

Text

Description automatically generated

*Hình 24.2 Thuật toán Decision Tree(ID3)*

Chart, treemap chart

Description automatically generated

*Hình 24.3 Ma trận nhầm lẫn của mô hình thuật toán Decision Tree(ID3)*

* Thông qua ma trận nhầm lẫn của mô hình thuật toán Decision Tree(ID3), ta biết được
* Độ nhạy (Precision) của mô hình thuật toán: 98.773%
* Độ đặc hiệu (Recall) của mô hình thuật toán: 99.76%%
* F1-Score: 99.26%

Text

Description automatically generated

A picture containing text, dark

Description automatically generated

*Hình 24.4 Cây Decision Tree(ID3)*

#### Cây Cart

Table

Description automatically generated with medium confidence

*Hình 25.1 Thuật toán Decision Tree(CART)*

* Độ chính xác của thuật toán: 98.563%
* Thời gian chạy của thuật toán: 0.0529ss

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

*Hình 25.2 Ma trận nhầm lẫn của mô hình thuật toán Decision Tree(CART)*

* Thông qua ma trận nhầm lẫn của mô hình thuật toán Decision Tree(CART), ta biết được
* Độ nhạy (Precision) của mô hình thuật toán: 98.773%
* Độ đặc hiệu (Recall) của mô hình thuật toán: 99.757%%
* F1-Score: 99.26%

Text

Description automatically generated

A picture containing chart

Description automatically generated

*Hình 25.1 Cây Decision Tree(CART)*

### Thuật toán Naive Bayes

Table

Description automatically generated

*Hình 26.1 Thuật toán Naïve Bayes*

* Độ chính xác của thuật toán: 93.709%
* Thời gian chạy của thuật toán: 0.021s

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

*Hình 26.2 Ma trận nhầm lẫn của mô hình thuật toán Naïve Bayes*

* Thông qua ma trận nhầm lẫn của mô hình thuật toán Naive Bayes, ta biết được
* Độ nhạy (Precision) của mô hình thuật toán: 97.292%
* Độ đặc hiệu (Recall) của mô hình thuật toán: 96.195%%
* F1-Score: 96.74%

### So sánh, đánh giá

*Sử dụng đồ thị BarPlot để có cái nhìn tổng quan về thời gian chạy và độ chính xác giữa các thuật toán*

* Vẽ biểu đồ so sánh thời gian chạy của các thuật toán

A picture containing text

Description automatically generated

Chart, bar chart

Description automatically generated

*Hình 27 Biểu đồ thời gian chạy của các thuật toán*

**Kết luận về biểu đồ thời gian chạy:**

* Thuật toán Naïve Bayes là thuật toán có thời gian chạy nhanh nhất đối với tập dữ liệu. Chỉ với 0.0196s
* Thuật toán Random Forest là chậm nhất với 1.5772s
* Vẽ biểu đồ so sánh độ chính xác của các thuật toán

A picture containing text

Description automatically generated

Chart, bar chart

Description automatically generated

*Hình 28. Biểu đồ độ chính xác của các thuật toán*

**Kết luận về biểu đồ độ chính xác:**

* Các thuật toán đều cho kết quả độ chính xác rất cao, cân bằng với nhau, hầu hết là độ chính xác bằng 99%
* Thuật toán Naïve Bayes có độ chính xác thấp nhất trong số 5 thuật toán, với độ chính xác là 93.7%

# ***Link tham khảo***

**Ma trận nhầm lẫn:** [Precision, Recall và F1-score là gì? – Cái Hữu Thức's notes (wordpress.com)](https://caihuuthuc.wordpress.com/2020/02/23/precision-recall-va-f1-score-la-gi/)

**Tìm hiểu thuật toán:** [CART: Classification and Regression Trees for Clean but Powerful Models | by Saul Dobilas | Towards Data Science](https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85)

**Các kiểu dữ liệu:** [What Is Ordinal Data? [Definition, Analysis & Examples] (careerfoundry.com)](https://careerfoundry.com/en/blog/data-analytics/what-is-ordinal-data/)

**Các kiểu máy học:** [4 Machine Learning Approaches that Every Data Scientist Should Know | by Orhan G. Yalçın | Towards Data Science](https://towardsdatascience.com/4-machine-learning-approaches-that-every-data-scientist-should-know-e3a9350ec0b9#:~:text=Unsupervised%20learning%20is%20a%20type%20of%20machine%20learning,datasets%2C%20which%20do%20not%20contain%20labels.%20Figure%204.)