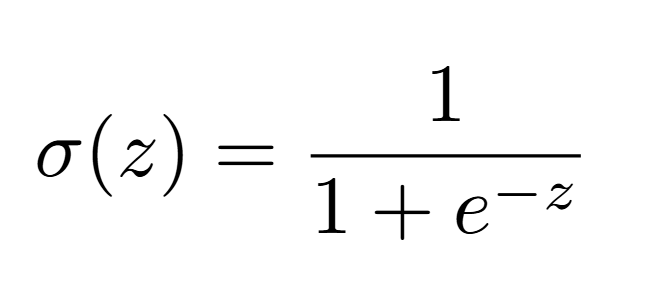
**Dự đoán nghỉ việc của nhân viên bằng Logistic Regression**

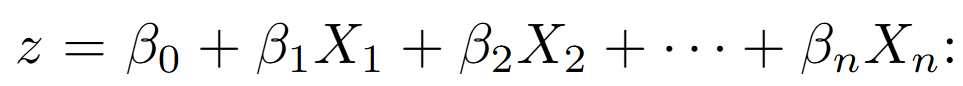
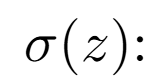
# **1. Mục Tiêu**

Dự án này nhằm xây dựng một mô hình **Hồi quy Logistic** để dự đoán khả năng nghỉ việc của nhân viên dựa trên các đặc điểm nhân khẩu học, tài chính và công việc. Hồi quy Logistic là một mô hình phân loại nhị phân, sử dụng hàm sigmoid để ước tính xác suất một sự kiện xảy ra. Trong bài toán này, mô hình giúp xác định khả năng một nhân viên có thể nghỉ việc hay không.

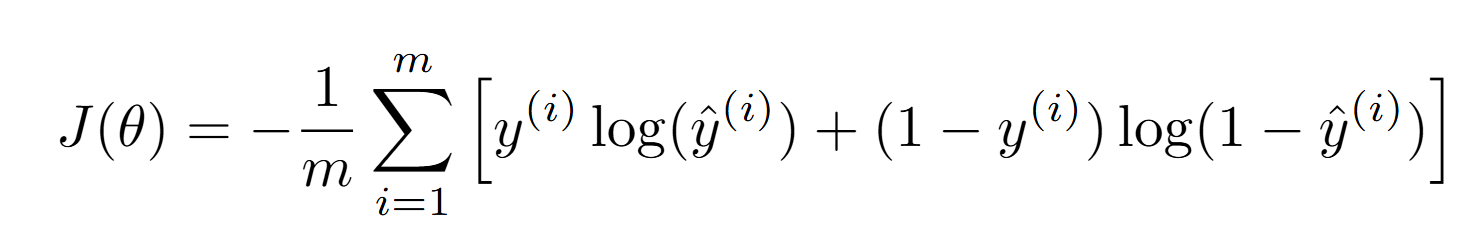
Hàm sigmoid được sử dụng trong Hồi quy Logistic có dạng:



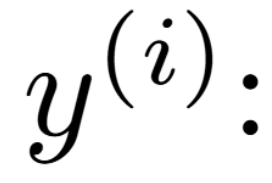
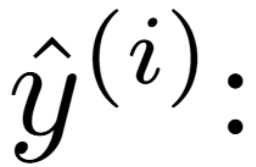
Trong đó:

* là tổng trọng số của các biến đầu vào.
* là xác suất của một nhân viên nghỉ việc.

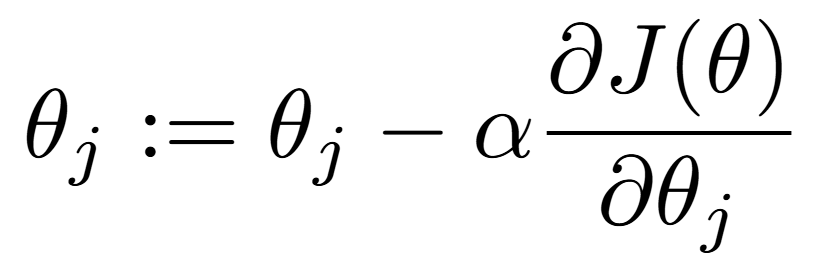
Hàm mất mát (Binary Cross-Entropy) được sử dụng để tối ưu mô hình:



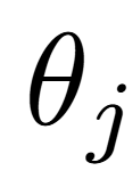
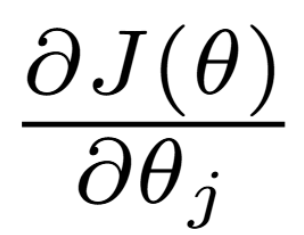
Trong đó:

* Giá trị thực tế của mẫu thứ i.
* Giá trị dự đoán của mẫu thứ i.
* m: Số lượng mẫu trong tập dữ liệu.

Quá trình tối ưu hóa sử dụng thuật toán Gradient Descent:



Trong đó:

*  : Tham số cần tối ưu.
* : Tốc độ học (learning rate).
* : Đạo hàm của hàm mất mát J(θ) theo tham số θj .

Lý do chọn Hồi quy Logistic cho bài toán này:

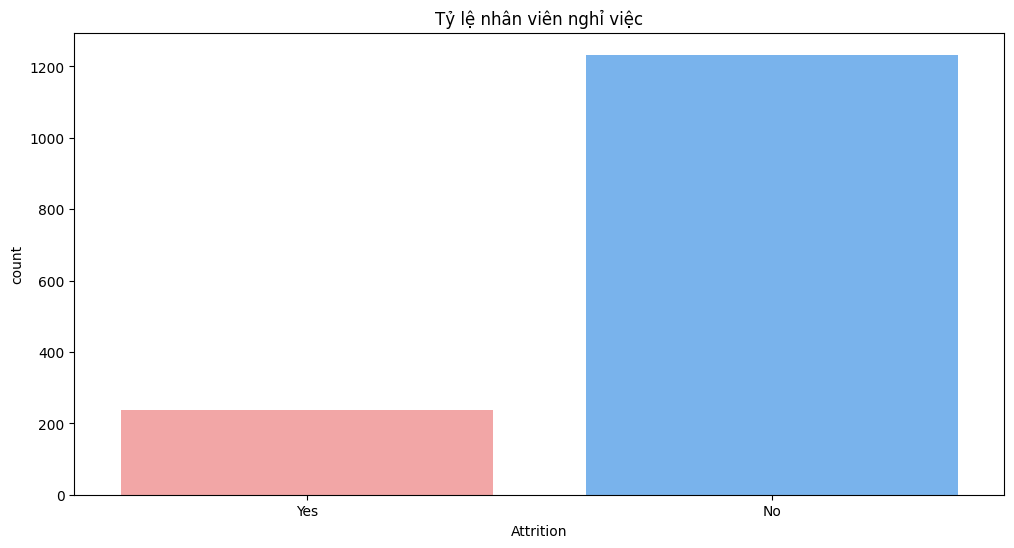
1. **Khả năng diễn giải tốt**: Hệ số của mô hình có thể giúp xác định mức độ ảnh hưởng của từng yếu tố đến xác suất nghỉ việc.
2. **Tính toán nhanh chóng**: Phù hợp với tập dữ liệu vừa phải mà không đòi hỏi tài nguyên tính toán cao.
3. **Tổng quát hóa tốt**: Tránh được overfitting nếu dữ liệu không quá phức tạp.
4. **Khả năng mở rộng**: Có thể áp dụng thêm L1/L2 regularization để cải thiện hiệu suất.

# **2. Data/Input/Output**

## **2.1 Data**

* **Dữ liệu sử dụng:** Tập dữ liệu [*IBM HR Analytics Employee Attrition & Performance*](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset) từ Kaggle.
* **Kích thước dữ liệu:** 1470 mẫu, 35 cột.

### **2.1.1 Sự phân bố của “Nhân viên nghỉ”**



Dữ liệu cho thấy sự mất cân bằng đáng kể giữa hai nhóm nhân viên nghỉ việc và không nghỉ việc:

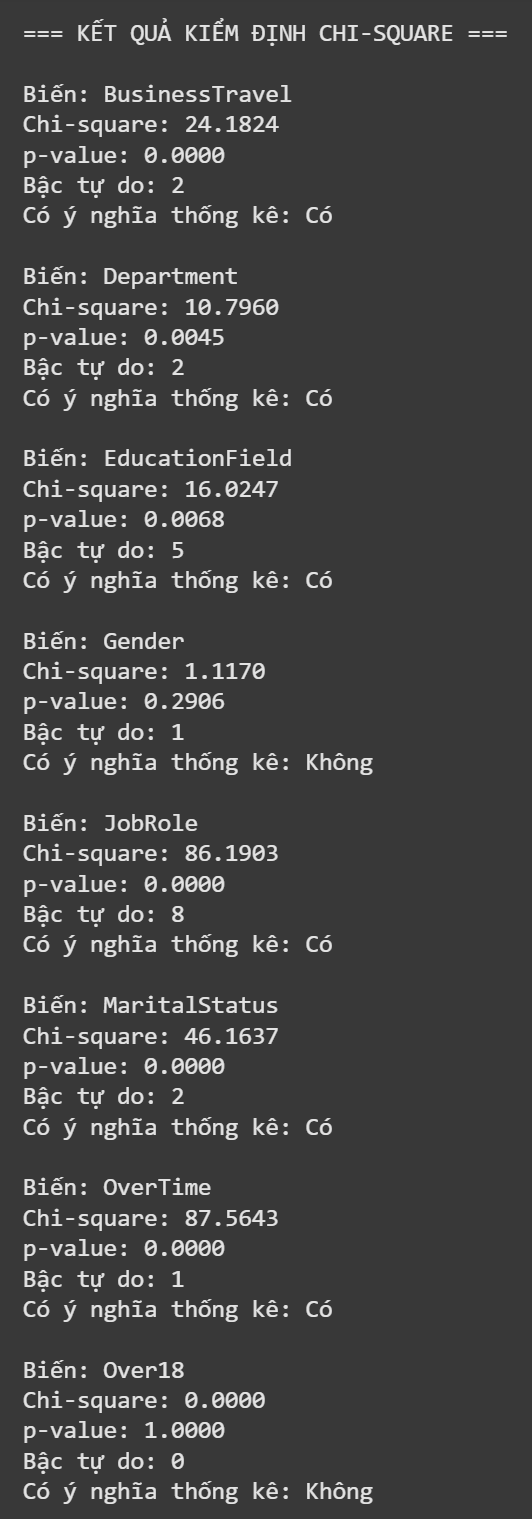
* **1,233 nhân viên (83.88%)** không nghỉ việc.
* **237 nhân viên (16.12%)** nghỉ việc.

### **2.1.2 Phân tích Ma trận Tương quan**

Dựa trên ma trận tương quan, các biến có tương quan mạnh với Attrition (|r| > 0.1) bao gồm:

* OverTime (r = 0.2461) - tương quan dương mạnh nhất
* MaritalStatus (r = 0.1621)
* JobSatisfaction (r = -0.1035)
* EnvironmentSatisfaction (r = -0.1034)
* JobInvolvement (r = -0.1300)
* TotalWorkingYears (r = -0.1711)
* YearsAtCompany (r = -0.1344)
* YearsWithCurrManager (r = -0.1562)

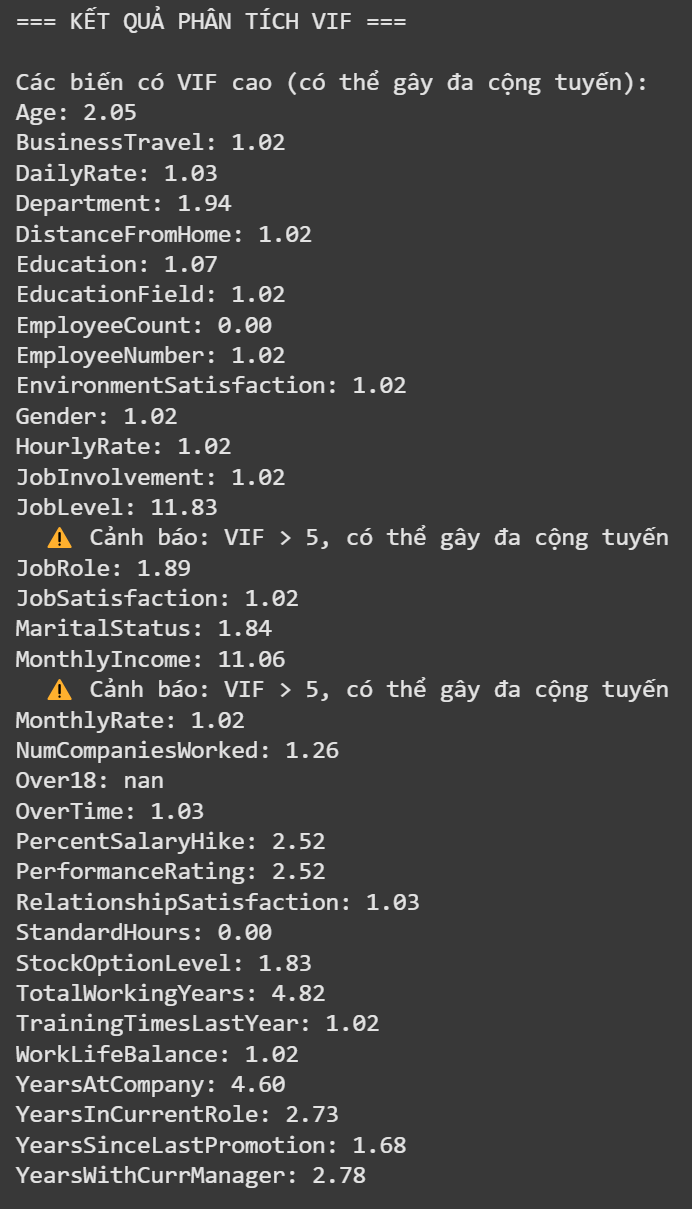
### **2.1.3 Kiểm định Chi-square**

****

Kết quả kiểm định Chi-square cho các biến phân loại:

* Có ý nghĩa thống kê (p < 0.05): OverTime, MaritalStatus, Department, JobRole, BusinessTravel, EducationField
* Không có ý nghĩa thống kê: Gender, Over18

### **2.1.4 Phân tích VIF**

****

Các biến có VIF cao (VIF > 5):

* JobLevel (VIF = 11.83)
* MonthlyIncome (VIF = 11.06)

### **2.1.5 Kết quả Lựa chọn Biến**

**Biến được chọn (24 biến):**

* Biến phân loại: OverTime, MaritalStatus, Department, JobRole, BusinessTravel, EducationField
* Biến số: Age, DistanceFromHome, Education, YearsAtCompany, YearsWithCurrManager, TotalWorkingYears, NumCompaniesWorked, JobSatisfaction, EnvironmentSatisfaction, JobInvolvement, RelationshipSatisfaction, WorkLifeBalance, PercentSalaryHike, PerformanceRating, StockOptionLevel, TrainingTimesLastYear

**Biến bị loại:**

* Không có ý nghĩa thống kê: Gender, Over18
* VIF cao: JobLevel, MonthlyIncome
* Không cần thiết: EmployeeCount, StandardHours, EmployeeNumber, DailyRate, HourlyRate, MonthlyRate
* Tương quan cao với biến khác: YearsInCurrentRole

## 

## **2.2 Input**

* **Nguồn dữ liệu:** IBM HR Analytics Employee Attrition & Performance
* **Kích thước ban đầu:** 1,471 mẫu với 35 biến
* **Biến mục tiêu:** Attrition (Yes/No)

### **2.2.1 Tiền Xử lý Dữ liệu**

#### **1. Mã hóa Biến**

Sử dụng LabelEncoder để mã hóa biến mục tiêu Attrition

Phân phối ban đầu:

* Nghỉ việc (Yes): 16.1%
* Không nghỉ việc (No): 83.9%

#### **2. Loại bỏ Biến**

**Biến không có phương sai:**

* EmployeeCount (VIF = 0.00)
* StandardHours (VIF = 0.00)

**Biến không có ý nghĩa thống kê:**

* Over18 (p-value = 1.0000)
* Gender (p-value = 0.2906)

**Biến có VIF cao:**

* JobLevel (VIF = 11.83)
* MonthlyIncome (VIF = 11.06)

**Biến không cần thiết:**

* EmployeeNumber (ID)
* DailyRate, HourlyRate, MonthlyRate (thông tin trùng lặp)

#### **3. Xử lý Biến Phân loại**

**Mã hóa các biến phân loại bằng LabelEncoder:**

* BusinessTravel
* Department
* EducationField
* JobRole
* MaritalStatus
* OverTime

#### **4. Chia Dữ liệu**

**Tỷ lệ train/test: 80/20**

* Kích thước tập training: (1,177 mẫu, 24 biến)
* Kích thước tập test: (294 mẫu, 24 biến)

#### **2.5 Xử lý Outlier**

Sử dụng phương pháp IQR (Interquartile Range)

Cắt giá trị ngoại lai ở mức Q1 - 1.5IQR và Q3 + 1.5IQR

#### **2.6 Chuẩn hóa Dữ liệu**

Sử dụng StandardScaler cho các biến số

Chuyển đổi dữ liệu về phân phối chuẩn với mean = 0 và std = 1

#### **2.7 Xử lý Mất Cân bằng Dữ liệu**

**Sử dụng SMOTE (Synthetic Minority Over-sampling Technique)**

**Kết quả sau xử lý:**

* Nghỉ việc: 50%
* Không nghỉ việc: 50%

Kích thước tập training sau SMOTE: (1,956 mẫu, 24 biến)

### **2.2.2. Biến Được Sử Dụng**

**24 biến cuối cùng được sử dụng trong mô hình:**

* **Thông tin cá nhân:** Age, MaritalStatus
* **Thông tin công việc:** Department, JobRole
* **Thông tin thâm niên:** YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
* **Mức độ hài lòng:** EnvironmentSatisfaction, JobSatisfaction, RelationshipSatisfaction
* **Các yếu tố khác:** BusinessTravel, DistanceFromHome, OverTime, NumCompaniesWorked

## **2.3 Output**

### **2.3.1 Hiệu suất Mô hình**

#### **1. Mô hình Cơ bản**

**Accuracy:** 72.11%

**Precision:** 26%

**Recall:** 59%

**F1-Score:** 36%

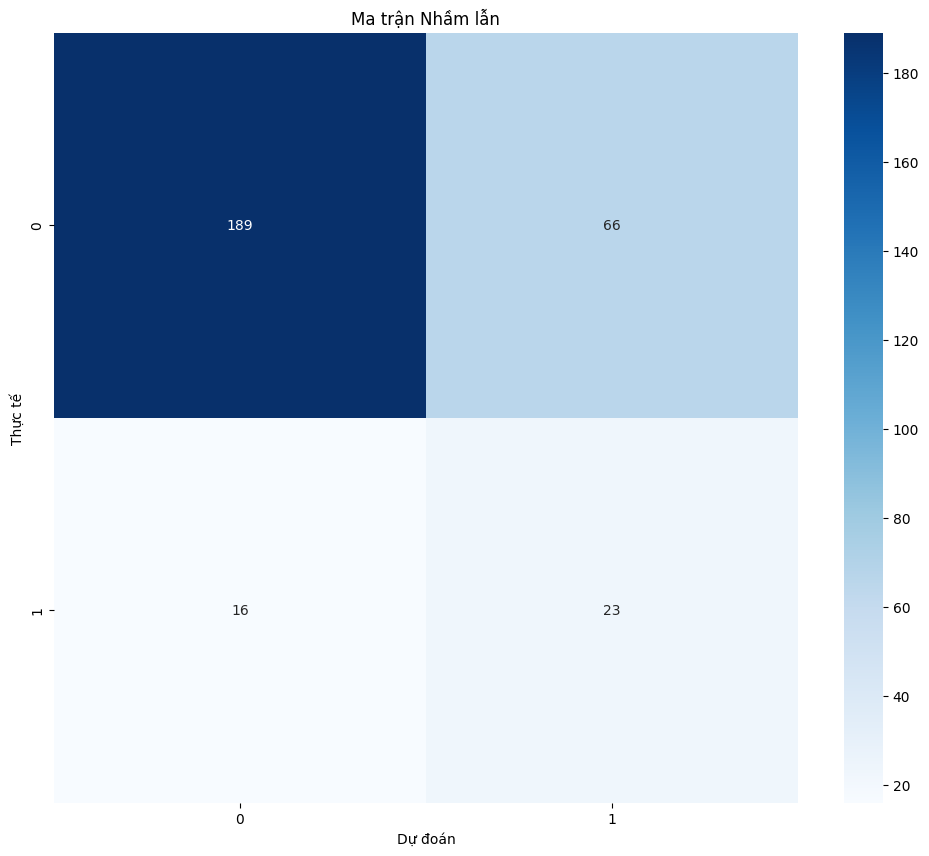
**Thời gian huấn luyện:** 0.0223s

**Kết quả dự đoán trên tập test**:

* **Dự đoán đúng:** 212 mẫu
* **Dự đoán sai:** 82 mẫu

**Tỷ lệ dự đoán đúng:** 72.11%

**Confusion Matrix:**



**True Negative (TN):** 189

**False Positive (FP):** 66

**False Negative (FN):** 116

**True Positive (TP):** 23

#### **2. Mô hình GridSearchCV**

**Accuracy:** 68.71%

**Precision:** 24%

**Recall:** 64%

**F1-Score:** 35%

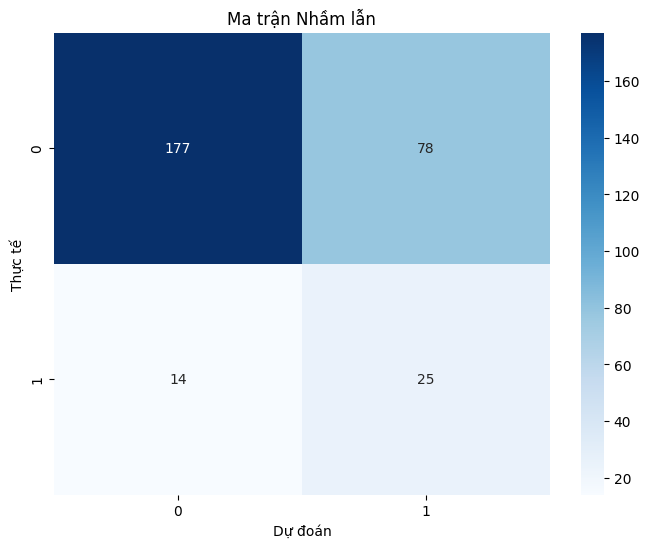
**Thời gian huấn luyện:** 6.5810s

**Kết quả dự đoán trên tập test:**

* **Dự đoán đúng:** 202 mẫu
* **Dự đoán sai:** 92 mẫu

**Tỷ lệ dự đoán đúng:** 68.71%

**Confusion Matrix:**



**True Negative (TN):** 177

**False Positive (FP):** 78

**False Negative (FN):** 14

**True Positive (TP):** 25

**Tham số tối ưu:**

* **C:** 0.01
* **max\_iter:** 1000
* **penalty:** l2
* **solver:** liblinear

# **3. Phương pháp xử lý (Model)**

## **3.1 Phương pháp Hồi quy Logistic**

**Hồi quy Logistic được chọn làm phương pháp chính cho bài toán dự đoán nghỉ việc vì:**

* Phù hợp với biến mục tiêu nhị phân (nghỉ việc/không nghỉ việc)
* Dễ diễn giải kết quả
* Hiệu quả với dữ liệu có kích thước vừa phải
* Có khả năng xử lý cả biến số và biến phân loại

## **3.2 Các phiên bản Mô hình**

### **3.2.1 Mô hình Cơ bản**

Sử dụng LogisticRegression với các tham số mặc định

Đặc điểm:

* Regularization: L2 (Ridge)
* Solver: liblinear
* Max iterations: 100
* Multi-class: auto
* Class weight: balanced

### **3.2.2 Mô hình GridSearchCV**

Tối ưu hóa tham số thông qua GridSearchCV

Không gian tìm kiếm tham số:

* C: [0.01, 0.1, 1, 10]
* max\_iter: [100, 500, 1000]
* penalty: ['l1', 'l2']
* solver: ['liblinear', 'saga']
* Cross-validation: 5-fold

## **3.3 Quy trình Huấn luyện**

**Chuẩn bị dữ liệu:**

1. Chia tập dữ liệu: 80% training, 20% testing
2. Xử lý mất cân bằng dữ liệu bằng SMOTE
3. Chuẩn hóa dữ liệu số bằng StandardScaler

**Huấn luyện mô hình:**

* Mô hình cơ bản: Huấn luyện trực tiếp
* Mô hình GridSearchCV: Tìm kiếm tham số tối ưu

**Đánh giá mô hình:**

* Độ chính xác (Accuracy)
* Precision, Recall, F1-score
* Ma trận nhầm lẫn
* Thời gian huấn luyện

## **3.4 Các Tham số Quan trọng**

**C (Inverse of regularization strength):**

* Điều chỉnh độ phức tạp của mô hình
* Giá trị nhỏ: Mô hình đơn giản hơn, tránh overfitting
* Giá trị lớn: Mô hình phức tạp hơn, có thể overfitting

**Penalty:**

* L1 (Lasso): Tạo sparse model, chọn lọc đặc trưng
* L2 (Ridge): Giảm thiểu hệ số, tránh overfitting

**Solver:**

* liblinear: Hiệu quả cho dữ liệu nhỏ
* saga: Phù hợp với dữ liệu lớn

## **3.5 Lý do Chọn Phương pháp**

**Đơn giản và Hiệu quả:**

* Dễ triển khai và bảo trì
* Thời gian huấn luyện nhanh
* Kết quả dễ diễn giải

**Phù hợp với Dữ liệu:**

* Xử lý tốt dữ liệu có kích thước vừa phải
* Có thể xử lý cả biến số và biến phân loại
* Hiệu quả với dữ liệu mất cân bằng

**Khả năng Mở rộng:**

* Dễ dàng thêm biến mới
* Có thể kết hợp với các kỹ thuật nâng cao
* Phù hợp cho việc triển khai trong thực tế

# **4. Kết quả và phân tích**

## **4.1 Hiệu suất mô hình**

Bảng dưới đây thể hiện các chỉ số đánh giá của hai mô hình: **Mô hình cơ bản** (Logistic Regression với tham số mặc định) và **Mô hình GridSearchCV** (được tối ưu tham số).

| **Chỉ số** | **Mô hình cơ bản** | **Mô hình GridSearchCV** |
| --- | --- | --- |
| **Accuracy** | 72.11% | 68.71% |
| **Precision** | 26% | 24% |
| **Recall** | 59% | 64% |
| **F1-Score** | 36% | 35% |
| **Thời gian huấn luyện** | 0.0223s | 6.5810s |

* **Mô hình cơ bản có độ chính xác cao hơn (72.11% so với 68.71%)**, nhưng khả năng phát hiện nhân viên nghỉ việc (Recall) **thấp hơn** so với mô hình tối ưu hóa tham số.
* **GridSearchCV cải thiện Recall từ 59% lên 64%**, giúp phát hiện các trường hợp nghỉ việc tốt hơn. Tuy nhiên, Precision lại giảm nhẹ từ 26% xuống 24%, dẫn đến **F1-score gần như không thay đổi**.
* **Thời gian huấn luyện của GridSearchCV cao hơn đáng kể**, mất **6.5810 giây**, trong khi mô hình cơ bản chỉ mất **0.0223 giây**.

## **4.2 Tham số tối ưu từ GridSearchCV**

Quá trình tìm kiếm tham số tối ưu với GridSearchCV đã chọn được các giá trị sau:

* **C**: 0.01
* **max\_iter**: 1000
* **penalty**: l2
* **solver**: liblinear

Những tham số này giúp cải thiện khả năng phân biệt nhân viên nghỉ việc nhưng không làm tăng đáng kể độ chính xác tổng thể.

## **4.3 Đặc trưng quan trọng nhất**

Các đặc trưng quan trọng được chọn bởi hai mô hình như sau:

* **Mô hình cơ bản**:  
  + **OverTime** (làm thêm giờ)
  + **YearsAtCompany** (số năm làm việc tại công ty)
  + **TotalWorkingYears** (tổng số năm kinh nghiệm làm việc)
  + **YearsInCurrentRole** (số năm trong vai trò hiện tại)
  + **YearsWithCurrManager** (số năm làm việc với quản lý hiện tại)
* **Mô hình GridSearchCV**:  
  + **OverTime** (làm thêm giờ)
  + **MaritalStatus** (tình trạng hôn nhân)
  + **TotalWorkingYears** (tổng số năm kinh nghiệm làm việc)
  + **YearsInCurrentRole** (số năm trong vai trò hiện tại)
  + **JobSatisfaction** (mức độ hài lòng với công việc)

##### **Nhận xét:**

* **OverTime (Làm thêm giờ) là yếu tố quan trọng nhất trong cả hai mô hình**, cho thấy những nhân viên thường xuyên làm thêm có tỷ lệ nghỉ việc cao hơn.
* **YearsAtCompany, TotalWorkingYears, và YearsInCurrentRole** đều đóng vai trò quan trọng, cho thấy thời gian làm việc tại công ty ảnh hưởng lớn đến quyết định nghỉ việc.
* **Mô hình GridSearchCV thay thế YearsWithCurrManager bằng MaritalStatus và JobSatisfaction**, có thể do những yếu tố này tác động mạnh hơn khi tối ưu tham số.

## **4.4 Phân tích phân bố dữ liệu**

Các biểu đồ phân bố cho thấy:

* Nhân viên trẻ tuổi có tỷ lệ nghỉ việc cao hơn.
* Những người làm việc trong công ty dưới **5 năm** có xu hướng nghỉ việc nhiều hơn.
* Nhân viên có **JobSatisfaction thấp và OverTime cao** thường có tỷ lệ nghỉ việc lớn hơn.
* Những người có **số năm làm việc với quản lý thấp** cũng có xu hướng nghỉ việc nhiều hơn, có thể do chưa có sự gắn kết với quản lý.

Kết quả này phù hợp với thực tế, khi nhân viên mới hoặc làm việc quá tải dễ dẫn đến tình trạng nghỉ việc.

# **Kết luận**

Dự án đã cho thấy hiệu quả của việc sử dụng mô hình Logistic Regression trong việc dự đoán khả năng nghỉ việc của nhân viên. Mô hình cơ bản đạt độ chính xác cao hơn và thời gian huấn luyện nhanh hơn, phù hợp để sử dụng trong các tình huống yêu cầu phản hồi nhanh và độ chính xác tổng thể tốt. Trong khi đó, mô hình được tối ưu tham số bằng GridSearchCV cho kết quả tốt hơn về khả năng phát hiện các trường hợp nghỉ việc, nhờ điểm Recall cao hơn, dù thời gian huấn luyện lâu hơn. Những kết quả này nhấn mạnh tầm quan trọng của việc cân bằng giữa độ chính xác tổng thể và hiệu suất của từng lớp, đặc biệt là lớp có số liệu ít hơn. Dựa trên những kết quả đạt được, cần tiếp tục thử nghiệm các phương pháp khác như xử lý dữ liệu mất cân bằng, mở rộng tập dữ liệu, và áp dụng các thuật toán tiên tiến hơn để cải thiện hiệu suất dự đoán trong tương lai.