Phân tích các yếu tố ảnh hưởng đến sự hao mòn nhân lực và dự đoán khả năng nghỉ việc của nhân viên bằng các thuật toán học máy

Võ Đức Dương
duong922003@gmail.com

Tóm tắt nội dung - *Attrition Rate (tỉ lệ hao mòn nhân sự)* cao là tình trạng đáng lo ngại cho một doanh nghiệp vì nó ảnh hưởng tiêu cực đến năng lực tài chính cũng như khả năng phát triển lâu dài. Do đó, doanh nghiệp cần chủ động chuẩn bị bằng cách xác định những nhân viên có khả năng nghỉ việc và lý do họ rời bỏ tổ chức.

Trong nghiên cứu này, bộ dữ liệu IBM Human Resource Analytics Employee Attrition được khai thác bằng nhiều phương pháp khai phá dữ liệu và trích xuất đặc trưng khác nhau, giúp làm sạch và cải thiện chất lượng dữ liệu, từ đó lựa chọn được các đặc điểm phù hợp, nâng cao hiệu quả của các mô hình. Có tổng cộng 34 đặc trưng trong tập dữ liệu gốc. Kết quả phân tích cho thấy các yếu tố ảnh hưởng nhiều nhất là *độ tuổi*, *thu nhập hàng tháng*, *số năm làm việc* và *mức độ hài lòng với môi trường làm việc*. Trong khi đó, các yếu tố như ngành học, phòng ban, giới tính hay đánh giá hiệu suất làm việc hầu như không có nhiều tác động đến xu hướng rời bỏ công ty của nhân viên. Tổng cộng 25 đặc trưng quan trọng nhất được lựa chọn, 4 mô hình được lựa chọn để điều chỉnh tham số và đánh giá chi tiết, bao gồm Logistic Regression, CatBoost, Stacking (kết hợp Logistic Regression với CatBoost) và Neural Network.

Kết quả cuối cùng cho thấy mô hình mạng học sâu cho kết quả tốt nhất trên cả hai chỉ số ROC-AUC (0.7495) và F1-Score (0.5263), mô hình CatBoost cho độ chính xác cao nhất đạt 0.8571. Trong số các mô hình học máy cổ điển, mô hình Stacking đạt chỉ số F1-Score cao nhất (0.5) cùng với độ chính xác đạt 0.8503.

I. Giới thiệu

Hiện nay, tình hình nghỉ việc của nhân viên đang ở mức đáng chú ý. Theo thống kê của Vieclam24h.vn - website việc làm phổ biến hàng đầu Việt Nam, khoảng 75% người lao động có dự định nhảy việc trong 6 tháng đầu năm 2023 [1]. Theo một báo cáo khác của bộ Nội Vụ, tổng số công chức, viên chức thôi việc từ giữa năm 2022 đến giữa năm 2023 là 18.991 người, trong đó chủ yếu ở nhóm viên chức sự nghiệp giáo dục - đào tạo và sự nghiệp y tế [2]. Employee Attrition (Hao mòn nhân sự) xảy ra khi nhân viên nghỉ việc mà công ty không tuyển thay thế hay mất rất nhiều thời gian để tuyển thay thế. Attrition Rate cao ảnh hưởng xấu đến doanh nghiệp, vì khi một lượng lớn nhân viên nghỉ việc, doanh nghiệp sẽ phải tổ chức tuyển dụng [9]. Chi phí tổ chức phỏng vấn, tuyển dụng, đào tạo rất đắt đỏ, ảnh hướng đến nguồn lực tài chính của doanh nghiệp. Đặc biệt, với nghỉ việc tự nguyện, những nhân viên nghỉ việc thường là những nhân viên giỏi, việc họ nghỉ việc sẽ khiến doanh nghiệp mất đi nhân lực trình độ cao. Ngoài ra, với những doanh nghiệp có dây chuyền sản xuất liên tục và nối tiếp, việc nhân viên nghỉ việc đột ngột có thể làm gián đoạn, trì trệ khả năng sản xuất của doanh nghiệp.

Do những ảnh hưởng từ Attrition Rate cao mang lại, việc phải có những sự chuẩn bị kĩ càng từ phía doanh nghiệp là hết sức cần thiết. Trong đó, doanh nghiệp cần trả lời được hai câu hỏi. *Một là*, những nhân viên có đặc điểm như thế nào thì sẽ nghỉ việc trong thời gian gần? *Hai là*, với những đặc điểm cho sẵn, liệu một nhân viên sẽ nghỉ việc trong thời gian gần? Với câu hỏi đầu tiên, đầu ra mong đợi của doanh nghiệp là các đặc điểm sẽ làm nhân viên nghỉ việc, chẳng hạn như lương, thưởng, sự thách thức, sự tự do, áp lực,... Biết được những đặc điểm này sẽ giúp công ty thay đổi hay cải thiện nơi làm việc, chế độ đãi ngộ,... từ đó làm giảm được Attrition Rate, giúp công ty giảm chi phí tuyển dụng và giữ được nhân lực giỏi. Với câu hỏi thứ hai, đầu ra mong đợi của doanh nghiệp là khả năng một nhân viên sẽ nghỉ việc. Biết được khả năng nghỉ việc sẽ giúp doanh nghiệp biết những đặc điểm cần cải thiện, thay đổi, hay chuẩn bị trước được nhân lực thay thế. Việc này giúp dây chuyền làm việc của công ty không bị trì trệ và giúp bộ phận tuyển dụng nắm được tình hình, từ đó có kế hoạch tuyển dụng tốt hơn.

Để trả lời được những câu hỏi trên, người ta thường thu thập dữ liệu với số lượng lớn sau đó rút trích đặc trưng từ dữ liệu. Với lượng dữ liệu lớn đến cực lớn, việc rút trích đặc trưng bằng tay tỏ ra không hiệu quả cả về mặt thời gian và độ chính xác. Thay vào đó, các mô hình máy học sẽ là công cụ hỗ trợ rất tốt để phân tích và dự đoán trên dữ liệu. Các mô hình này có thể tìm ra những quy luật đằng sau những thông số có được từ dữ liệu, thực hiện dự đoán ở tốc độ rất nhanh, độ chính xác cao và không bị mệt mỏi hay thiên kiến.

Trong bài báo cáo này, các phân tích về dữ liệu sẽ được triển khai để xem xét các đặc điểm của nhóm nhân viên sẽ nghỉ việc và sẽ không nghỉ việc. Sau đó, các mô hình máy học trong nhiều nhóm khác nhau (mô hình Naive Bayes, mô hình Bagging, mô hình Boosting và mô hình mạng học sâu) sẽ được sử dụng để cố gắng giải quyết bài toán dự đoán khả năng nghỉ việc của nhân viên. Bộ dữ liệu được dùng là bộ IBM Human Resource Analytic Employee Attrition có từ Kaggle Dataset Repository. Dự án (bao gồm bài báo cáo này và mã nguồn) đã được public

ở github.

Cấu trúc của các phần tiếp theo của bài báo cáo như sau: Phần II nói về bộ dữ liệu được dùng, phân tích các yếu tố có trong tập dữ liệu và thực hiện tiền xử lý dữ liệu; Tiếp đó, Phần III nói về các mô hình máy học được sử dụng, cài đặt thực nghiệm và độ đo; Sau đó, Phần IV nói về các thử nghiệm tinh chỉnh mô hình; Tiếp theo, Phần V phân tích các lỗi của mô hình và vạch ra hướng phát triển trong tương lai; Cuối cùng, Phần VI là phần kết luận của bài báo cáo.

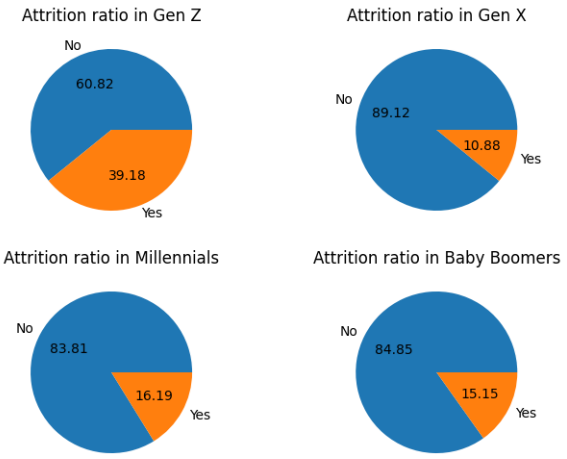
II. Dữ liệu

A. Về bộ dữ liệu được dùng

Bộ dữ liệu IBM Human Resource Analytic Employee Attrition (IBM-HRAEA) là bộ dữ liệu giả tưởng được tạo ra bởi các nhà khoa học ở IBM và được công khai trên Kaggle[1]. Bộ dữ liệu bao gồm 1470 dòng dữ liệu với 35 cột, bao gồm các thông tin nhân khẩu học, lương, kinh nghiệm và mức độ hài lòng của nhân viên. Mỗi một dòng tương ứng với một nhân viên, trong đó biến mục tiêu *Attrition* chỉ ra nhân viên có rời bỏ công ty hay không. Trong 1470 nhân viên, có 86% nhân viên rời bỏ (Attrition = "Yes") và 14% nhân viên ở lại công ty (Attrition = "No")

B. Phân tích ảnh hưởng của một số thuộc tính đến khả năng nghỉ việc của nhân viên

1) Tuổi và thế hệ: Độ tuổi (hay thế hệ) khác nhau cho thấy những xu hướng khác nhau trong rời bỏ công ty hay tổ chức

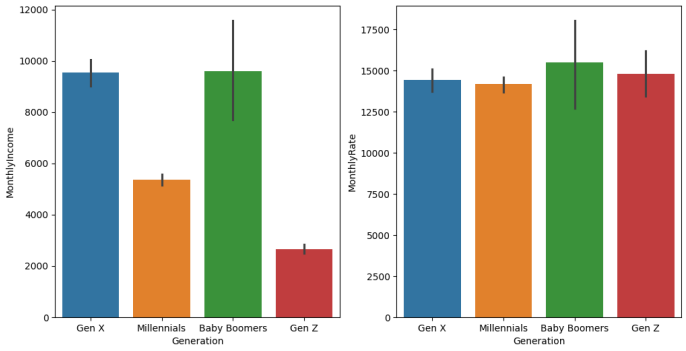


Hình 1. Tỉ lệ nghỉ việc trong từng nhóm tuổi

Có 4 nhóm thế hệ được phân tích, bao gồm *Gen Z* (dưới 24 tuổi), *Millennials* (từ trên 24 đến dưới 40 tuổi), *Gen X* (từ trên 40 tuổi đến dưới 56 tuổi) và *Baby Boomers* (trên 56 tuổi)

Biểu đồ *Hình 1* cho thấy sự chênh lệch trong tỉ lệ nghỉ việc của nhóm *Gen Z* so với 3 thế hệ lớn tuổi hơn khi đạt đến gần 40%, trong khi tỉ lệ này chỉ từ hơn 10% đến khoảng 16% ở các nhóm còn lại, cho thấy nhóm thế hệ trẻ có xu hướng rời bỏ công ty nhiều hơn. Ngoài ra, các phân tích (được đề cập ở mục sau) còn cho thấy mối quan hệ giữa nhóm tuổi và một số yếu tố khác trong ảnh hưởng đến sự nghỉ việc của nhân viên.

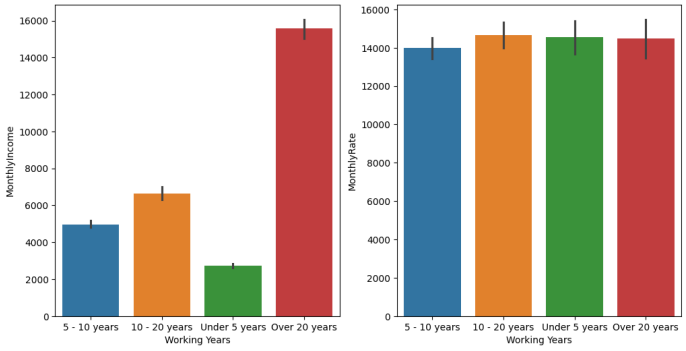
2) Mức Thu Nhập Hàng Tháng: Mức thu nhập hàng tháng có ảnh hưởng lớn đến sự nghỉ việc của nhân viên.



Hình 2. Thu nhập (trái) và mức lương (phải) hàng tháng của nhân viên

Mức lương hàng tháng được trả ở 4 nhóm tuổi không có quá nhiều khác biệt. Trong khi đó, mức thu nhập hàng tháng cao nhất ở nhóm Baby Boomers và X, thấp nhất ở nhóm Gen Z. Bên cạnh đó, mặc dù Millennials có mức thu nhập trung bình thấp hơn gần một nửa so với Baby Boomers nhưng tỉ lệ nghỉ việc không cao. Đối chiếu với kết quả phân tích theo thế hệ, có thể thấy rằng nhóm độ tuổi cao hơn thì có mức thu nhập hàng tháng cao hơn, môi trường làm việc ổn định và ít có xu hướng rời bỏ công ty hơn.

3) Số Năm Làm Việc: Phân tích cho thấy số năm làm việc cũng có ảnh hưởng đến xu hướng nghỉ việc của nhân viên.

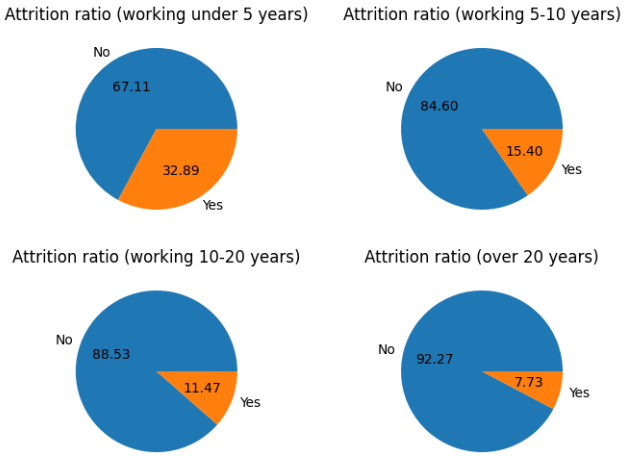


Hình 3. Thu nhập (trái) và mức lương (phải) hàng tháng theo từng nhóm kinh nghiệm làm việc

Biểu đồ *Hình 3* cho thấy tương quan giữa thu nhập hàng tháng và số năm kinh nghiệm làm việc: số năm làm việc càng cao thì thu nhập hàng tháng cũng cao hơn. Đáng chú ý rằng mặc dù mức lương được trả hàng tháng ở 4 nhóm kinh nghiệm không chênh lệch nhiều thì tỉ lệ nghỉ việc giữa các nhóm này (*Hình 4*) lại có sự khác biệt rõ rệt.

Dễ nhận thấy tỉ lệ nghỉ việc tỉ lệ nghịch với số năm làm việc. Có nghĩa rằng các nhân viên có số năm kinh nghiệm làm việc

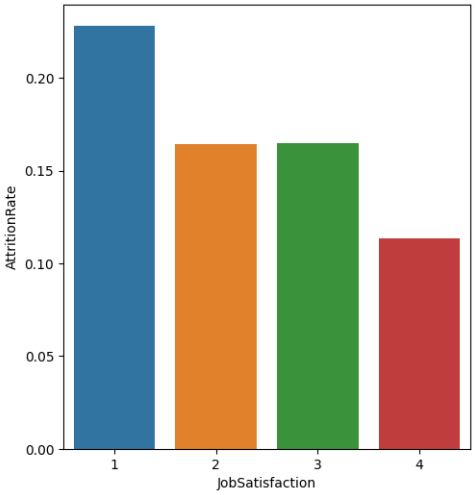
[1]www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset



Hình 4. Tỉ lệ nghỉ việc theo từng nhóm kinh nghiệm làm việc

càng thấp thì càng có xu hướng rời bỏ công ty nhiều hơn và ngược lại. Điều này cũng cho thấy những nhân viên có nhiều kinh nghiệm làm việc thường có môi trường làm việc ổn định.

4) Các yếu tố khác: Ngoài độ tuổi, thu nhập và số năm làm việc có ảnh hưởng nhiều nhất, một số yếu tố khác cũng đóng góp vào sự nghỉ việc của nhân viên, chẳng hạn như mức độ hài lòng với công việc thấp hơn cho thấy nhân viên có nhiều khả năng muốn rời bỏ tổ chức hơn.

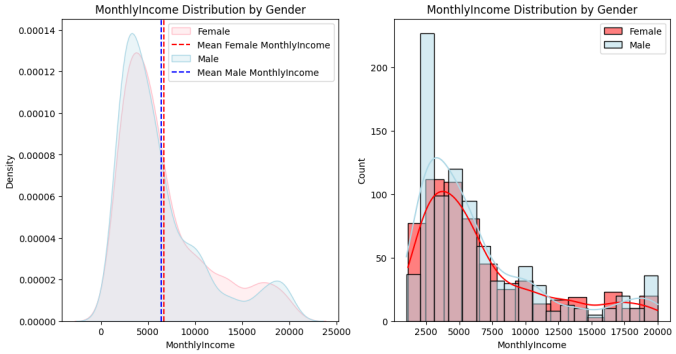


Hình 5. Tỉ lệ nghỉ việc theo mức độ hài lòng công việc

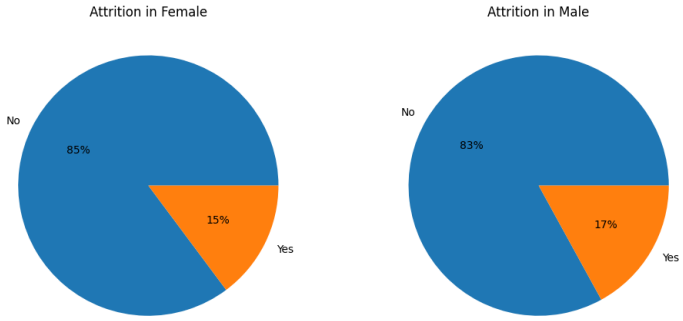
Giống với sự hài lòng về công việc, sự hài lòng về môi trường làm việc, quan hệ, mức độ cân bằng giữa công việc với cuộc sống hay mức độ tham gia vào công việc cũng thể hiện các đặc điểm tương tự.

Bên cạnh đó, có nhiều yếu tố không thể hiện rõ sự ảnh hưởng hay đặc điểm của nó đến sự nghỉ việc của nhân viên. Chẳng hạn như giới tính (*Hình 7*)

Phân bố mức thu nhập (ngoài ra còn có mức lương, mức độ hài lòng, hiệu suất công việc hay tăng ca), thậm chí cả tỉ lệ nghỉ

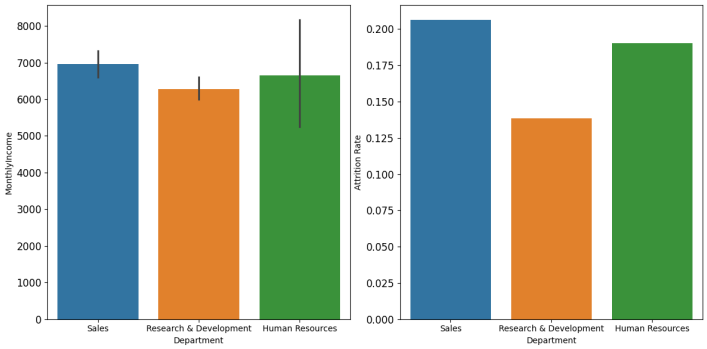


Hình 6. Phân bố mức thu nhập theo giới tính



Hình 7. Tỉ lệ nghỉ việc theo giới tính

việc gần như tương đồng ở cả hai giới, cho thấy rằng giới tính không ảnh hưởng đến xu hướng rời bỏ tổ chức của nhân viên.



Hình 8. Thu nhập hàng tháng (trái) và tỉ lệ nghỉ việc (phải) theo từng phòng ban

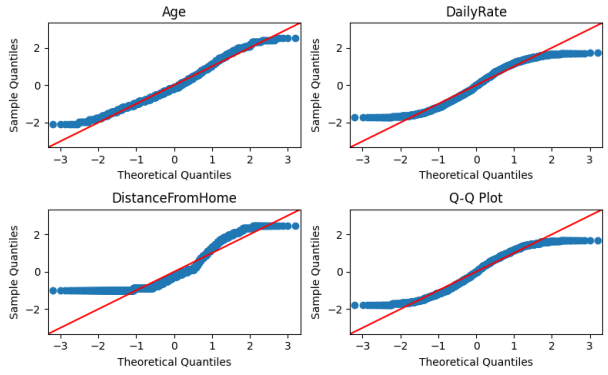
Mặc dù tỉ lệ nghỉ việc có sự chênh lệch giữa các phòng ban, nhưng mức thu nhập hàng tháng về cơ bản không quá khác biệt. Ngoài ra, mặc dù R&D có mức thu nhập hàng tháng bình quân thấp nhất nhưng cũng có tỉ lệ nghỉ việc thấp nhất. Điều tương tự cũng xảy ra với nhóm Sales và Human Resources, cho thấy rằng phòng ban không có ảnh hưởng rõ ràng hoặc không ảnh hưởng đến sự nghỉ việc của nhân viên.

C. Tiền xử lý dữ liệu

1) Kiểm tra dữ liệu khuyết và dữ liệu trùng lặp: Tập dữ liệu đầy đủ, không có dữ liệu bị thiếu, không có bất kì dữ liệu nào bị trùng lặp.

2) Xử lý kí tự đặc biệt: Trong tập dữ liệu có hai cột dạng phân loại chứa ký tự đặc biệt là Department (Research & Development) và BusinessTravel (Travel_Rarely, Travel_Frequently, NonTravel). Các ký tự đặc biệt như dấu "&"hoặc "_"không mang ý nghĩa đặc biệt trong ngữ cảnh phân loại và không ảnh hưởng đến mục tiêu dự báo của bài toán. Do đó, ta giữ nguyên các giá trị này và áp dụng phương pháp *Label Encoding* để chuyển đổi sang dạng số, phục vụ cho việc huấn luyện mô hình.

3) Loại bỏ giá trị ngoại lai: Có 3 phương pháp phát hiện giá trị ngoại lai được sử dụng, bao gồm *Z-Score*, *IQR* và *Quantile*. Mỗi phương pháp áp dụng cho các đặc trưng nhất định phụ thuộc vào phân phối dữ liệu của đặc trưng đó.



Hình 9. Phân phối của một số đặc trưng dạng số

Thuộc tính *Age* được áp dụng phương pháp Z-Score do có phân phối gần với phân phối chuẩn. Các thuộc tính gần với phân phối đều (ví dụ HourlyRate) và các thuộc tính có phân phối lệch phải (ví dụ DistanceFromHome) sẽ sử dụng phương pháp IQR. Phương pháp Quantile được sử dụng cho các thuộc tính còn lại.

Với mỗi thuộc tính, ta xác định giá trị *lower_limit* và *upper_limit* tương. Công thức cho các giá trị này lần lượt như sau.

Đối với phương pháp Z-Score:

$$lower\_limit = Mean - 3 * std \quad (1)$$

$$upper\_limit = Mean + 3 * std \quad (2)$$

Đối với phương pháp IQR:

$$IQR = Q3 - Q1 \quad (3)$$

$$lower\_limit = Q3 + 1.5 * IQR \quad (4)$$

$$upper\_limit = Q1 - 1.5 * IQR \quad (5)$$

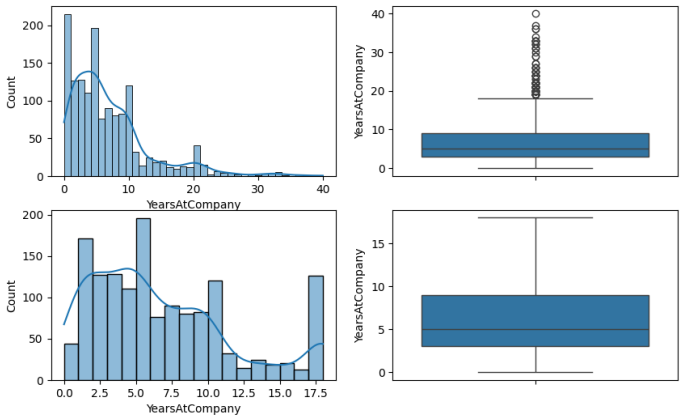
Đối với phương pháp Quantile:

$$lower\_limit = Quantile(0.01) \quad (6)$$

$$upper\_limit = Quantile(0.99) \quad (7)$$

Các giá trị ngoại lai sẽ được thay thế (capping) bằng giá trị giới hạn. Có nghĩa rằng nếu $x > upper\_limit$ thì thay thế $x = upper\_limit$, và nếu $x < lower\_limit$ thì thay thế $x = lower\_limit$.

Mặc dù vậy, có một vấn đề đối với việc thay thế hay loại bỏ các giá trị ngoại lai, bởi vì chúng đại diện cho dữ liệu thực tế chứ không phải phát sinh do lỗi dữ liệu.



Hình 10. Phân phối của biến *YearsAtCompany* trước và sau khi loại bỏ outlier

Trong *Hình 10*, Years > 18 đều được coi là outlier và các giá trị này được thay thế bằng 18. Điều này vừa làm mất đi ý nghĩa ban đầu, vừa làm thay đổi phân phối giá trị của biến. Mặt khác, mô hình phải có khả năng dự đoán cho các khoảng giá trị khác nhau, chẳng hạn như độ tuổi. Do đó, các giá trị này vẫn sẽ được giữ lại để huấn luyện mô hình.

4) Biến đổi giá trị: Tương tự với việc phát hiện giá trị ngoại lai, biến đổi giá trị cũng sẽ dựa vào phân phối của từng biến để áp dụng phương pháp phù hợp. Các biến biến gần với phân phối đều hoặc phạm vi giá trị hẹp sẽ áp dụng chuẩn hóa (MinMax Scaling). Các biến lệch phải sẽ được logarit hóa (Log Transformation) để giảm độ lệch. Các biến còn lại, đặc biệt các biến gần với phân phối chuẩn sẽ được tiêu chuẩn hóa (Standard Scaling) để đưa về phân phối chuẩn.

5) Xử lý mất cân bằng dữ liệu: Do sự chênh lệch về số lượng mẫu của mỗi lớp trong tập dữ liệu, ta cần xử lý mất cân bằng trước khi tiến hành huấn luyện mô hình để giảm thiên lệch về lớp đa số (ở đây là Attrition = No), tăng khả năng tổng quát hóa.

Có nhiều hướng tiếp cận để xử lý mất cân bằng dữ liệu. Đồ án và bài báo cáo này trình bày phương pháp ADASYN cho các mô hình cổ điển và phương pháp đánh trọng số cho mô hình học sâu.

ADASYN tập trung vào việc tạo ra các mẫu dữ liệu của lớp thiểu số ở những vùng khó học hơn, tức là những điểm gần với ranh giới lớp. Trước tiên, ADASYN xác định số lượng mẫu synthetic cần tạo cho mỗi điểm thuộc lớp thiểu số dựa

trên độ khó học (learning difficulty) tại điểm đó. Với mỗi một điểm dữ liệu x_i thuộc lớp thiểu số, ta tìm ***k*** hàng xóm gần nhất với nó. Độ khó học được xác định bởi công thức $r_i = \frac{\text{số hàng xóm thuộc lớp đa số}}{k}$, r_i càng cao thì x_i càng khó học và càng cần nhiều mẫu tổng hợp.
Tiếp theo, ADASYN tính toán tổng số mẫu cần tạo: $G = (N_{maj} - N_{min}) * \beta$, với β là siêu tham số điều chỉnh mức độ oversampling. Trong đồ án này, giá trị β được đặt bằng 0.5. Dựa vào r_i, ADASYN phân bổ tổng G mẫu synthetic cho từng điểm x_i: $g_i = \frac{r_i}{\Sigma r_j} \times G$.

Cuối cùng, với mỗi một điểm x_i, ADASYN tạo mẫu mới theo công thức: $x_{new} = x_i + \delta * (x_{zi} - x_i)$, với x_{zi} là một trong số các láng giềng của x_i và $\delta \in [0, 1]$ là hệ số ngẫu nhiên. Trong thư viện *sklearn*, siêu tham số δ được điều chỉnh một cách gián tiếp thông qua tham số *random_state*.

Đối với mô hình mạng học sâu, tập dữ liệu huấn luyện không được xử lý một cách trực tiếp. Thay vào đó, mô hình được điều chỉnh thông qua tham số *weight* (trọng số các lớp) của hàm lỗi *CrossEntropyLoss*: Loss $= -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(p_{i,y_i})$ với $\log(p_{i,y_i})$ là xác suất (softmax) mà mô hình dự đoán đúng cho lớp y_i.

6) Phân chia tập dữ liệu: Bộ dữ liệu được chia thành 3 tập khác nhau, bao gồm tập *train*, *test* và *valid*. Trong đó, tập train được dùng để huấn luyện mô hình, chiếm 80% tổng số dữ liệu, tập valid và test mỗi tập chiếm 10% dữ liệu. Tập valid được sử dụng để điều chỉnh mô hình và tập test được sử dụng cho bước đánh giá cuối cùng. Tỉ lệ giữa các lớp được giữ nguyên ở cả 3 tập thông qua tham số *stratify* trong thư viện *sklearn*. Các tập dữ liệu này sau khi xử lý thì được lưu trữ và sử dụng cố định cho toàn bộ các mô hình.

III. Xây dựng mô hình

Nhiều phương pháp máy học được thử nghiệm trong bài báo cáo, nhằm khai thác và khám phá triệt để hiệu quả của các phương pháp này lên bộ dữ liệu đã trình bày. Trong đó, các phương pháp máy học mà bài báo cáo đề xuất có thể được chia thành năm nhóm chính: Mô hình tuyến tính, mô hinh phân cụm, mô hình dạng cây, SVM và mạng nơ-ron nhân tạo.

A. Các mô hình học máy cổ điển

1) Lựa chọn mô hình: Ở bước này, có tổng cộng 7 mô hình được lựa chọn, thuộc các nhóm mô hình Linear - Logistic Regression, Bagging - Random Forest, XGBoost Random Forest, Probabilistic - Gaussian Naive Bayes, Boosting - CatBoost, XGBoost và LightGBM. Ngoại trừ Logistic Regression được đặt tham số *max_iter* = 2000, các mô hình khác không được cài đặt hay điều chỉnh bất kì siêu tham số nào nhằm so sánh nhanh hiệu quả của chúng. Mô hình Support Vector Machine không được chọn do không diễn giải được đóng góp của từng đặc trưng vào mô hình. Mỗi mô hình thực hiện kiểm chứng chéo với *n_split* = 10 và *n_repeats* = 5, tổng cộng 50 lần đánh giá.

Kết quả được thể hiện trong bảng I cho thấy mô hình Logistic Regression và Gaussian Naive Bayes cho hiệu quả khá tốt với chỉ số ROC-AUC trung bình tương ứng là 0.72 và 0.68. Đối với các mô hình còn lại, CatBoost cho kết quả tốt nhất là 0.66. Đồng thời, thực hiện t-test cho thấy có sự khác biệt giữa kết quả của các mô hình này, cho thấy rằng mô hình CatBoost có tiềm năng và hiệu quả tốt hơn. 2 mô hình được lựa chọn cuối cùng là Logistic Regression và CatBoost. Ngoài ra, có một mô hình cũng được lựa chọn để thử nghiệm là mô hình stacking với base learner là Logistic Regression và meta model là CatBoost.

Mô hình	Trung bình ROC-AUC	Độ lệch chuẩn
Logistic Regression	**0.7201**	0.0551
Random Forest	0.6100	0.0467
GaussianNB	0.6825	0.0463
XGBoost	0.6563	0.0481
XGBRF	0.6293	0.0499
CatBoost	0.6633	0.0535
LightGBM	0.6557	0.0549

Bảng I
KẾT QUẢ HUẤN LUYỆN NHANH CÁC MÔ HÌNH

2) Lựa chọn đặc trưng: Quá trình lựa chọn đặc trưng diễn ra trên nhiều bước thử nghiệm. Trong đó có một số đặc trưng được loại bỏ ở ngay sau bước EDA, bao gồm các biến chỉ có một giá trị như Over18, EmployeeCount và StandardHours, cùng với biến không ý nghĩa đối như EmployeeNumber (mã số nhân viên).

Ban đầu, cả 30 đặc trưng đều được sử dụng để huấn luyện và lựa chọn mô hình. Tiếp theo đó, từ mô hình CatBoost được chọn tiếp tục triển khai đồng thời (song song) 3 thuật toán khác nhau, bao gồm *RFE (Recursive Feature Elimination)*, Multicollinearity Elimination bằng VIF và Dimensionality Reduction bằng PCA. Mỗi thuật toán cho kết quả là số lượng và các đặc trưng khác nhau. Kết quả sau khi được tổng hợp sẽ tiến hành so sánh và bổ sung thêm một số đặc trưng để tiến hành huấn luyện.

Tiếp theo, các mô hình sau khi huấn luyện (trên cả 30 đặc trưng) sẽ tiến hành trích xuất độ quan trọng của từng đặc trưng và so sánh với nhau. Các đặc trưng ít quan trọng hoặc không có đóng góp sẽ được loại bỏ. Toàn bộ các đặc trưng còn lại được đưa vào huấn luyện.

Bước lựa chọn đặc trưng cuối cùng được thực hiện sau khi đánh giá hiệu suất của các mô hình (với số lượng đặc trưng huấn luyện khác nhau). Kết quả cho thấy 25 đặc trưng từ bước phân tích độ quan trọng cho hiệu suất tốt nhất ở tất cả các mô hình. Xếp hạng trung bình độ quan trọng của 10 đặc trưng quan trọng nhất và 5 đặc trưng ít quan trọng nhất được thể hiện trong bảng II.

Có thể thấy những đặc trưng ảnh hưởng nhiều nhất có sự thống nhất với phân tích ban đầu, bao gồm *độ tuổi*, *thu nhập hàng tháng*, *số năm làm việc* và *mức độ hài lòng với môi trường làm việc*. 5 đặc trưng bao gồm trình độ đào tạo, phòng ban, giới tính, ngành học và đánh giá hiệu suất có độ quan trọng rất thấp và hầu như không ảnh hưởng đến kết quả của mô hình, đây là các đặc trưng bị loại bỏ khỏi tập dữ liệu.

3) Điều chỉnh siêu tham số: Quá trình điều chỉnh siêu tham số được diễn ra một cách tự động thông qua framework Optuna.

Đặc trưng	RF	XGB	LGBM	CatB	LR	Hạng
EnvironmentSatisfaction	14	10	10	2	2	7.6
Age	2	7	5	13	15	8.4
TotalWorkingYears	3	2	15	19	7	9.2
JobInvolvement	16	6	17	5	3	9.4
MonthlyIncome	1	13	1	14	19	9.6
NumCompaniesWorked	10	12	9	10	10	10.0
DistanceFromHome	7	15	6	11	13	10.2
StockOptionLevel	21	4	20	4	1	10.4
WorkLifeBalance	17	14	16	1	4	10.4
YearsWithCurrManager	12	8	12	9	16	11.2
...	...	...	...	...	...	...
Education	25	23	22	23	20	22.6
Department	28	27	28	27	8	23.6
EducationField	23	28	25	25	23	24.8
Gender	29	29	27	29	17	26.2
PerformanceRating	30	30	30	30	11	26.2

Bảng II
XẾP HẠNG TRUNG BÌNH ĐỘ QUAN TRỌNG CỦA MỘT SỐ ĐẶC TRƯNG

Đối với mô hình CatBoost, có tổng cộng 14 siêu tham số được điều chỉnh, được thể hiện trong bảng III.

Tham số	Khoảng giá trị tìm kiếm	Giá trị tối ưu
iterations	[500, 2000], step=100	1500
learning_rate	[1e-3, 0.3] (log scale)	0.08445356
depth	[4, 10]	7
random_strength	[1, 20]	9
max_leaves	[20, 64]	57
min_data_in_leaf	[1, 20]	5
scale_pos_weight	[0.5, 5.0]	1.0
bootstrap_type	{Bayesian, Bernoulli, MVS }	Bayesian
l2_leaf_reg	[1e-3, 10.0] (log scale)	4.45969349
border_count	[32, 255]	108
od_type	{IncToDec, Iter}	IncToDec
od_wait	[10, 100]	53
bagging_temperature	[0.0, 1.0] (Bootstrap = Bayesian)	0.75091736
subsample	[0.5, 1.0] (Bootstrap = Bernoulli)	None

Bảng III
TỐI ƯU SIÊU THAM SỐ CHO MÔ HÌNH CATBOOST

Đối với mô hình Logistic Regression, có tổng cộng 4 siêu tham số được điều chỉnh, được thể hiện trong bảng IV.

Tham số	Khoảng giá trị tìm kiếm	Giá trị tối ưu
C	[1e-4, 1e-2] (log scale)	0.13555782
penalty	[l1, l2]	l1
solver	[liblinear, saga]	saga
max_iter	[1000, 2000] (step = 100)	1000

Bảng IV
TỐI ƯU SIÊU THAM SỐ CHO MÔ HÌNH LOGISTIC REGRESSION

Điều chỉnh siêu tham số của mô hình stacking về cơ bản giống với các mô hình thành phần nhưng bỏ đi các tham số grow_policy, bootstrap_type, od_type, od_wait và bagging_temperature. Kết quả được thể hiện trong bảng V.

GaussianNB là một mô hình xác suất đơn giản, không có nhiều siêu tham số để điều chỉnh, quá trình fine-tuning không đem lại cải thiện rõ rệt về hiệu suất. Do đó, mô hình GaussianNB được sử dụng với cấu hình mặc định.

B. Mô hình mạng học sâu

Mạng nơ-ron được xây dựng thủ công với hình thức là một mạng Fully Connected Multi-Layer Perceptron nhưng bổ sung các lớp Batch Normalization, Dropout, đồng thời sử dụng hàm kích hoạt LeakyReLU thay vì ReLU. Kiến trúc đầy đủ của mạng được thể hiện ở bảng VI với *input_dim* là số lượng đặc trưng được sử dụng từ bộ dữ liệu và *output_dim* là số lớp dự đoán.

Tham số	Mô hình	Giá trị tối ưu
C	Logistic Regression	0.05925733
penalty	Logistic Regression	l2
solver	Logistic Regression	liblinear
max_iter	Logistic Regression	1000
iterations	CatBoost	1700
learning_rate	CatBoost	0.02611771
depth	CatBoost	5
random_strength	CatBoost	14
max_leaves	CatBoost	27
min_data_in_leaf	CatBoost	9
scale_pos_weight	CatBoost	1.78224989
l2_leaf_reg	CatBoost	0.10917219
border_count	CatBoost	58

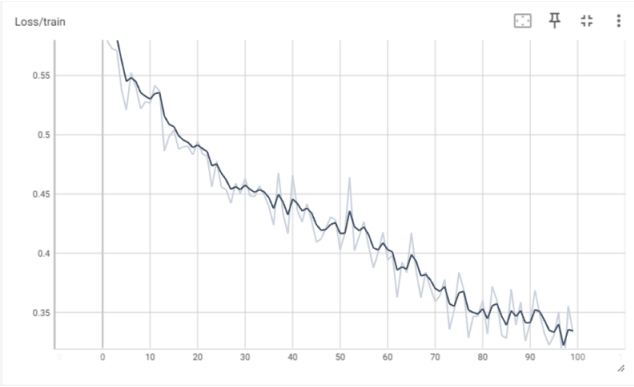
Bảng V
TỐI ƯU SIÊU THAM SỐ CHO MÔ HÌNH STACKING

STT	Lớp	Kích thước	Ghi chú
1	Linear	input_dim → 32	Lớp ẩn 1
	BatchNorm1d	32	Chuẩn hóa
	LeakyReLU(0.1)	32	Kích hoạt
2	Linear	32 → 64	Lớp ẩn 2
	BatchNorm1d	64	
	LeakyReLU(0.1)	64	
3	Linear	64 → 128	Lớp ẩn 3
	BatchNorm1d	128	
	LeakyReLU(0.1)	128	
	Dropout(0.3)	128	Giảm overfitting
4	Linear	128 → 256	Lớp ẩn 4
	BatchNorm1d	256	
	LeakyReLU(0.1)	256	
	Dropout(0.3)	256	
5	Linear	256 → 512	Lớp ẩn 5
	BatchNorm1d	512	
	LeakyReLU(0.1)	512	
	Dropout(0.3)	512	
6	Linear	512 → output_dim	Lớp đầu ra
	LogSoftmax(dim=1)	output_dim	Phân phối xác suất

Bảng VI
CẤU TRÚC MẠNG HỌC SÂU CHO BÀI TOÁN

Quá trình huấn luyện được thực hiện trên 100 epoch. Ở mỗi epoch đều ghi lại giá trị độ lỗi, độ chính xác và độ đo ROC-AUC vào tensorboard. Kết thúc quá trình huấn luyện, 3 bộ trọng số của mạng được lưu lại, bao gồm mạng có độ chính xác cao nhất, mạng có chỉ số ROC-AUC cao nhất và mạng sau epoch huấn luyện cuối cùng.

Đối với mạng học sâu, số lớp ẩn và kích thước các mạng đều là siêu tham số và được điều chỉnh liên tục. Hàm lỗi được sử dụng là Cross Entropy với trọng số lớp được đặt bằng 0.25 cho lớp 0 và 0.75 cho lớp 1. Optimizer được sử dụng là Adam với trọng số phân rã là 1e-4.



Hình 11. Giá trị độ lỗi được ghi lại trong quá trình huấn luyện

IV. Cài đặt thực nghiệm

A. Môi trường

Ngoại trừ bước triển khai mô hình, các phân tích và huấn luyện đều được thực hiện bằng Jupyter Notebook.

B. Thư viện

- numpy: Thao tác với dữ liệu số, chuẩn hóa, và tạo đặc trưng đầu vào.
- pandas: Thao tác với dữ liệu dạng bảng, xử lý, làm sạch, và phân tích dữ liệu.
- seaborn, matplotlib: Biểu diễn và trực quan hóa các phân tích trong dữ liệu cùng với các kết quả của mô hình.
- scikit-learn (sklearn): Sử dụng các công cụ tiền xử lý, lựa chọn đặc trưng, phân chia tập dữ liệu, và các thuật toán học máy cơ bản.
- imblearn: Sử dụng thuật toán ADASYN để xử lý mất cân bằng.
- optuna: Fine-tuning tự động trên khoảng giá trị tìm kiếm.
- torch (PyTorch): Xây dựng kiến trúc mạng và các bộ Dataset, DataLoader tương ứng cho mô hình học sâu.
- tensorboard: Theo dõi độ lỗi và các độ đo trong quá trình huấn luyện.
- xgboost, catboost, lightgbm: các thuật toán boosting hiện đại.
- pickle: Lưu lại tập dữ liệu sau khi đã xử lý
- joblib: Lưu mô hình cùng với các bộ encoder, scaler.

C. Độ đo

1) Độ chính xác: Là tỉ lệ số dự đoán đúng trên tổng số lượng dự đoán.

2) ROC-AUC: Chỉ số AUC là một độ đo quan trọng trong bài toán phân loại, đo lường chất lượng dự đoán của mô hình bất kể ngưỡng phân loại. Về cơ bản, nó định lượng sự cân bằng giữa True Positive Rate và False Positive Rate. Điểm AUC-ROC càng gần 1 thì mô hình càng phân biệt tốt hơn giữa các lớp tích cực và tiêu cực.

3) F1-Score: Đối với mục tiêu của bài toán, bên cạnh độ chính xác thì ta quan tâm nhiều hơn đến số lượng dự đoán đúng của lớp Positive (tức là dự đoán nhân viên nào sẽ nghỉ việc). Do đó các mô hình được huấn luyện với mục tiêu là cực đại hóa giá trị AUC.

V. Kết quả thực nghiệm

Kết quả thực nghiệm trên các mô hình được trình bày ở bảng VII. Có thể thấy, mô hình học sâu cho hiệu năng tốt nhất với giá trị cao nhất ở hai chỉ số ROC-AUC và F1-Score, chứng tỏ khả năng giải quyết rất tốt của các mạng nơ-ron nhân tạo kể cả đối với các bài toán cổ điển như phân loại.

Xét riêng nhóm các mô hình học máy cổ điển, mặc dù kết quả giữa các mô hình không có chênh lệch quá lớn nhưng mô hình Stacking đã cho thấy khả năng của nó trong việc kết hợp ưu điểm của các mô hình thành phần.

Mô hình	Số đặc trưng	Accuracy	ROC-AUC	F1-Score
CatBoost	30	0.8095	0.6746	0.4400
Logistic Regression	30	0.8095	0.6569	0.4167
GaussianNB	30	0.6803	0.6688	0.3896
Neural Network	30	0.8231	0.7181	0.5000
LR + CatBoost	30	0.8163	0.6786	0.4490
CatBoost	27	0.7891	0.6271	0.3673
Logistic Regression	27	0.6531	0.6173	0.3377
CatBoost	25	**0.8571**	0.6497	0.4324
Logistic Regression	25	0.8367	0.7085	0.5000
LR + CatBoost	25	0.8503	0.6988	0.5000
Neural Network	25	0.8163	**0.7495**	**0.5263**
CatBoost	24	0.8027	0.5643	0.2564
Logistic Regression	24	0.6735	0.6648	0.3846
CatBoost	21	0.8163	0.5724	0.2703
Logistic Regression	21	0.6735	0.6648	0.3846

Bảng VII
Kết quả thực nghiệm

VI. Hướng phát triển

Mặc dù các mô hình được tinh chỉnh nhiều và đạt được độ chính xác trên 80% nhưng chỉ số F1-Score vẫn chỉ ở mức trên 0.5, phản ánh rằng các mô hình chưa thực sự hoạt động tốt đối với các dữ liệu thuộc lớp Positive (ở đây là các nhân viên nghỉ việc). Có một số nguyên nhân mà tôi xin đề xuất, bao gồm:

- *Thiếu hụt dữ liệu*: Số lượng đặc trưng của tập dữ liệu khá lớn và các giá trị trải ở các khoảng khác nhau, trong khi chỉ có 1470 mẫu trong tập dữ liệu. Điều này khiến cho các giá trị của các đặc trưng bị phân tán, thiếu hụt giá trị hoặc chứa nhiều giá trị ngoại lai, khiến mô hình không học đủ tốt về đặc trưng đó.
- *Chênh lệch giữa 2 lớp mục tiêu*: Mặc dù tỉ lệ số lượng nhân viên nghỉ việc ở công ty ở mức trung bình (14%), nhưng do có quá ít dữ liệu về các nhân viên nghỉ việc khiến cho mô hình trở nên khó khăn khi phân biệt các nhân viên này với nhóm nhân viên ở lại.
- *Có đặc trưng gây nhiễu*: Mặc dù quá trình lựa chọn đặc trưng triển khai nhiều bước để giảm bớt các đặc trưng không cần thiết nhưng một số đặc trưng vẫn không đủ mạnh để mô hình có thể phân loại một cách rõ ràng, chẳng hạn như tình trạng hôn nhân.

Với những thực nghiệm và phân tích trên, báo cáo vạch ra những hướng phát triển có thể cho các công trình tiếp theo. Đầu tiên, dữ liệu có những vấn đề nhất định cần phải xem xét như

mất cân bằng, nhiều ngoại lai, số mẫu khá ít và dữ liệu này là dữ liệu giả tưởng. Việc có một bộ dữ liệu thật sự với nhiều mẫu dữ liệu hơn, thu nhập đồng đều ở nhiều nơi khác nhau sẽ giúp việc phân tích được thực tế và có ích hơn. Thứ hai, các mô hình được thực nghiệm trong bài báo chưa được chuyên biệt hoá và thiếu các đề xuất hay cơ sở nhắm đến nhân sự, do đó kết quả có thể chưa đủ tốt. Các công trình trong tương lai có thể tập trung hơn vào việc xây dựng các mô hình có tính chuyên biệt hoá cao, từ đó cho ra hiệu suất tốt hơn và có thể áp dụng trong thực tế.

VII. Kết luận

Bài báo cáo đã trình bày một hướng tiếp cận sử dụng các mô hình học máy và học sâu nhằm giải quyết bài toán dự đoán nhân viên nghỉ việc dựa trên dữ liệu nhân sự. Thông qua quá trình phân tích dữ liệu, tiền xử lý, lựa chọn đặc trưng và huấn luyện nhiều mô hình khác nhau, kết quả thực nghiệm cho thấy mô hình mạng nơ-ron nhân tạo đạt hiệu năng cao nhất, đặc biệt ở các chỉ số ROC-AUC và F1-Score. Điều này cho thấy tiềm năng của các mô hình học sâu trong việc xử lý các bài toán phân loại có tính chất phức tạp.

Mặc dù vậy, các chỉ số như F1-Score vẫn còn hạn chế, phản ánh sự khó khăn trong việc phân biệt lớp nhân viên nghỉ việc do dữ liệu không cân bằng và chất lượng đặc trưng chưa thực sự tối ưu. Từ đó, đồ án này đề xuất một số hướng phát triển như sử dụng bộ dữ liệu thực tế, tăng số lượng mẫu, xử lý mất cân bằng lớp và phát triển các mô hình chuyên biệt hoá hơn cho lĩnh vực nhân sự.

Tổng thể, đồ án này là bước đầu trong việc ứng dụng các mô hình trí tuệ nhân tạo vào phân tích nhân sự, mở ra nhiều tiềm năng để cải tiến và áp dụng vào thực tiễn trong tương lai.

Tài liệu tham khảo

[1] Việc Làm 24h. [*Báo cáo thị trường lao động 2023 – Toạ độ nhân tài*, 2023.]

[2] Lao Động. [*Gần 19.000 công chức, viên chức thôi việc trong 1 năm qua*, 2023.]

[3] Evidently AI. [*Explain ROC Curve and AUC Score*, 2023.]

[4] Viblo. [*Tự động điều chỉnh siêu tham số với Optuna và PyTorch*, 2023.]

[5] Analytics Vidhya. [*Feature Engineering: How to Detect and Remove Outliers with Python Code*, 2021.]

[6] Stack Overflow. [*Variance Inflation Factor in Python*, 2017.]

[7] Machine Learning Mastery. [*RFE Feature Selection in Python*, 2021.]

[8] MIAI. [*Principal Component Analysis (PCA) – Tuyệt chiêu giảm chiều dữ liệu*, 2021.]

[9] Alduayj, Sarah S. and Rajpoot, Kashif. *Predicting employee attrition using machine learning*, 2018. 2018 International Conference on Innovations in Information Technology (IIT). DOI: 10.1109/innovations.2018.8605976.