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.

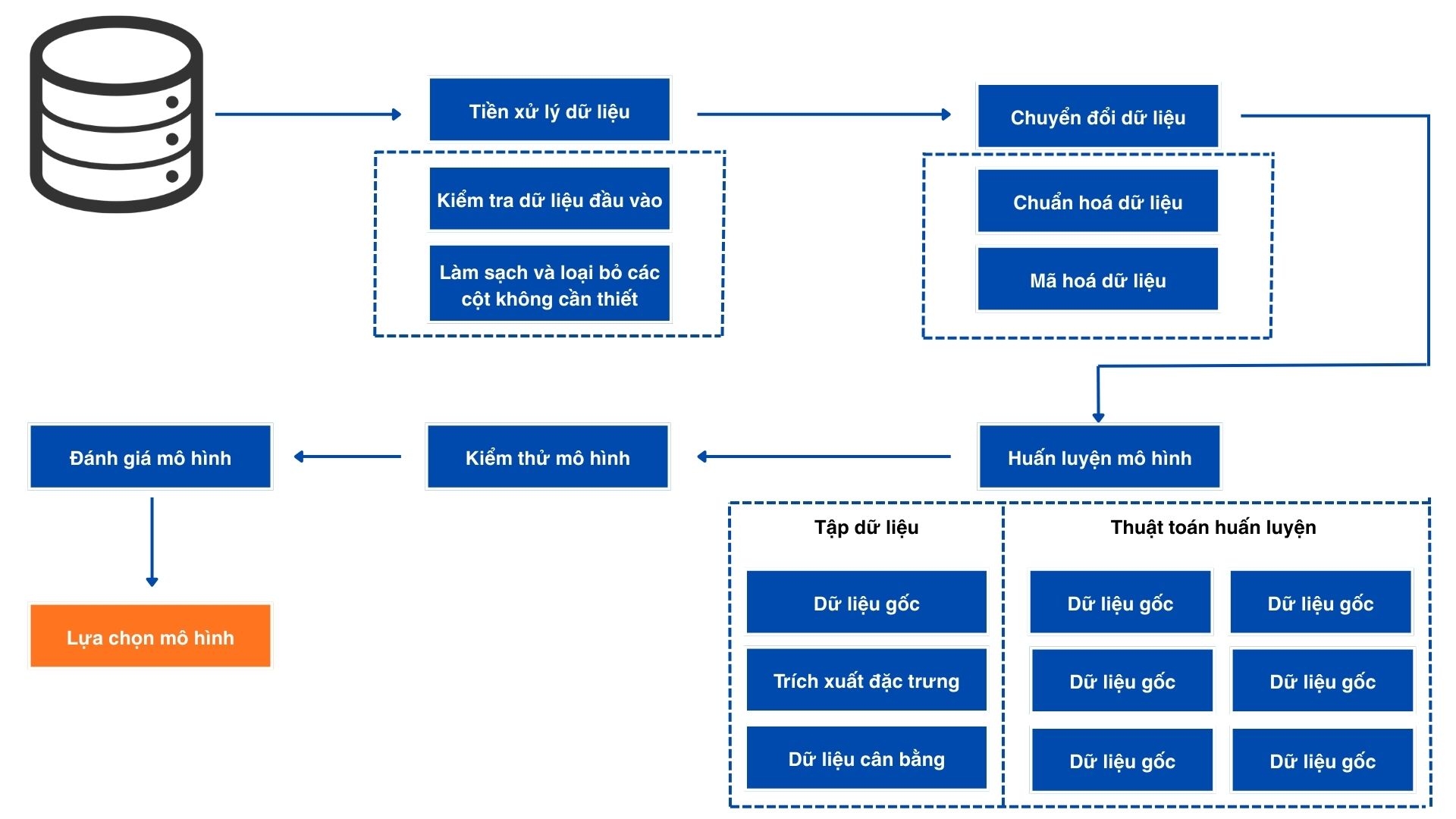
## Tổng quan nghiên cứu

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: áp dụng các thuật toán như Logistic Regression, Decision Tree, Random Forest, XGBoost,… 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 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

## Các phương pháp đánh giá và ứng dụng các thuật toán học máy

### Phương pháp đánh giá truyền thống

#### KPI:

* **Khái niệm:** là một công cụ đánh giá hiệu quả thực hiện công việc của một cá nhân, bộ phận hoặc một tổ chức. Chỉ số KPI được thể hiện qua nhiều tiêu chí khác nhau liên quan đến số liệu, tỷ lệ hay chỉ tiêu định lượng… phụ thuộc vào đặc tính nghiệp vụ, chuyên môn của từng đối tượng.
* **Ý nghĩa và vai trò:** Đánh giá KPI giúp quản lý dễ dàng theo dõi được hiệu suất làm việc của mỗi nhân viên một cách trực quan hơn. Nhờ đó, có thể đưa ra những nhận định minh bạch, chính xác về kế hoạch làm việc và chế độ lương thưởng hay kỷ luật.
* **Mô hình thực hiện:**

#### OKR

#### Đánh giá 360 độ

#### Bars

### Ứng dụng thuật toán học máy

#### Phân cụm

#### Khai phá dữ liệu và luật kết hợp

#### Máy học

#### Xử lý ngôn ngữ lớn

### Giới thiệu về các mô hình dự đoán

# 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 2

* **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 |  | 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

### 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

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

Mô tả được tạo tự động

Hình

* **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. 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 .

## 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

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

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

* **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, giấy, bức thư, tài liệu

Mô tả được tạo tự động

### 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.

* **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.

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

Mô tả được tạo tự động

Hình

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

Mô tả được tạo tự động

Hình

* **Xử lý mất cân bằng dữ liệu**

## Khai phá dữ liệu

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

Mô tả được tạo tự động

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

Mô tả được tạo tự động

Ảnh có chứa Kế hoạch, biểu đồ

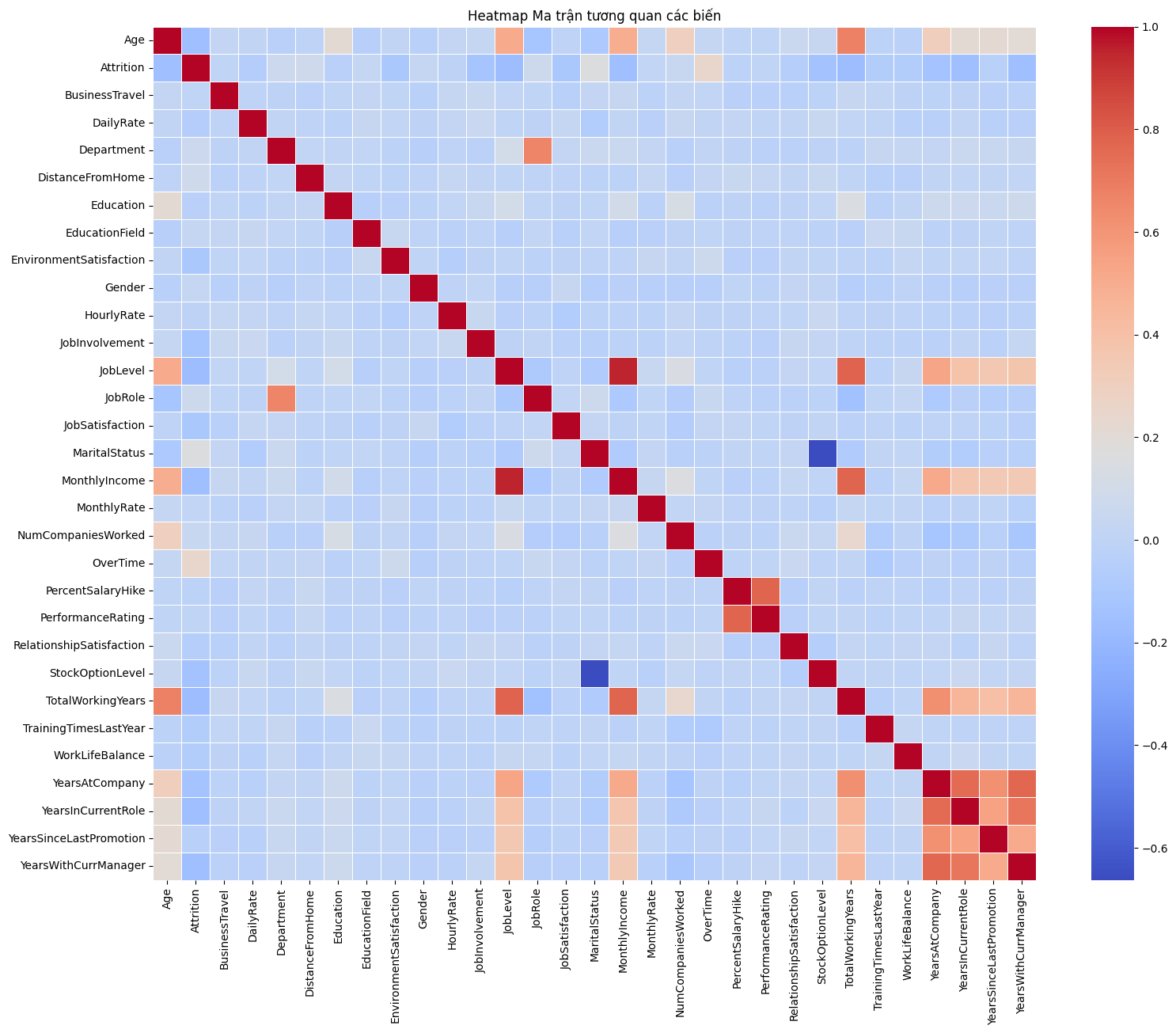
Mô tả được tạo tự động

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

Mô tả được tạo tự động

Ảnh có chứa ảnh chụp màn hình, biểu đồ

Mô tả được tạo tự động



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

### Phân tích đơn biến, đa biến

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

# XÂY DỰNG MÔ HÌNH DỰ ĐOÁN

## Phân chia dữ liệu huấn luyện và kiểm tra

## Huấn luyện mô hình

# 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