ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**NGÔ GIA KHANG**

**ĐINH HỮU AN**

**Ứng dụng Machine Learning trong**

**dự đoán rủi ro nghỉ việc của nhân viên**

**KHÓA LUẬN TỐT NGHIỆP**

**NGÀNH: CÔNG NGHỆ THÔNG TIN**

**tRÌNH ĐỘ ĐÀO TẠO: ĐẠI HỌC**

**TP. HỒ CHÍ MINH, THÁNG 0 NĂM 2026**

ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**NGÔ GIA KHANG**

**ĐINH HỮU AN**

**Ứng dụng Machine Learning trong**

**dự đoán rủi ro nghỉ việc của nhân viên**

**KHÓA LUẬN TỐT NGHIỆP**

**NGÀNH: CÔNG NGHỆ THÔNG TIN**

**tRÌNH ĐỘ ĐÀO TẠO: ĐẠI HỌC**

NGƯỜI HƯỚNG DẪN: . PHẠM THI VƯƠNG

**TP. HỒ CHÍ MINH, THÁNG 0 NĂM 2026**

# LỜI CAM ĐOAN

*Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng tôi, các số liệu và kết quả nghiên cứu nêu trong khóa luận là trung thực và chưa từng được công bố trong bất kì một công trình nào khác.*

Tác giả khóa luận

|  |  |
| --- | --- |
| **Đinh Hữu An** | **Ngô Gia Khang** |

# LỜI CẢM ƠN

# NHẬN XÉT CỦA GIẢNG VIÊN

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**……………………………………………………………………………………………………………**

**Hồ Chí Minh, ngày.....tháng.....năm 2025**

**Giảng viên hướng dẫn**

**Phạm Thi Vương**

# MỤC LỤC

[**LỜI CAM ĐOAN** 2](#_Toc206237187)

[**LỜI CẢM ƠN** 3](#_Toc206237188)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 4](#_Toc206237189)

[**MỤC LỤC** 5](#_Toc206237190)

[**DANH MỤC THUẬT NGỮ** 6](#_Toc206237191)

[**DANH MỤC CÁC BẢNG** 7](#_Toc206237192)

[**DANH MỤC HÌNH ẢNH** 8](#_Toc206237193)

[**CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI** 9](#_Toc206237194)

[1.1. Tổng quan 9](#_Toc206237195)

[1.2. Tính cấp thiết 9](#_Toc206237196)

[1.3. Mục tiêu, đối tượng và phạm vi nghiên cứu 9](#_Toc206237197)

[1.4. Phương pháp nghiên cứu 9](#_Toc206237198)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN NGHIÊN CỨU** 10](#_Toc206237199)

[2.1. Tổng quan về đánh giá dự đoán nghỉ việc của nhân sự 10](#_Toc206237200)

[2.2. Các phương pháp đánh giá và ứng dụng các thuật toán học máy 10](#_Toc206237201)

[2.2.1. Phương pháp đánh giá truyền thống 10](#_Toc206237202)

[2.2.2. Ứng dụng thuật toán học máy 10](#_Toc206237203)

[2.2.3. Giới thiệu về tập dữ liệu mẫu và các mô hình dự đoán 10](#_Toc206237204)

[**CHƯƠNG 3: TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU** 11](#_Toc206237205)

[3.1. Tóm tắt tập dữ liệu mẫu 11](#_Toc206237206)

[3.1.1. Các trường thuộc tính 11](#_Toc206237207)

[3.1.2. Các giá trị 11](#_Toc206237208)

[3.2. Các bước làm sạch và tiền xử lý dữ liệu 11](#_Toc206237209)

[3.2.1. Kiểm tra và xử lý dữ liệu thiếu 11](#_Toc206237210)

[3.2.2. Giảm chiều dữ liệu 11](#_Toc206237211)

[3.2.3. Mã hóa biến phân loại 11](#_Toc206237212)

[3.3. Phân tích dữ liệu 11](#_Toc206237213)

[3.3.1. Phân tích tương quan giữa các biến trong bộ dữ liệu 11](#_Toc206237214)

[3.3.2. Phân tích đơn biến, đa biến 11](#_Toc206237215)

[3.3.3. Phân tích như cầu tăng cường mẫu của tập dữ liệu 11](#_Toc206237216)

[**CHƯƠNG 4: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN** 12](#_Toc206237217)

[4.1. Phân chia dữ liệu huấn luyện và kiểm tra 12](#_Toc206237218)

[4.2. Huấn luyện mô hình 12](#_Toc206237219)

[**CHƯƠNG 5: THỬ NGHIỆM VÀ ĐÁNH GIÁ** 13](#_Toc206237220)

[5.1. Kiểm thử mô hình 13](#_Toc206237221)

[5.2. Kết quả đạt được 13](#_Toc206237222)

[5.3. Hạn chế 13](#_Toc206237223)

[**CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 14](#_Toc206237224)

[6.1. Kết luận 14](#_Toc206237225)

[6.2. Hướng phát triển 14](#_Toc206237226)

# DANH MỤC THUẬT NGỮ

# DANH MỤC CÁC BẢNG

# DANH MỤC HÌNH ẢNH

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## Lời mở đầu

Trong bối cảnh thời đại chuyển đổi số, nền kinh tế hội nhập và cạnh tranh ngày càng gay gắt, nguồn nhân lực đóng vai trò quan trọng then chốt trong kết quả kinh doanh và sự thành công của doanh nghiệp, hơn nữa đó là tài sản quan trọng nhất, là cơ sở nền tảng phát triển và tồn tại của mỗi doanh nghiệp. Việc ứng dụng công nghệ công tin vào việc quản lý nhân sự trở thành một xu thế tất yếu, phổ biến. Quản lý nhân sự không còn là quản lý thông tin nhân viên hay chấm công mà còn quản lý cả hiệu suất làm việc của nhân viên giúp theo dõi tiến trình công việc một cách chủ động, hiệu quả và đáng tin cậy bằng các phương pháp đánh giá hiệu suất công việc.

Với sự phát triển mạnh mẽ của công nghệ trí tuệ nhân tạo, đặc biệt là Machine Learning (ML), các tổ chức hiện nay có thể ứng dụng các thuật toán học máy để phân tích hiệu suất làm việc và dự đoán nguy cơ nghỉ việc từ dữ liệu nhân sự. Tuy nhiên, tại Việt Nam, việc ứng dụng các thuật toán học máy trong quản trị nhân sự vẫn còn mới mẻ và chưa được khai thác sâu rộng, đặc biệt ở các doanh nghiệp vừa và nhỏ.

Qua đề tài này đề xuất các mô hình dự đoán thông qua việc xử lý dữ liệu và chọn lọc các đặc trưng cùng với các thuật toán học máy phù hợp để dự đoán rủi ro nghỉ việc của nhân viên. Từ những kết quả phân tích và dự đoán có thể xây dựng các chiến lược quản lý nhân sự mới, chủ động dự đoán nhu cầu lực lượng lao động và giảm thiểu nguy cơ tiềm năng trong tương lai.

Chính vì vậy, đề tài này mang tính cấp thiết cả về mặt lý thuyết lẫn thực tiễn, góp phần nâng cao hiệu quả quản lý nhân sự bằng công nghệ hiện đại.

## Tổng quan về đề tài

* **Lý do chọn đề tài:**
* Xuất phát từ thực tiễn tỷ lệ nghỉ việc ngày càng tăng tại nhiều doanh nghiệp, đặc biệt là trong bối cảnh kinh tế chưa ổn định và cạnh tranh nhân lực khốc liệt.
* Ứng dựng công nghệ thông tin, đặc biệt là trí tuệ nhân tạo và học máy trong quản trị nguồn nhân lực.
* **Ý nghĩa:**
* Về mặt lý thuyết: Bổ sung và hệ thống hóa kiến thức về ứng dụng học máy trong quản trị nhân sự, đặc biệt là bài toán dự đoán nguy cơ nghỉ việc.
* Về mặt thực tiễn: Đề xuất các mô hình dự đoán có thể hỗ trợ doanh nghiệp quản lý nhân sự hiệu quả, giảm thiểu rủi ro về biến động lao động, đồng thời giúp hoạch định chiến lược nhân sự dài hạn.

## Mục tiêu, đối tượng và phạm vi nghiên cứu

* **Mục tiêu:**
* Ứng dụng các kỹ thuật Machine Learning để đánh giá hiệu suất làm việc của nhân viên dựa trên dữ liệu thu nhập được.
* Hỗ trợ doanh nghiệp trong việc dự báo khả năng nghỉ việc của nhân viên, đánh giá hiệu suất nhân viên từ đó giúp doanh nghiệp có thể can thiệp kịp thời và giữ chân nhân sự phù hợp.
* **Đối tượng nghiên cứu:**
* **Phạm vi nghiên cứu:**
* Tìm hiểu các phương pháp đánh giá hiệu suất làm việc của nhân viên.
* Tìm hiểu về các dữ liệu mẫu.
* Tìm hiểu về việc xử lý dữ liệu, ứng dụng học máy trong xây dựng mô hình dự báo.
* Tìm hiểu và khảo sát để tìm ra thuật toán sử dụng để xây dựng mô hình ra quyết định trong AI.
* Thử nghiệm và triển khai mô hình dự đoán.
* Đánh giá hiệu quả và đề xuất các hướng cải tiến giúp đánh giá và dự đoán tốt hơn.

## Phương pháp nghiên cứu

Để thực hiện đề tài “Ứng dụng Machine Learning trong dự đoán nguy cơ nghỉ việc của nhân viên”, các phương pháp nghiên cứu sau được áp dụng:

* Phương pháp phân tích định lượng: Xử lý dữ liệu bằng Python và các thư viện ML
* Phương pháp mô hình hóa: Ứng dụng các thuật toán Machine Learning để xây dựng mô hình dự đánh giá hiệu suất công việc và dự báo nghỉ việc.
* Phương pháp phân tích và đánh giá mô hình: Đánh giá hiệu quả của mô hình dự đoán thông qua các chỉ số như: Accuraccy, Precision Recall,…
* Phương pháp tổng hợp tài liệu: Tham khảo các bài nghiên cứu, case study, tài liệu liên quan đến HR Analytics, Machine Learning và các dữ liệu mẫu có sẵn để ứng dụng vào xây dựng mô hình

# CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN NGHIÊN CỨU

## Cơ sở lý thuyết

Trong quản trị nguồn nhân lực, vấn đề nghỉ việc của nhân viên luôn là một trong những thách thức lớn mà doanh nghiệp phải đổi mặt. Lực lượng nhân lực có tay nghề cao và kinh nghiệm lâu năm đóng góp rất nhiều vào sự thành công của một tổ chức. Vì vậy, nếu nhân viên rời bỏ hay nghỉ việc không chỉ làm công ty mất đi một nhân tài mà còn dẫn đến nhiều hệ luỵ khác ví dụ sản phẩm họ phát triển hay khách hàng của doanh nghiệp. Điều này ảnh hưởng đến hoạt động cũng như sự phát triển của doanh nghiệp. Tỷ lệ nghỉ việc cao không chỉ làm tăng chi phí tuyển dụng và đào tạo, mà còn ảnh hưởng trực tiếp đến năng suất, chất lượng công việc, môi trường làm việc và sự mất ổn định trong tổ chức, ngoài ra còn có thể ảnh hưởng xấu tới danh tiếng của doanh nghiệp trên thị trường lao động.

Do đó, việc dự đoán nguy cơ nghỉ việc một cách chủ động và kịp thời trở thành một công cụ quan trọng giúp doanh nghiệp có thể tìm ra những đặc điểm chung của nhân sự nghỉ việc để từ xây dựng chiến lược quản lý, cải thiện phục lợi, môi trường làm việc và giữ chân nguồn nhân lực quý giá.

* **Các yếu tố có thể ảnh hưởng đến nguy cơ nghỉ việc của nhân sự:**

Thường xuất phát từ những lý do khách quan từ chính bản thân họ, ngoài ra có thể xuất phát từ những tác động từ môi trường xung quanh, có thể kể đến như:

* thu nhập, chức vụ, môi trường làm việc, tình trạng gia đình hoặc mức độ hài lòng với các đồng nghiệp.
* **Vai trò của việc dự đoán nghỉ việc:**
* Giúp doanh nghiệp chủ động nhận diện các yếu tố ảnh hưởng đến quyết định nghỉ việc của nhân viên.
* Tạo kế hoạch cho việc xây dựng chính sách hỗ trợ nhân sự hợp lý, tập trung vào nhóm nhân viên có nguy cơ cao.
* Tiết kiệm chi phí thông qua việc giảm thiếu tỷ lệ nghỉ việc và tối ưu hoá hoạt động tuyển dụng đào tạo.
* Nâng cao năng lực cạnh tranh nhờ giữ chân được đội ngũ nhân sự chất lượng cao.

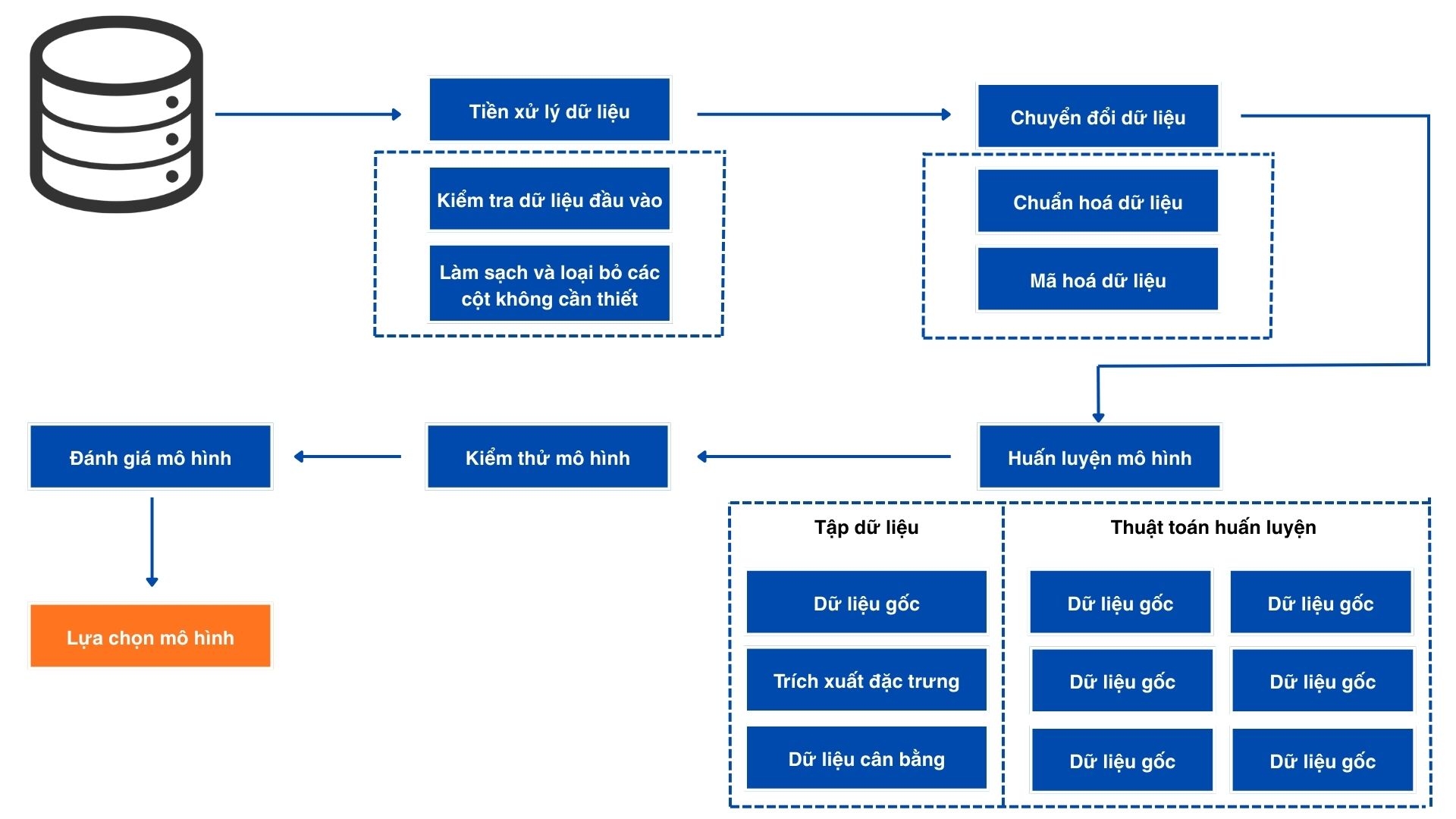
## Tổng quan nghiên cứu

Qua các bài báo cáo cũng như nghiên cứu trong những năm gần đây, ta có thể thấy được việc ứng dụng phân tích dữ liệu vào quản lý nguồn nhân lực dã mang lại giá trị to lớn. Ứng dụng phân tích dữ liệu trong quản lý nguồn nhân lực không chỉ giúp doanh nghiệp nắm bắt xu hướng và hành vi nhân viên, mà còn cung cấp cơ sở khoa học để đưa ra các quyết định quản trị chiến lược, từ đó nâng cao hiệu quả và năng lực cạnh tranh.

Nhiều nghiên cứu đã khai thác các thuật toán học máy như Decision Tree, Random Forest, Support Vector Machine, Logistic Regression,…để dự đoán khả năng nghỉ việc của nhân viên. Kết quả cho thấy, các mô hình này có thể xác định những yếu tố quan trọng như thu nhập, cấp bậc, độ tuổi hay trình độ học vấn, từ đó hỗ trợ doanh nghiệp quản lý và giữ chân nhân tài hiệu quả hơn. Tuy vậy, phần lớn các nghiên cứu mới dừng lại ở việc xác định yếu tố tác động và dự đoán, mà chưa đi sâu phân tích xu hướng biến động nhân sự hay ảnh hưởng dài hạn đối với nhu cầu lao động trong tương lai.

Xuất phát từ khoảng trống đó, nghiên cứu này đề xuất một mô hình dự đoán rủi ro nghỉ việc dựa trên xử lý dữ liệu mất cân bằng, lựa chọn đặc trưng và áp dụng các thuật toán học máy phù hợp. Thông qua đó, doanh nghiệp có thể chủ động xây dựng chiến lược nhân sự, tối ưu nguồn lực và nâng cao hiệu suất tổ chức.

* **Các hướng tiếp cận trong dự đoán nghỉ việc:**
* Phương pháp đánh giá truyền thống: Dựa trên các chỉ số nhân sự như tỷ lệ nghỉ việc, mức thu nhập, mức độ hài lòng, tình trạng hôn nhân, …
* Phương pháp ứng dụng các thuật toán học máy: Ap dụng các thuật toán như Logistic Regression, Decision Tree, Random Forest, XGBoost, Artificial Neural Networks… để mô hình hóa và dự đoán nguy cơ nghỉ việc dựa trên tập hợp các đặc trưng (tuổi, giới tính, thu nhập, thâm niên, mức độ hài lòng, hiệu suất,…).
* **Quy trình nghiên cứu:**



Hình 2.1.

**Bước 1: Thu thập dữ liệu từ trang web Kaggle**

Dữ liệu được thu thập từ nền tảng Kaggle, có 1,470 quan sát với 35 đặc điểm, thể hiện những thông tin liên quan đến sự nghỉ việc của công ty công nghệ IBM.

**Bước 2: Tiền xử lý dữ liệu**

Tiền xử lý dữ liệu là bước…

**Bước 3: Phân loại và chuyển đổi dữ liệu**

Phân loại kiểu dữ liệu thành hai loại đó là. Sử dụng để biến đổi các biến này thành các giá trị 0 và 1 tương ứng…

**Bước 4: Khai phá dữ liệu**

Đây là bước tìm ra các mối quan hệ giữa các biến…

**Bước 5: Huấn luyện mô hình**

Từ dữ liệu gốc lựa chọn ra các thuộc tính quan trọng để tạo ra tập thuộc tính đặc trưng giúp việc huấn luyện mô hình chính xác và tối ưu hơn

**Bước 6: Kiểm thử mô hình**

Các mô hình sẽ cho ra các chỉ số như …. Từ đó tiến hành so sánh hiệu quả dự báo

**Bước 7: Đánh giá và lựa chọn mô hình phù hơp**

Thông qua các so sánh dựa trên các chỉ số mà mô hình đưa ra, chọn mô hình phù hợp nhất

## Giới thiệu về các mô hình dự đoán

### Mô hình hồi quy (Logistic Regression)

* Khái niệm:
* Đây là một thuật toán học máy có giám sát phổ biến và được sử dụng rộng rãi trong các bài như dự đoán khách hàng rời bỏ, phát hiện gian lận hoặc chẩn đoán y học. Mô hình sẽ dự đoán xác suất của một biến phụ thuộc dạng phân loại, kết quả đầu ra của mô hình thường là dạng nhị phân ( Có/Không, Đúng/Sai,…).
* Khác với hồi quy tuyến tính (Linear Regression) dự đoán giá trị liên tục, mô hình hồi quy sử dụng hàm biến đổi đầu ra về khoảng từ 0 đến 1, thể hiện xác suất một quan sát thuộc về một lớp cụ thể.
* Công thức tổng quát

Trong đó:

* : Xác suất biến phụ thuộc Y nhận giá trị 1 khi có tập biến độc lập X
* : Các hệ số hồi quy thể hiện mức độ ảnh hưởng của từng biến độc lập
* Ưu điểm:
* Dễ hiểu, dễ triển khai và giải thích kết quả (mối quan hệ giữa biến độc lập và xác suất).
* Hiệu quả với dữ liệu tuyến tính và không quá phức tạp.
* Tính toán nhanh, không đòi hỏi tài nguyên lớn.
* Có thể biểu diễn xác suất dự đoán (giúp đánh giá mức độ tin cậy).
* Nhược điểm:
* Kém hiệu quả với dữ liệu phi tuyến.
* Dễ bị ảnh hưởng với dữ liệu mất cân bằng hoặc chưa xử lý các giá trị ngoại lai
* Cần xác định và xử lý các biến đặc trưng trước khi huấn luyện

### Cây quyết định (Decision Tree)

* Khái niệm:
* Đây là một thuật toán học máy có giám sát được sử dụng cho cả phân loại và hồi quy. Mô hình hoạt động bằng cách chia dữ liệu thành các nhánh dựa trên giá trị của các đặc trưng, cho đến khi đạt đến một quyết định cuối cùng.
* Cây quyết định giúp mô hình dễ diễn giải vì quá trình quyết định có thể mô tả giống như một chuỗi các câu hỏi phân loại, phù hợi với các bài toán phân loại khách hàng, dự đoán rủi ro.
* Cấu trúc của cây bao gồm:
* Nút gốc (Root Node): Điểm bắt đầu của toàn bộ dữ liệu
* Nút quyết định (Decision Node): Nơi dữ liệu được chia nhỏ dựa trên một đặc điểm cụ thể
* Lá (Leaf Node): Chứa kết quả hoặc nhãn được gán sau khi đã dự đoán.
* Ưu điểm:
* Dễ hiểu, trực quan
* Không cần chuẩn hóa dữ liệu hoặc giả định tuyến tính.
* Có thể xử lý biến định tính và định lượng.
* Có thể phát triển và tối ưu hóa thành các mô hình khác (Random Forest, XGBoots)
* Ưu điểm:
* Dễ bị overfitting nếu cây quá sâu.
* Kết quả có thể không chính xác nếu dữ liệu mất cân bằng.
* Khó tối ưu nếu có nhiều biến liên tục và tương quan.

### Rừng ngẫu nhiên (Random Forest)

* Khái niệm:
* Đây là một thuật toán học máy có giám sát, thuộc nhóm mô hình tập hợp và được tạo ra bằng cách kết hợp nhiều cây quyết định (Decision Tree) để nâng cao độ chính xác và khả năng tổng quát của mô hình.
* Rừng ngẫu nhiên hoạt động dựa trên nguyên tắc mỗi cây trong rừng được huấn luyện trên một mẫu dữ liệu ngẫu nhiên và một tập con ngẫu nhiên của các đặc trưng, từ đó giúp giảm hiện tượng overfitting so với một cây đơn lẻ.
* Công thức tổng quát:

Giả sử có K cây quyết định , , ,.., :

* Đối với phân loại (Classification):

Kết quả cuối cùng là nhãn được dự đoán nhiều nhất – lấy theo “phiếu bầu đa số”.

* Đối với hồi quy (Regression):

Kết quả cuối cùng là trung bình dự đoán của các cây.

* Ưu điểm:
* Độ chính xác cao, giảm overfitting.
* Xử lý tốt dữ liệu có nhiều đặc trưng.
* Có thể đo mức độ quan trọng của từng đặc trưng.
* Nhược điểm:
* Mô hình lớn, phức tạp hơn Decision Tree
* Tốn tài nguyên tính toán hơn.

### K gần nhất (K-Nearest Neighbors)

* Khái niệm:
* KNN (K-Nearest Neighbors) là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học, 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.
* Các bước trong huấn luyện
* 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 mà bạn đã nhận được ở bước 5.
* Ưu điểm:
* Thuật toán đơn giản, dễ dàng triển khai.
* Độ phức tạp tính toán nhỏ.
* Xử lý tốt với tập dữ liệu nhiễu
* Nhước điểm:
* Với K nhỏ dễ gặp nhiễu dẫn tới kết quả đưa ra không chính xác
* Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.
* Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.

## Các chỉ số để đánh giá mô hình

Sau khi mô hình được huấn luyện và dự đoán xong, để xác định được độ chính xác và tin cậy, ta cần kiểm tra và đánh giá các chỉ số mà mô hình đem lại. Các chỉ số đó bao gồm: Accuracy, Precision, Recall, F1-Score và Ma trận nhẫm lẫn.

### Ma trận nhầm lẫn (Confusion Matrix): Ma trận nhầm lẫn là một kỹ thuật giúp đo lường hiệu suất của một phương pháp máy học. Tính toán ma trận nhầm lẫn mang ý nghĩa so sánh kết quả dự đoán phân loại so với kết quả phân loại thực tế, tính toán và so sánh giữa những phân loại đúng và phân loại sai.

### Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:

A diagram of a positive values

AI-generated content may be incorrect.

Hình 2.2.

* TP (True Positive): Số lượng nhân viên nghỉ việc được dự đoán đúng
* TN (True Negative): Số lượng nhân viên không nghỉ việc được dự đoán đúng
* FP (False Positive): Số lượng nhân viên nghỉ việc được dự đoán sai
* FN (False Negative): Số lượng nhân viên không nghỉ việc được dự đoán sai
* Accuracy
* Precision
* Recall
* F1-Score

# TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU

## Tóm tắt tập dữ liệu mẫu

* **Giới thiệu:**

Tập dữ liệu IBM HR Analytics Employee Attrition & Performance được thu thập trên trang web <https://www.kaggle.com/>.

Do các nhà khoa học dữ liệu thuộc tập đoàn IBM tạo ra, nhằm mục đích phân tích và tìm ra các đặc điểm nghỉ việc trong quản lý nhân sự.

Tập dữ liệu này bao gồm 35 cột thuộc tính và 1470 dòng giá trị.

### Các cột thuộc tính

Ảnh có chứa văn bản, thực đơn, ảnh chụp màn hình, tài liệu

Mô tả được tạo tự động

Hình 3.1

* **Bảng mô tả tổng quan dữ liệu**

|  |  |  |  |
| --- | --- | --- | --- |
| **Số thứ tự** | **Cột thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| 1 | Age | Int64 | Độ tuổi của nhân viên |
| 2 | Attrition | Object | Nhân viên đã nghỉ việc hay chưa   * Yes: Có * No: Không |
| 3 | BusinessTravel | Object | Tần suất đi công tác của nhân viên   * Travel\_Rarely: Thỉnh thoảng * Travel\_Frequently: Thường xuyên * Non-Travel: Không đi |
| 4 | DailyRate | Int64 | Tỷ lệ trả lương theo ngày cho nhân viên |
| 5 | Department | Object | Phòng ban làm việc |
| 6 | DistanceFromHome | Int64 | Khoảng cách từ nhà đến công ty |
| 7 | Education | Int64 | Trình độ học vấn   * Thang đo từ 1 đến 5 |
| 8 | EducationField | Object | Chuyên ngành đào tạo |
| 9 | EmployeeCount | Int64 | Số lượng nhân viên |
| 10 | EmployeeNumber | Int64 | Mã nhân viên |
| 11 | EnvironmentSatisfaction | Int64 | Mức độ hài lòng với môi trường làm việc |
| 12 | Gender | Object | Giới tính   * Male: Đàn ông * Female: Phụ nữ |
| 13 | HourlyRate | Int64 | Tỷ lệ trả lương theo giờ cho nhân viên |
| 14 | JobInvolvement | Int64 | Mức độ gắn bó với công việc hiện tại   * Thang đo từ 1 đến 5 |
| 15 | JobLevel | Int64 | Cấp bậc trong công ty   * Thang đo từ 1 đến 5 |
| 16 | JobRole | Object | Chức vụ trong công ty |
| 17 | JobSatisfaction | Int64 | Mức độ hài lòng với công việc   * Thang đo từ 1 đến 4 |
| 18 | MaritalStatus | Object | Tình trạng hôn nhân |
| 19 | MonthlyIncome | Int64 | Thu nhập mỗi tháng |
| 20 | MonthlyRate | Int64 | Tỷ lệ trả lương hàng tháng cho nhân viên |
| 21 | NumCompaniesWorked | Int64 | Số lượng công ty đã làm trước đó |
| 22 | Over18 | Object | Có trên 18 tuổi hay chưa |
| 23 | OverTime | Object | Làm việc tăng ca   * Yes: Có * No: Không |
| 24 | PercentSalaryHike | Int64 | Phần trăm tăng lương |
| 25 | PerformanceRating | Int64 | Đánh giá hiệu suất làm việc   * Thang đo từ 1 đến 4 |
| 26 | RelationshipSatisfaction | Int64 | Mức độ hài lòng về mối quan hệ với đồng nghiệp   * Thang đo từ 1 đến 4 |
| 27 | StandardHours | Int64 | Số giờ làm chuẩn |
| 28 | StockOptionLevel | Int64 | Mức quyền chọn cổ phiếu   * Thang đo từ 1 đến 3 |
| 29 | TotalWorkingYears | Int64 | Tổng số năm kinh nghiệp làm việc |
| 30 | TrainingTimesLastYear | Int64 | Số lần tham gia đào tạo trong năm vừa qua |
| 31 | WorkLifeBalance | Int64 | Cân bằng giữa công việc và cuộc sống |
| 32 | YearsAtCompany | Int64 | Số năm làm việc tại công ty hiện tại |
| 33 | YearsInCurrentRole | Int64 | Số năm trong chức vụ hiện tại |
| 34 | YearsSinceLastPromotion | Int64 | Số năm kể từ lần thăng chức gần nhất |
| 35 | YearsWithCurrManager | Int64 | Số năm làm việc với quản lý hiện tại |

Bảng 3.1

### Thống kê số liệu trong tập dữ liệu

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, tài liệu

Mô tả được tạo tự động

Hình 3.2

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 3.3

* **Phân loại các biến trong tập dữ liệu:**

Tập dữ liệu trên có hai loại biến đó là biến phân loại và biến số.

* Biến phân loại mang ý nghĩa gán nhẵn, nhóm, mô tả đặc điểm và thuộc tính của nhân viên ví dụ như giới tính, tên, chức vụ, địa chỉ.
* Đối với biến số, đây là các biến đô lường thể hiện giá trị cụ thể về thu nhập, thời gian, kinh nghiệm. Dưới đây là phân loại các biến:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 3.4.

## Các bước làm sạch và tiền xử lý dữ liệu

### Kiểm tra và xử lý dữ liệu thiếu

Dữ liệu thiếu có thể dẫn đến kết quả thống kê mà vô mình dự đoán bị sai lệch, ngoài ra nhiều thuật toán học máy không thể xử lý được nếu dữ liệu thị thiếu. Cách khắc phục là xoá những dòng chứa giá trị thiếu, thay thế bằng giá trị trung bình/ trung vị, hoặc dùng mô hình ước lượng.

Ảnh có chứa văn bản, thực đơn, bức thư, tài liệu

Mô tả được tạo tự động

Hình 3.5

Kết quả kiểm tra cho thấy không có dữ liệu thiếu trong tập dữ liệu

### Xử lý dữ liệu dư thừa, trùng lặp

Dữ liệu dư thừa và trùng lặp có thể làm sai thống kê, gây nhiễu và tăng độ phức tạp không cần thiết. Ngoài ra còn dẫn tới mô hình xử lý chậm hơn, dễ overfitting. Xác định các cột chứa các giá trị cố định trong tập dữ liệu, loại bỏ các cột vô nghĩa và không quan trọng tới mô hình

Ảnh có chứa văn bản, Phông chữ, tài liệu, bức thư

Mô tả được tạo tự động

Hình 3.6

Qua quan sát dữ liệu thì có thể thấy các cột dữ liệu không cần thiết như sau:

* **Over18**: Tất cả nhân viên đều đã trên 18 tuổi
* **EmployeeCount**: Luôn có giá trị là 1
* **EmployeeNumber**: Luôn có giá trị khác nhau như việc nhân viên có mã nhân viên
* **StandardHours**: Luôn có giá trị là 80

Tiến hành loại bỏ các cột dư thừa, kết quả sau khi đã loại bỏ:

Ảnh có chứa văn bản, thực đơn, ảnh chụp màn hình, tài liệu

Mô tả được tạo tự động

Hình 3.7

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Hình 3.8

* **Chuẩn hoá dữ liệu**

Các biến có đơn vị khác nhau như tuổi và thu nhập. Nếu không chuẩn hoá thì sẽ có sự chêch lệch rõ ràng từ biến có giá trị lớn so với biến nhỏ.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, đen và trắng

Mô tả được tạo tự động

Hình 3.9

### Mã hóa biến phân loại

Các biến phân loại thể hiện đặc điểm của nhân viên, vì vậy để thuật toán có thể xử lý thì cần mã hoá các biến này thành dạng số. Mã hoá Attrition thành 0/1

….

* **Xử lý giá trị ngoại lai**

Trong quá trình thu thập hoặc nhập liệu có thể xảy ra lỗi và có thể xuất hiện các giá trị ngoại lai. Giá trị ngoại lai là giá trị nhỏ hoặc lớn bất thường so với các giá trị còn lại, nó làm ảnh hưởng tới trung bình, phương sai và mô hình khiến cho mô hình dự đoán sai và kém chính xác. Dùng phương pháp IQR\* phát hiện các giá trị ngoại lai và tạo ra bộ dữ liệu loại bỏ giá trị ngoại lai

Ảnh có chứa văn bản, thực đơn, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Hình 3.10

Với ngưỡng giá trị tại phân vị là 25% và 75%, sau đó đánh dấu các hàng có ngoại lệ. Kết quả cho thấy có 691 hàng có ngoại lệ được phát hiện so với 1470 mẫu ban đầu. Sau khi loại bỏ các ngoại lệ ta thu được mẫu mới chứa 779 giá trị.

Ảnh có chứa văn bản, Phông chữ, biên lai, màu trắng

Mô tả được tạo tự động

Hình 3.11

* **Xử lý mất cân bằng dữ liệu**

Khi phân tích và dự đoán một bài toán nào đó thường có rất nhiều vấn đề, một trong số đó là việc dữ liệu bị mất cân bằng hoặc nghiêng về một hướng. Trong dự đoán nghỉ việc, thông thường ở thuộc tính Attrition tình trạng nghỉ việc (Yes) chiếm rất ít so với không nghỉ (No). Điều này dẫn đến mô hình không xử lý chính xác và luôn dự đoán No mà chỉ số accuracy vẫn cao và có thể mô hình sẽ không phát hiện được các trường hợp hiếm. Phương pháp được sử dụng ở đây là SMOTE.

Ảnh có chứa văn bản, thực đơn, tài liệu, ảnh chụp màn hình

Mô tả được tạo tự động

Hình 3.12

## Khai phá dữ liệu

### Tổng quan về tình hình nghỉ việc của nhân viên

Thông kế số lượng nhân viên nghỉ việc và không nghỉ việc:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Hình 3.13

Kết quả cho thấy tỷ lệ nghỉ việc của nhân viên chiếm 16,12%

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình, biểu đồ

Mô tả được tạo tự động

Hình 3.14

### Trực quan hoá phân bố dữ liệu theo từng biến.

* **Về thông tin cá nhân của nhân viên**

Ảnh có chứa hàng, Sơ đồ, biểu đồ, văn bản

Mô tả được tạo tự động

Hình 3.15

* Nhân viên có độ tuổi phân bố tập trung ở 25 tới 45 tuổi, trong đó nhân viên ở độ tuổi 30 tới 35 chiếm nhiều nhất.
* Tỷ lệ giới tính giữa nam và nữ phân bố khá đồng đều
* Tình trạng hôn nhân của nhân viên đa số là đã kết hôn, kế tiếp là tình trạng độc thân. Ngoài ra còn có những trường hợp nhân viên đã ly hôn.
* **Về quản trị nhân sự:**

Ảnh có chứa hàng, văn bản, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Hình 3.16

* Như đã nói ở trên, tỷ lệ nghỉ việc của nhân viên rất mất cân bằng. Phần lớn nhân viên không nghỉ việc, chỉ một số nhỏ nghỉ việc
* Nhân viên chọn không làm thêm giờ chiếm số lượng lớn, nhưng một tỷ lệ đáng kể nhân viên có làm thêm
* **Về công việc và phòng ban**

Ảnh có chứa văn bản, biểu đồ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động

Hình 3.17

* Phần lớn nhân viên hiếm khi phải đi công tác, chỉ có số ít nhân viên phải đi công tác thường xuyên và hầu như không có trường hợp không đi công tác.
* Chủ yếu nhân viên thuộc phòng ban Nghiên cứu và Phát triển, ngoài ra có số lượng ít thuộc phòng Bán hàng và rất ít nhân viên ở phòng Quản trị nhân sự.
* Chức vụ phân bổ không đồng đều, một số chức vụ chiếm áp đảo
* Cấp bậc tập trung ở level 1-2, số lượng level cao hơn không đáng kể
* **Thu nhập và lương**

Ảnh có chứa biểu đồ, hàng, Sơ đồ, văn bản

Mô tả được tạo tự động

Hình 3.18

* Tỷ lệ trả lương theo ngày, giờ và tháng phân bố khá đồng đều.
* Thu nhập hàng tháng bị lệch về phải, đa số dưới 10000, một số ít nhân viên có thu nhập cao trên 15000.
* Phần trăm tăng lương đa số năm trong khoảng 11 tới 15%
* **Trình độ học vấn và đào tạo**

Ảnh có chứa văn bản, hàng, Sơ đồ, ảnh chụp màn hình

Mô tả được tạo tự động

Hình 3.19

* Trình độ học vấn tập trung ở bậc 3
* Ngành đào tạo phần lớn là
* Đa số nhân viên đều tham gia đào tạo từ 2 tới 3 lần trong năm vừa qua, một số nhỏ nhân viên được đào tạo trên 5 lần
* **Sự hài lòng và đánh giá**

Ảnh có chứa văn bản, biểu đồ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động

Hình 3.20

* Tất cả đều khá đồng đều, nhưng vẫn có xu hướng tập trung ở mức cao 3 và 4, cho thấy nhân viên khá hài lòng với công việc, môi trường, đồng nghiệp cũng như có thể cân bằng giữa cuộc sống và công việc.
* Đánh giá hiệu suất nhân viên gần như toàn bộ đều là 3, một ít là 4
* **Về kinh nghiệm và thâm niên làm việc**

Ảnh có chứa biểu đồ, hàng, văn bản, Sơ đồ

Mô tả được tạo tự động

Hình 3.21

* Nhân viên đa số sống gần công ty từ 1 tới 5km, một số ít phải di chuyển xa tới 20km để tới nơi làm việc.
* Nhiều người từng làm ở 1 đến 2 công ty, chỉ có một số ít nhân viên có kinh nghiệm làm việc ở nhiều nơi
* Nhân viên đa số có thời gian dưới 20 năm làm việc, mức đỉnh ở khoảng 5 tới 10 năm
* Kinh nghiệm với vai trò hiện tại tập trung từ 0 tới 5 năm
* Nhân viên chỉ mới làm việc với người quản lý của họ từ 0 tới 5 năm

**Kết luận**

Ảnh có chứa văn bản, biểu đồ, Kế hoạch, sơ đồ

Mô tả được tạo tự động

Hình 3.22

Qua trực quan hoá biểu đồ Histogram, ta có thể thấy được kết quả khách quan như sau:

* Dữ liệu mất cân bằng ở Attrition, cần xử lý.
* Thu nhập & kinh nghiệm đều có phân bố lệch phải cho thấy nhiều nhân viên mới, lương thấp, ít người lâu năm, lương cao.
* Nhiều biến dạng thang đo (satisfaction, involvement) phân bố khá đồng đều nhưng hơi nghiêng về mức hài lòng cao.

### Trực quan hoá mối quan hệ giữa các biến và tình trạng nghỉ việc

Ảnh có chứa Kế hoạch, biểu đồ

Mô tả được tạo tự động

Hình 3.23

* **Về thông tin cá nhân của nhân viên**
* Tình trạng nhân viên trẻ độ tuổi từ 20 tới 30 tuổi có tỷ lệ nghỉ việc cao hơn rõ rệt
* Nhân viên độc thân có tỷ lệ nghỉ việc cao hơn những nhân viên đã kết hôn hoặc đã ly hôn
* Nhân viên ở xa có xu hướng nghỉ việc nhiều hơn

Ý nghĩa: Nhìn chung, nhân viên trẻ, độc thân, sống xa công ty có rủi ro nghỉ việc cao

* **Về trạng thái nhân sự**
* Tình trạng nghỉ việc chiếm khoảng 16,12% tổng số. Nhưng tăng ca là yếu tố nổi bật nhất, Nhân viên phải làm thêm giờ thường xuyên có tỷ lệ nghỉ việc cao hơn rất nhiều.

Ý nghĩa: Quản lý khối lượng công việc chưa tốt sẽ dẫn đến tình trạng nghỉ việc ở nhân viên.

* **Công việc và phòng ban**
* Những nhân viên thuộc bộ phận Bán hàng có tỷ lệ nghỉ việc cao hơn số với các phòng ban khác.
* Tần suất đi công tác nhiều có tỷ lệ nghỉ việc cao.
* Cấp bậc và chức vụ cũng ảnh hưởng tới tỷ lệ nghỉ việc. Đa số các cấp bậc thấp từ 1 đến 2 có khả năng nghỉ việc cao. Về chức vụ

Ý nghĩa: Những nhân viên cấp bậc thấp, bộ phận Bán hàng và phải thường xuyên công tác có rủi ro nghỉ việc cao

* **Trình độ học vấn và đào tạo**
* Không cho thấy khác biệt quá rõ, nhưng nhân viên từ ngành HR/Technical có tỷ lệ nghỉ việc cao hơn
* Số lần đào tạo không phải yếu tố chính bởi vì dù có trải qua nhiều lần đào tạo vẫn có trường hợp nghỉ việc

Ý nghĩa: Chuyên ngành có thể ảnh hưởng tới tỷ lệ nghỉ việc.

* **Mức độ hài lòng và đánh giá**
* Những trường hợp không hài lòng với công việc, môi trường và đồng nghiệp cũng như không thể cân bằng được giữa cuộc sống và công việc đều dẫn tới nghỉ việc
* Hiệu suất làm việc của nhân viên hầu hết đều cao, không phân biệt được rõ.

Ý nghĩa: Có thể thấy được mức độ hài lòng và cân bằng cuộc sống là yếu tố ảnh hưởng lớn.

* **Kinh nghiệm và thâm niên làm việc**
* Những nhân viên có ít kinh nghiệm, thường dưới 10 năm có khả năng dễ nghỉ việc hơn
* Số năm ở công ty và thời gian đảm nhận vai trò hiện tại từ 0 tới 3 đầu tỷ lệ nghỉ việc cao hơn.
* Vấn đề thăng chức lâu cũng dẫn tới tình trạng nghỉ việc
* Mối quan hệ ít gắn bó với quản lý sẽ dễ nghỉ việc hơn

Ý nghĩa: Nhân viên mới vào, ít kinh nghiệm, chưa được thăng chức, hoặc quan hệ với quản lý chưa bền vững sẽ dễ nghỉ việc

* **Thu nhập và lương**
* Nhân viên có thu nhập thấp dưới 5000 sẽ nghỉ việc nhiều
* Tỷ lệ trả lương theo giờ, ngày và tháng phân bố khá ngẫu nhiên, không ảnh hưởng nhiều
* Tỷ lệ tăng lương không khác biệt rõ ràng
* Nhân viên có mức chọn cổ phần cao có xu hướng gắn bó với công ty hơn

Ý nghĩa: Lương thấp và không có phúc lợi cổ phiếu dễ khiến nhân viên nghỉ.

**Kết luận**

Qua trực quan hoá các biểu đồ và nhìn nhận, ta có được những đánh giá các yếu tố dẫn tới tình trạng nghỉ việc như sau:

* Nhân viên trẻ tuổi, độc thân và ở xa nới làm việc
* Khối lượng công việc phải tăng ca và đi công tác thường xuyên là yếu tố lớn
* Vai trò và cấp bậc thấp dễ nghỉ việc hơn
* Nhân viên thường rời bỏ công ty vì lương thấp và nhận được ít phúc lợi
* Sự hài lòng thấp và không thể cân bằng cuộc sống là yếu tố quan trọng
* Nhân viên có thâm niên ngắn và ít cơ hội thăng tiến.

Ảnh có chứa văn bản, biểu đồ, Kế hoạch, ảnh chụp màn hình

Mô tả được tạo tự động

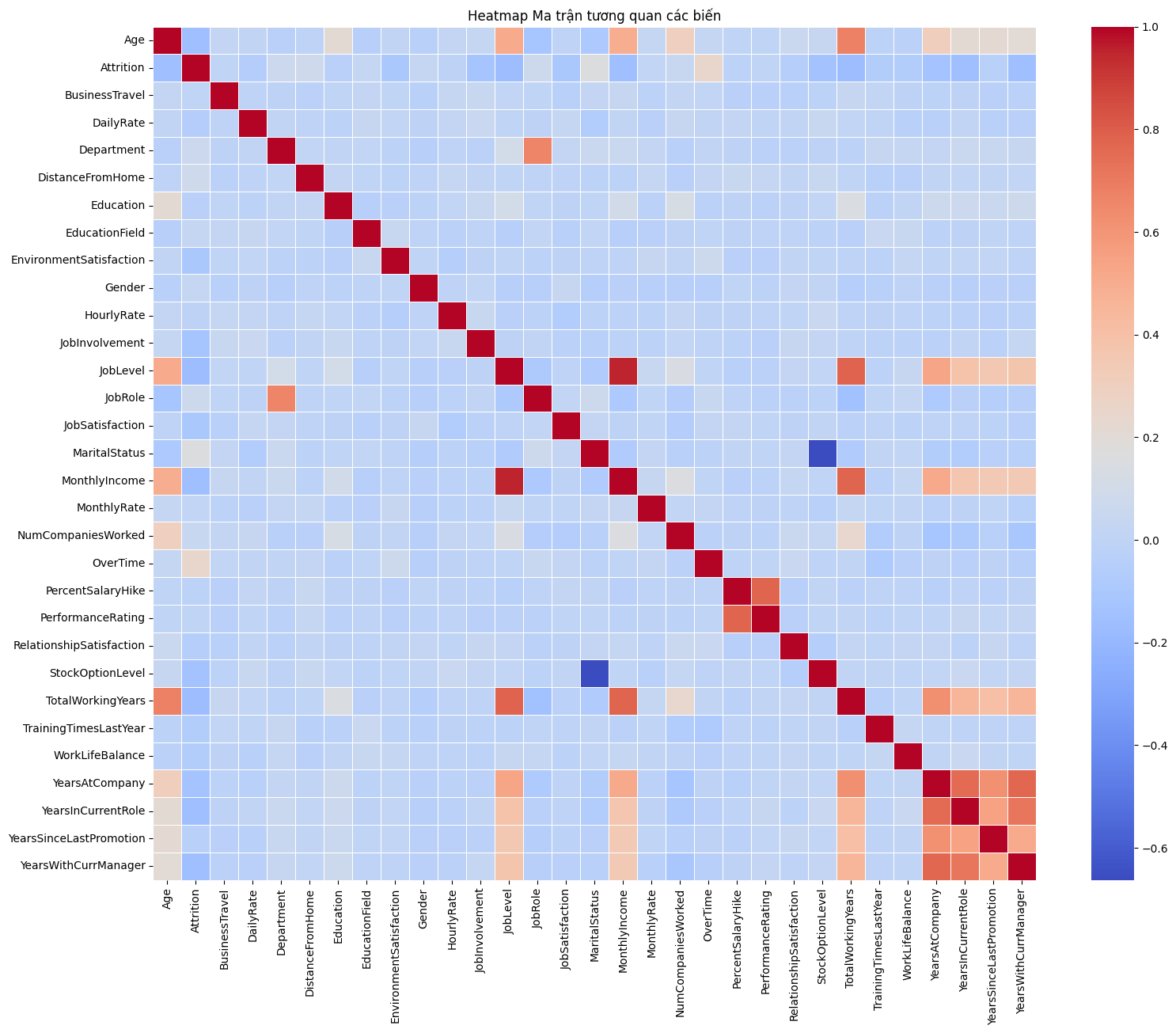
Hình 3.24

Ảnh có chứa ảnh chụp màn hình, biểu đồ

Mô tả được tạo tự động

Hình 3.25

### Trực quan hoá ma trận tương quan giữa các biến



Hình 3.26

* Màu đỏ đậm: tương quan dương mạnh (2 biến tăng cùng nhau).
* Màu xanh đậm: tương quan âm mạnh (một biến tăng, biến kia giảm).
* Màu xanh nhạt/đỏ nhạt: ít hoặc không có tương quan.
* Đường chéo đỏ = 1.0 vì một biến luôn tương quan tuyệt đối với chính nó.

### Phân tích tương quan giữa các biến trong bộ dữ liệu

* Nhóm biến thâm niên/kinh nghiệm (YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion, TotalWorkingYears):

Có tương quan cao với nhau, cho thấy cùng đo lường gắn bó/thời gian.

* Thu nhập và cấp bậc (MonthlyIncome, JobLevel, TotalWorkingYears):

Càng thâm niên, càng cấp bậc cao thì lương càng cao.

* Attrition (nghỉ việc):

Không có biến nào tương quan mạnh tuyệt đối, nhưng có quan hệ yếu – trung bình:

* Dương với OverTime (tăng ca nhiều dẫn tới dễ nghỉ).
* Âm với MonthlyIncome, TotalWorkingYears (ít kinh nghiệm, lương thấp dẫn tới nghỉ việc cao).
* Âm nhẹ với StockOptionLevel (có cổ phiếu sẽ gắn bó lâu dài).
* Biến ít giá trị (DailyRate, HourlyRate, MonthlyRate)

Gần như độc lập với tất cả biến khác, có thể xem xét loại bỏ.

### Phân tích đơn biến, đa biến

* Đơn biến
* Attrition: mất cân bằng (đa số “Stay”, ít “Leave”).
* Age: tập trung 25–40 tuổi, ít người >55.
* MonthlyIncome: phân bố lệch phải, nhiều người thu nhập thấp, ít người thu nhập rất cao.
* JobRole: có một số nhóm nhiều mẫu (Sales Executive, Research Scientist) và một số ít (Manager, HR).
* Đa biến
* Attrition - OverTime: nhóm làm OT cao có tỷ lệ nghỉ việc nhiều hơn.
* Attrition - Income: nhân viên lương thấp có tỷ lệ nghỉ cao hơn.
* Attrition - JobRole: một số nhóm nghề (Sales, Lab Tech) có tỷ lệ nghỉ việc cao hơn nhóm Manager/Research Director.
* Age - Attrition: nhân viên trẻ (từ 25 đến 35 tuổi) dễ nghỉ hơn so với nhóm trên 40 tuổi.

### Phân tích như cầu tăng cường mẫu của tập dữ liệu

* **Phân tích và tìm ra các đặc trưng quan trọng trong tập dữ liệu**

Tập dữ liệu ban đầu có 35 thuộc tính, tuy nhiên trong đó có thể có nhiều thuộc tính không liên quan hoặc không ảnh hưởng tới việc dự đoán. Điều này có thể khiến mô hình bị nhiễu và overfitting, ảnh hưởng tới hiệu suất mô hình. Chính vì vậy, cần phân tích và tìm ra các đặc trưng quan trọng trong tập dữ liệu giúp tăng hiệu suất, giảm thời gian huấn luyện và tránh đa cộng tuyến.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Hình 3.27

Các bước thực hiện phân tích:

* Chia dữ liệu thành tập đặc trưng đầu vào và tập nhãn mục tiêu
* Áp dụng mô hình huấn luyện XGBoost và tạo SHAP(\* chưa biết giải thích thuật ngữ này ở phần nào) explainer
* Tính giá trị SHAP cho từng đặc trưng
* Trực quan hoá các đặc trưng quan trọng
* Lấy giá trị tuyệt đối độ quan trọng
* Tạo bảng xếp hạng đặc trưng
* Lưu lại 12 đặc trung quan trọng nhất
* Lọc dữ liệu đã loại bỏ giá trị ngoại lai

Bảng xếp loại các đặc trưng quan trọng

|  |  |
| --- | --- |
| Feature | Importance |
|  |  |

Ảnh có chứa văn bản, ảnh chụp màn hình, Nhiều màu sắc

Mô tả được tạo tự động

Hình 3.28

# XÂY DỰNG MÔ HÌNH DỰ ĐOÁN

## Phân chia dữ liệu huấn luyện và kiểm tra

### Dữ liệu gốc

* Chia bộ dữ liệu ra làm 2 phần 70/30 để huấn luyện mô hình
* Số lượng mẫu trong tập huấn luyện là 1029 mẫu, trong đó số lượng mẫu theo lớp có:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 863 | 166 |

Bảng 4.1

* Số lượng mẫu trong tập kiểm tra là 441 mẫu, trong đó:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 371 | 71 |

Bảng 4.2

* Chia bộ dữ liệu đã loại bỏ giá trị ngoại lai và cân bằng dữ liệu ra làm 2 phần 70/30
* Số lượng mẫu trong tập huấn luyện là 545 mẫu, trong đó:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 448 | 97 |

* Số lượng mẫu trong tập kiểm tra là 234 mẫu, trong đó:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 193 | 41 |

* Chia bộ dữ liệu bộ dữ liệu đã lọc ra 12 đặc trưng quan trọng nhất
* Số lượng mẫu trong tập huấn luyện là 1029 mẫu, trong đó:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 863 | 166 |

* Số lượng mẫu trong tập kiểm tra là 441 mẫu, trong đó:

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 370 | 71 |

### Xử lý mất cân bằng dữ liệu bằng phương pháp SMOTE

Nhận thấy bộ dữ liệu đang bị mất cân bằng giữa các lớp. Tiến hành áp dụng phương pháp SMOTE thêm số lượng mẫu trong tập huấn luyện

* Bộ dữ liệu gốc

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 863 | 863 |

* Bộ dữ liệu đã loại bỏ giá trị ngoại lai

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 448 | 448 |

* Bộ dữ liệu khi đã lọc ra 12 đặc trưng quan trọng nhất

|  |  |
| --- | --- |
| Lớp 0 | Lớp 1 |
| 863 | 863 |

## Huấn luyện mô hình

### Huấn luyện mô hình Logistic Regression và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động

* Bộ dữ liệu đã qua

### Huấn luyện mô hình Decision Tree và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ

Mô tả được tạo tự động

### Huấn luyện mô hình Random Forest và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

**Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động**

### Huấn luyện mô hình KNN và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

**Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động**

### Huấn luyện mô hình XGBoost và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

**Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động**

### Huấn luyện mô hình LightGBM và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

### Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật Mô tả được tạo tự động

### Huấn luyện mô hình SVM và đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

### Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật Mô tả được tạo tự động

### Huấn luyện mô hình Tabnetvà đánh giá

* Bộ dữ liệu gốc:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 |  |  |  |  |
| 1 |  |  |  |  |
| Accuracy |  |  |  |  |
| Macro Avg |  |  |  |  |
| Weighted Avd |  |  |  |  |

**Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Mô tả được tạo tự động**

# THỬ NGHIỆM VÀ ĐÁNH GIÁ

## Kiểm thử mô hình

## Kết quả đạt được

## Hạn chế

# TÀI LIỆU THAM KHẢO

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết luận

## Hướng phát triển