Ứng dụng học máy trong phân tích dữ liệu vào quản lý nguồn nhân lực

Applying machine learning in data analytics of human resource management

Nguyễn Phát Đạt[1,2], Nguyễn Văn Hồ[1,2], Thái Kim Phụng[3*]

[1]Trường Đại học Kinh tế - Luật, Thành phố Hồ Chí Minh, Việt Nam
[2]Đại học Quốc Gia Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam
[3]Trường Công nghệ và Thiết kế, Đại học Kinh tế Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: phungthk@ueh.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.econ.vi.19.9.3193.2024

Ngày nhận: 16/01/2024

Ngày nhận lại: 17/03/2024

Duyệt đăng: 21/03/2024

Mã phân loại JEL:

C61; C63; C67

Từ khóa:

HRM; học máy; nhân viên rời bỏ; quản lý nguồn nhân lực

Keywords:

HRM; machine learning; employee attrition; human resource management

TÓM TẮT

Quản lý nguồn nhân lực (Human Resource Management - HRM) đóng vai trò quan trọng trong sự thành công của mỗi doanh nghiệp thông qua việc quản lý hiệu quả lực lượng lao động, từ đó làm nền tảng giúp doanh nghiệp phát triển bền vững. Sự thành công của mỗi doanh nghiệp đều có sự đóng góp của các nhân sự ở mọi cấp bậc. Tuy nhiên, hiện trạng tại một số doanh nghiệp có tỷ lệ nhân viên nghỉ việc nhiều, gây ra những cản trở trong công việc và có thể ảnh hưởng đến hiệu quả kinh doanh. Vì vậy, việc giữ chân nhân sự đóng vai trò quan trọng bởi quản lý tốt sẽ giúp nâng cao hiệu quả hoạt động của doanh nghiệp. Nghiên cứu này xây dựng phân tích dự báo nhân viên nghỉ việc trên tập dữ liệu nhân sự của IBM. Tác giả tiến hành thực nghiệm mô hình máy học để dự báo nhân viên nghỉ việc qua các thuật toán Logistics Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine, Neural Network và Random Forest để tìm ra mô hình tối ưu. Thông qua kết quả thực nghiệm, các tổ chức có thể sử dụng những kết quả này để xây dựng chiến lược HRM mang lại nhiều ý nghĩa cho doanh nghiệp.

ABSTRACT

Human Resource Management (HRM) plays a crucial role in achieving organizational success by effectively managing the workforce. Every business success has numerous contributions from employees at all levels. However, this becomes an intense dilemma when they leave, which leads to business delays and lower performance. Therefore, employee retention management plays a vital role, which, if well-controlled can enhance the business performance. This research suggests an employee attrition prediction model as well as reports to have an overall view of IBM's HR dataset. The authors proposed machine learning models to predict employees who left the company: Logistics Regression, K-nearest Neighbors, Decision Tree, Support Vector Machine, Neural Network, and Random Forest. In addition, dashboard reports are also created to support an executive view for business decision-making. By implementing the proposed models and building dashboards, organizations can make use of valuable output to drive suitable strategic HRM decisions and gain meaningful results for business.

1. Giới thiệu

Nguồn nhân lực có vai trò quyết định tới năng lực cạnh tranh, kết quả kinh doanh của doanh nghiệp và là tài sản quan trọng nhất, là cơ sở nền tảng phát triển và tồn tại của mỗi doanh nghiệp (Tran, 2015). Mỗi nhân viên dù đảm việc chức vụ lớn hay nhỏ nhưng vẫn đóng góp vào sự thành công chung của doanh nghiệp. Quản lý nguồn nhân lực (HRM) đóng vai trò quan trọng trong tổ chức vì nó chịu trách nhiệm giám sát tài nguyên quý giá nhất của doanh nghiệp - đó là lực lượng lao động. HRM có mối quan hệ mật thiết, ảnh hưởng to lớn đến sự thành công của một tổ chức hay doanh nghiệp (Guest, 1997). Bất kỳ tổ chức hoặc công ty nào cũng nhận thức rõ tầm quan trọng của nhân viên trong việc đạt được và duy trì năng lực lợi thế cạnh tranh (Dutta, Bandyopadhyay, & Bandyopadhyay, 2020). Trong môi trường kinh doanh hiện đại nhanh chóng và nhiều biến động, quản lý nhân sự hiệu quả ngày càng trở nên quan trọng hơn bao giờ hết. Sự thành công của một tổ chức phụ thuộc rất nhiều vào khả năng thu hút, giữ chân và phát triển một lực lượng lao động tài năng và cam kết làm việc lâu dài. Thế nên, nếu nhân viên rời bỏ hay nghỉ việc không chỉ làm công ty mất đi một nhân viên mà còn dẫn đến mất đi khách hàng của doanh nghiệp (Agarwal, 2013), điều này gây ảnh hưởng đến hoạt động kinh doanh sản xuất cũng như sự phát triển của các doanh nghiệp (Davidescu, Apostu, Paul, & Casuneanu, 2020). Phần lớn các doanh nghiệp đều không muốn một nhân viên làm việc nhiều năm hoặc một nhân viên mới gia nhập vào công ty nộp đơn thôi việc, bởi sẽ tốn nhiều chi phí để tuyển dụng thay thế hoặc tốn nhiều chi phí và thời gian để đào tạo một nhân viên mới. Sự nghỉ việc của nhân viên là một quá trình bình thường (Raza, Munir, Almutairi, Younas, & Fareed, 2022) bởi mỗi người nhân viên nghỉ việc điều có những lý do riêng của họ, có thể kể đến như: thu nhập, môi trường, thăng tiến, gia đình. Vì vậy, việc dự báo liệu rằng nhân viên có khả năng nghỉ việc hay không có vai trò quan trọng trong việc quản lý nhân sự ở bất kỳ doanh nghiệp nào, nó không chỉ ảnh hưởng tới việc quản trị và phát triển con người mà còn ảnh hưởng trực tiếp đến hoạt động kinh doanh của công ty nếu như nhiều nhân viên nghỉ việc trong một khoảng thời gian. Nếu có một mô hình dự báo tốt sẽ giúp doanh nghiệp hạn chế được việc này; bên cạnh đó, còn giúp bộ phận quản lý nhân sự, các nhà quản lý nắm được đặc điểm chung của nhân sự nghỉ việc để từ đó cải thiện các phúc lợi, môi trường làm việc, nâng cao sự trung thành, gắn bó của nhân viên tại công ty.

Từ những lý do trên, bài nghiên cứu thực hiện xây dựng mô hình học máy dự báo nhân sự nghỉ việc để hạn chế sự rời bỏ của những nhân viên tốt, đồng thời giúp nhà quản trị có những chiến lược quản lý và phát triển con người phù hợp. Bài nghiên cứu bao gồm các nội dung: (1) giới thiệu, (2) tổng quan tình hình nghiên cứu, (3) phương pháp thực nghiệm, (4) thu nhập dữ liệu và thực nghiệm mô hình, (5) tổng kết.

2. Tổng quan tình hình nghiên cứu

Nguồn nhân sự trên thế giới đang không ngừng bàn luận về dữ liệu lớn và tiềm năng biến đổi của phân tích nhân sự (Angrave, Charlwood, Kirkpatrick, Lawrence, & Stuart, 2016). Không chỉ thế, phân tích trong quản lý nguồn nhân lực đã có từ nhiều năm (Marler & Boudreau, 2017), trong những năm gần đây việc áp dụng phân tích dữ liệu trong quản lý nhân sự đã thu hút sự chú ý đáng kể, mang đến cho các chuyên gia quản lý nhân sự những hiểu biết quý giá về xu hướng nhân sự, hành vi của nhân viên và các yếu tố ảnh hưởng đến sự cam kết, sự hài lòng và năng suất của nhân viên. Kết hợp phân tích dữ liệu vào quản lý nhân sự có tiềm năng thay đổi cách tiếp cận ra quyết định của nhà quản lý bằng cách cung cấp một thông tin dựa trên dữ liệu để đưa ra quyết định chiến lược để tăng cường nguồn lực bền vững của tổ chức. Tận dụng phân tích dữ liệu trong quản lý nhân sự không chỉ cho phép tổ chức nhận ra đặc điểm các mẫu và xu hướng trong dữ liệu nhân sự, từ đó đưa ra quyết định thông qua việc tuyển dụng nhân sự, giữ chân nhân tài mà còn mang lại lợi thế cạnh tranh trên thị trường và đẩy mạnh sự thành công kinh doanh.

Đặc biệt, việc giữ chân nhân viên trở thành một trong những chiến lược hàng đầu, bởi sự ra đi của họ có thể gây ra những tác động tiêu cực đáng kể đến lợi thế cạnh tranh và có thể gây tốn kém nhiều chi phí cho tổ chức. Chi phí của sự ra đi nhân viên sẽ bao gồm chi phí liên quan đến vòng đời của nguồn nhân lực, sự mất tri thức, tinh thần nhân viên và văn hóa tổ chức (Setiawan, Suprihanto, Nugraha, & Hutahaean, 2020). Do đó, các tổ chức cần phát triển một khung công việc toàn diện cho phân tích dữ liệu nhân sự để đối phó với những thách thức và hạn chế này. Nếu như giải quyết được vấn đề đó, doanh nghiệp sẽ tiết kiệm được lượng lớn tài nguyên phải bỏ ra, tối ưu hóa kết quả kinh doanh (Hinkin & Tracey, 2000). Dự báo rời bỏ nói chung hay dự báo nhân viên rời bỏ nói riêng luôn đóng vai trò quan trọng do đối sự hình thành và phát triển của các doanh nghiệp.

Cùng với sự phát triển và ứng dụng của phân tích dữ liệu thì sự bùng nổ của Học máy (Machine Learning), đặc biệt là trong những năm gần đây là không thể phủ nhận. Các thuật toán và mô hình máy học ngày càng phát triển và được tối ưu hóa, ứng dụng trong nhiều mảng quản lý doanh nghiệp, trong đó có quản lý nguồn nhân lực. Nhiều nghiên cứu phân tích và dự đoán sự rời bỏ của nhân sự đã được xây dựng dựa trên các thuật toán khác thuật toán: Random Forest (RF) (Marvin, Jackson, & Alam, 2021), Support Vector Machine (SVM), K-nearest Neighbor (KNN), Decision Tree (DT) (Kaur & Dogra, 2022). Trong một nghiên cứu của Kamath, Jamsandekar, và Naik (2019), tác giả đã sử dụng thuật toán DT, RF, SVM, LG để phân tích tại sao một số nhân viên xuất sắc và có kinh nghiệm rời bỏ công ty sớm và cũng dự đoán những nhân viên có giá trị sẽ rời đi tiếp theo. Tác giả Ray và Sanyal (2019) sử dụng các mô hình học máy để đánh giá mối quan hệ giữa tuổi tác và bằng cấp tác động như thế nào đối với sự rời bỏ của nhân viên. Tác giả Raza và cộng sự (2022) đã sử dụng các thuật toán như SVM, LG, DT và Extra Tree Calssifier để dự báo khả năng rời bỏ, từ đó chỉ ra rằng các yếu tố tác động đến quyết định của nhân viên là thu nhập, tiền công mỗi giờ, cấp bậc công việc và tuổi. Các nghiên cứu cho kết quả dự báo tốt, giúp doanh nghiệp chủ động hơn trong việc quản lý và phát triển nguồn nhân lực, đảm bảo hoạt động kinh doanh được diễn ra thông suốt và phát triển bền vững. Tuy nhiên, mặc dù các nghiên cứu này đã chỉ ra các yếu tố ảnh hưởng đến quyết định nghỉ việc của nhân viên, nhưng chúng chưa đi sâu vào việc phân tích xu hướng thay đổi cũng như tác động của các yếu tố và tỷ lệ nghỉ việc đối với nhu cầu lao động trong tương lai.

Bài báo này đề xuất mô hình dự đoán thông qua việc xử lý mất cân bằng dữ liệu và chọn lựa đặc trưng cùng với các thuật toán học máy để dự đoán khả năng nghỉ việc của nhân viên. Bằng cách tận dụng kết quả phân tích dự đoán, tổ chức có thể phát triển các chiến lược nhân sự chủ động dự đoán nhu cầu lực lượng lao động trong tương lai và giảm thiểu nguy cơ tiềm tàng. Mô hình đề xuất nhằm cung cấp thông tin một cách bao quát và toàn diện từ dữ liệu nhân sự, tổ chức có thể sử dụng để phát triển các chiến lược nhân sự hiệu quả và nâng cao hiệu suất tổ chức tổng thể. Mô hình có thể được tùy chỉnh để đáp ứng nhu cầu cụ thể của mỗi tổ chức và có thể được áp dụng để giải quyết một loạt các thách thức nhân sự, chẳng hạn như tuyển dụng, giữ chân, cam kết của nhân viên và quản lý hiệu suất.

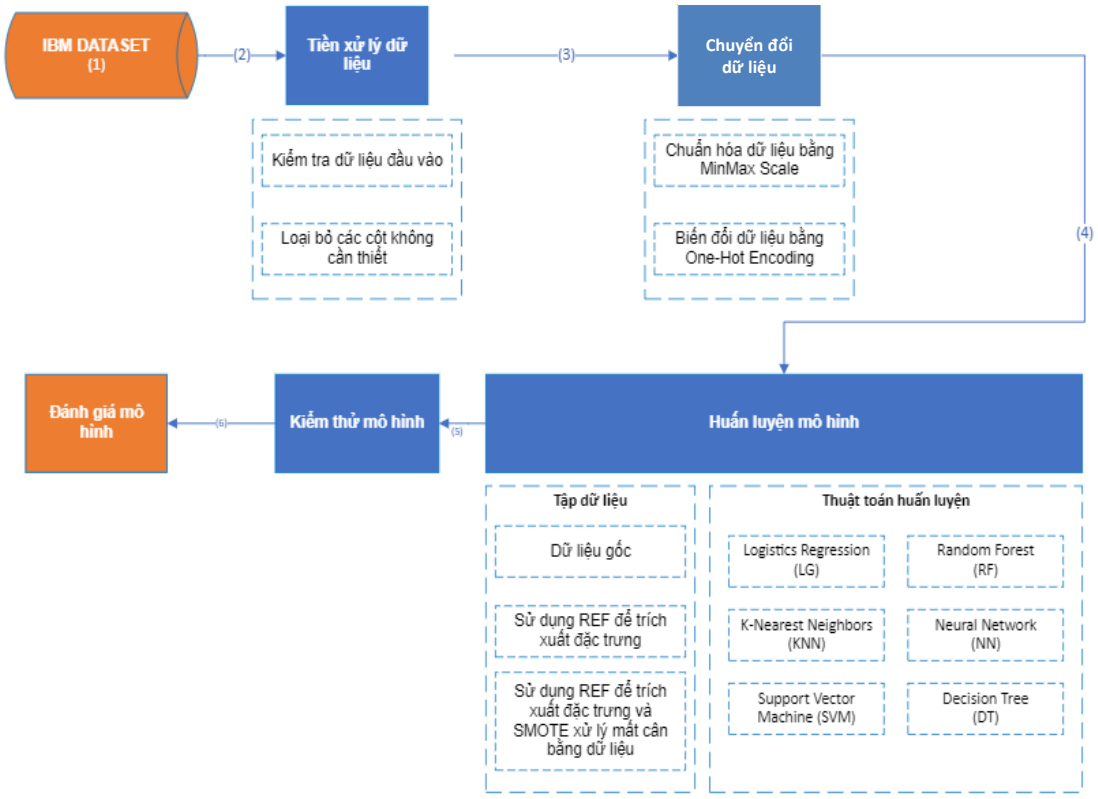
Tóm lại, việc áp dụng phân tích dữ liệu trong quản lý nhân sự có tiềm năng thay đổi cách tổ chức tiếp cận các chiến lược nhân sự. Bằng cách cung cấp những thông tin có giá trị mang nhiều ý nghĩa về xu hướng lực lượng lao động, hành vi của nhân viên và các yếu tố ảnh hưởng đến cam kết, sự hài lòng và năng suất của nhân viên, phân tích dữ liệu nhân sự giúp tổ chức đưa ra quyết định thông minh, phù hợp và nâng cao hiệu suất tổ chức tổng thể.

3. Phương pháp thực nghiệm

Nghiên cứu này được thực hiện dựa trên phương pháp nghiên cứu định tính và thực nghiệm. Trong đó, phương pháp định tính được sử dụng để khảo sát và tìm hiểu các nghiên cứu

thứ cấp, các công trình nghiên cứu đã được công bố về việc ứng dụng máy học và phân tích dữ liệu vào lĩnh vực quản lý nguồn nhân lực, từ đó, tìm ra các khoảng trống nghiên cứu để thực hiện nâng cao, cải thiện hiệu suất và xây dựng mô hình thực nghiệm phù hợp. Phương pháp thực nghiệm tiến hành thu thập và phân tích mô tả dữ liệu và xây dựng mô hình bằng các phương pháp máy học. Sau đó, tiến hành đánh giá các kết quả thực nghiệm để tìm mô hình dự báo phù hợp.

Nhóm tác giả đề xuất quy trình nghiên cứu gồm 06 bước như Hình 1 bên dưới:



Hình 1. Quy trình thực nghiệm

Nguồn: Tác giả

Bước 1: Thu thập dữ liệu

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 xử lý thiết yếu và quan trọng trong phân tích dữ liệu, có nhiều ảnh hưởng đến chất lượng của mô hình dự báo. Hơn thế nữa, tiền xử lý dữ liệu còn cung cấp cái nhìn tổng quan hơn tập dữ liệu đang thực nghiệm. Nghiên cứu này sẽ tiến hành một số bước xử lý trong quá trình này như kiểm tra kiểu dữ liệu, xác định giá trị bị null và xử lý giá trị null, loại bỏ những thuộc tính không cần thiết.

Bước 3: Chuyển đổi dữ liệu

Tác giả tiến hành chuyển đổi dữ liệu bao gồm: chuyển đổi các cột dữ liệu không có kiểu dữ liệu là "number", ở đây tác giả sử dụng One-Hot Encoding để biến đổi các biến này thành có biến giả dummy; sau đó, chuẩn hóa dữ liệu bằng phương pháp MinMax Scale.

Bước 4: Huấn luyện mô hình

Tác giả tiến hành thực nghiệm lần lượt 03 tập dữ liệu bao gồm: dữ liệu gốc, dữ liệu sau khi sử dụng Recursive Feature Elimination - REF để lựa chọn đặc trưng, dữ liệu sau khi sử dụng REF để lựa chọn đặc trưng và SMOTE - Synthetic Minority Oversampling Technique là xử lý mất cân bằng dữ liệu. Lựa chọn đặc trưng là quá trình chọn lọc những đặc trưng có ảnh hưởng đến biến phụ thuộc, loại bỏ các biến độc lập không quan trọng hoặc có sự tương quan với nhau. Nghiên cứu này sử dụng phương pháp RFE để chọn lọc trích xuất đặc trưng tham gia xây dựng mô hình ở giai đoạn tiếp theo.

Nghiên cứu thực nghiệm 06 mô hình máy học bao gồm: Logistic Regression, Random Forest, K-Nearest Neighbors, Neural Network, Decision Tree và Support Vector Machine trên từng tập dữ liệu đã đề cập ở trên.

Bước 5 - 6: Đánh giá mô hình

Tác giả chạy kiểm thử kết quả mô hình và tiến hành so sánh hiệu quả dự báo, từ đó lựa chọn mô hình tốt nhất thông qua các chỉ số như Accuracy, Recall, F1-score và Precision.

4. Thu thập dữ liệu và thực nghiệm mô hình

4.1. Tập dữ liệu

Nghiên cứu sử dụng bộ dữ liệu về lực lượng lao động của công ty IBM được chia sẻ trên trang website Kaggle (Pavansubhasht, 2017). Bộ dữ liệu này tập trung vào các yếu tố ảnh hưởng đến quyết định nghỉ việc của nhân viên và hiệu suất làm việc của họ. Bộ dữ liệu được công ty IBM tạo ra, bao gồm các trường thông tin như tuổi, giới tính, trình độ học vấn, lĩnh vực làm việc, thâm niên, số lần thăng chức, mức lương, đánh giá hiệu suất, và thông tin về việc nghỉ việc của nhân viên. Mỗi dòng trong bộ dữ liệu đại diện cho một nhân viên, bao gồm 1,470 quan sát với 35 cột như thể như Bảng 1 dưới đây:

Bảng 1

Các cột trong tập dữ liệu IBM

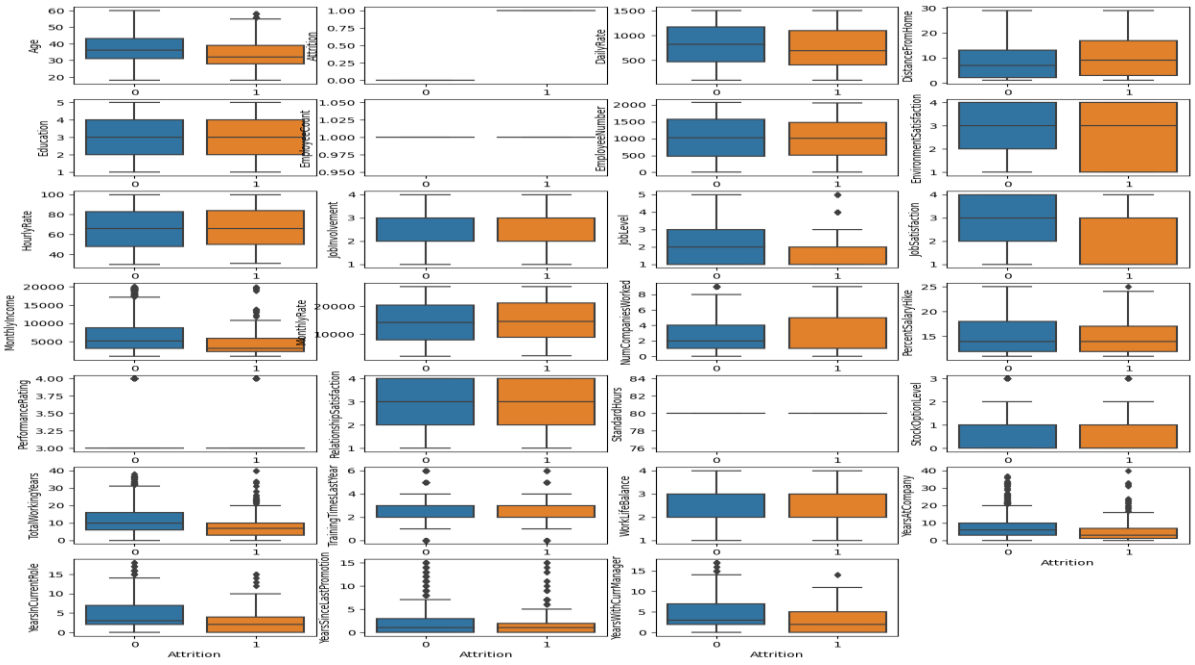
Số thứ tự	Cột dữ liệu	Kiểu dữ liệu	Diễn giải
1	Age	int64	Độ tuổi
2	Attrition	object	Có nghỉ việc hay không (0 là không nghỉ, 1 là có nghỉ việc)
3	BusinessTravel	object	Tuần suất nhân viên đi công tác
4	DailyRate	int64	Tỷ lệ thanh toán hằng ngày cho nhân viên
5	Department	object	Phòng Ban
6	DistanceFromHome	int64	Khoảng cách từ nhà đến công ty
7	Education	int64	Bậc giáo dục cao nhất
8	EducationField	object	Chuyên ngành đào tạo
9	EmployeeCount	int64	
10	EmployeeNumber	int64	Mã nhân viên

Số thứ tự	Cột dữ liệu	Kiểu dữ liệu	Diễn giải
11	EnvironmentSatisfaction	int64	Độ hài lòng về môi trường làm việc
12	Gender	object	Giới tính (Nam hoặc Nữ)
13	HourlyRate	int64	Tỷ lệ thanh toán hằng giờ cho nhân viên
14	JobInvolvement	int64	Mức độ tham gia vào công việc hiện tại
15	JobLevel	int64	Cấp bậc công việc
16	JobRole	object	Chức vụ
17	JobSatisfaction	int64	Độ hài lòng về công việc
18	MaritalStatus	object	Tình trạng hôn nhân
19	MonthlyIncome	int64	Thu nhập hằng tháng
20	MonthlyRate	int64	Tỷ lệ thanh toán hằng tháng cho nhân viên
21	NumCompaniesWorked	int64	Số lượng công ty đã làm việc trước đó
22	Over18	object	Nhân viên đã trên 18 tuổi
23	OverTime	object	Làm việc tăng ca
24	PercentSalaryHike	int64	Tỷ lệ phần trăm trăm tăng lương
25	PerformanceRating	int64	Đánh giá hiệu suất công việc
26	RelationshipSatisfaction	int64	Độ hài lòng về mối quan hệ
27	StandardHours	int64	Số giờ chuẩn làm việc
28	StockOptionLevel	int64	Nhân viên giữ quyền chọn cổ phiếu
29	TotalWorkingYears	int64	Tổng số năm kinh nghiệm của nhân viên
30	TrainingTimesLastYear	int64	Thời gian đào tạo năm ngoái
31	WorkLifeBalance	int64	Cân bằng giữa công việc và cuộc sống
32	YearsAtCompany	int64	Số năm làm việc tại công ty
33	YearsInCurrentRole	int64	Số năm làm việc ở vị trí hiện tại
34	YearsSinceLastPromotion	int64	Số năm kể từ lần thăng chức trước đó
35	YearsWithCurrManager	int64	Số năm làm việc với quản lý hiện tại

Nguồn: Tác giả

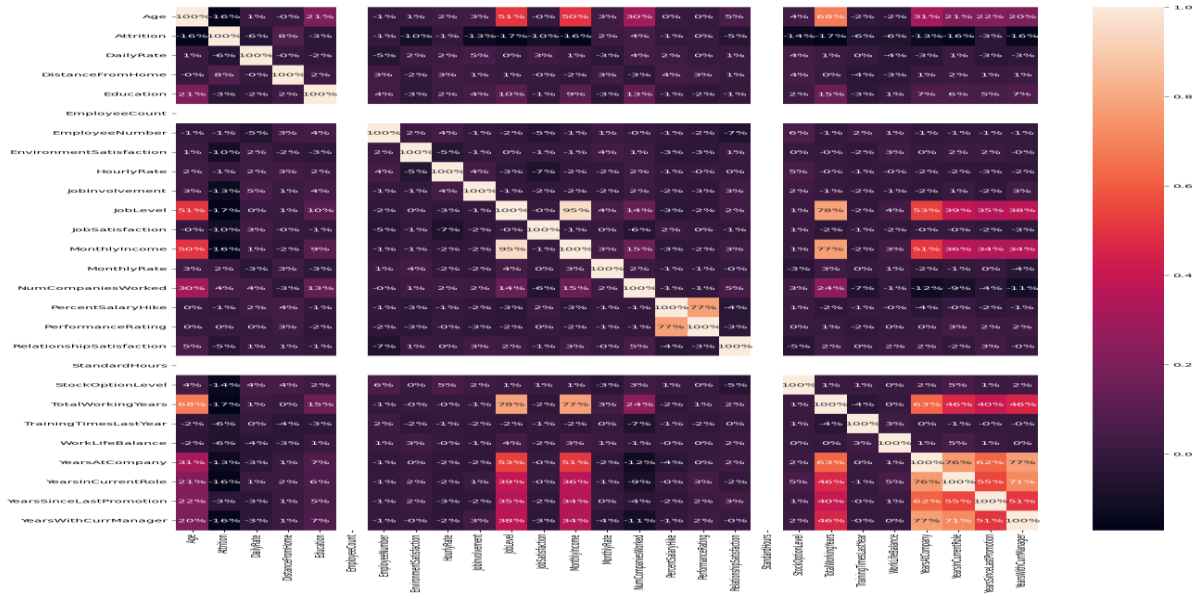
4.2. Tiền xử lý dữ liệu

Tác giả sử dụng biểu đồ box-plot để biểu diễn mối tương quan giữa các biến độc lập có kiểu dữ liệu số và biến phụ thuộc. Biểu đồ box-plot ở Hình 2 hiển thị số lượng nhân viên đã rời công ty. Điều này giúp cung cấp một biểu diễn trực quan về mối quan hệ giữa các biến này và cho phép hiểu rõ hơn về dữ liệu. Kết quả là, một số biến có ít hoặc không có mối tương quan với biến Attrition vì chúng chỉ có một giá trị duy nhất. Hơn thế nữa, bản đồ heatmap trong Hình 3 chỉ ra rằng không có mối tương quan nào giữa biến Attrition và các biến EmployeeCount, StandardHours, vì vậy cần phải loại bỏ hai thuộc tính này.



Hình 2. Biểu đồ box-plot tương quan giữa biến độc lập và phụ thuộc
(Nhân viên nghỉ việc là 0, còn làm việc là 1)

Nguồn: Tác giả



Hình 3. Biểu đồ heatmap biểu diễn sự tương quan giữa biến độc lập và phụ thuộc

Nguồn: Tác giả

Trong tập dữ liệu thực nghiệm, một số biến có kiểu dữ liệu là kiểu chữ cần phải chuyển đổi thành dạng số, ở đây sử dụng phương pháp One-Hot Endcoding để chuẩn hóa cột dữ liệu này. One-hot Encoding là phương pháp tạo ra tập dữ liệu mới, trong đó mỗi cột mới được tạo ra tương ứng với một giá trị riêng biệt, cột mới này sẽ có giá trị là 1 khi xuất hiện và 0 nếu không xuất hiện (Yu, Zhou, Chen, & Lai, 2022). Các cột dữ liệu cần chuyển đổi đó là ‘BusinessTravel’, ‘Department’, ‘EducationField’, ‘Gender’, ‘JobRole’, ‘MaritalStatus’.

Khi thực nghiệm mô hình dự báo bằng phương pháp học máy, khi sử dụng nhiều biến tham gia vào mô hình có thể sẽ làm tăng sự phức tạp của tính toán, và chi phí triển khai, ngoài

ra, không phải toàn bộ các biến đầu vào đều quan trọng trong việc xây dựng mô hình. Một số biến ít có tác động, thậm chí gây nhiễu, ảnh hưởng không tốt đến kết quả cuối cùng. Một số thuật toán được sử dụng để lựa chọn đặc trưng như: tương quan Pearson, Light XGB (Wang & Ni, 2019). Trong bài nghiên cứu, nhóm tác giả sử dụng phương pháp Recursive Feature Elimination (RFE) là một thuật toán hiệu quả để lựa chọn các biến đầu vào cho mô hình (Brownlee, 2020).

4.3. Thực nghiệm mô hình

Dự đoán nhân viên rời bỏ là bài toán xác định nhân viên có nghỉ việc ở công ty hay không, kết quả trả về là Có - Positive (giá trị 1) hoặc Không - Negative (giá trị 0). Với tập dữ liệu thực nghiệm đã xác định nhân viên nào nghỉ việc, vì vậy dự đoán nhân viên rời bỏ này thuộc học máy có giám sát. Theo Tatsat và Lookabaugh (2020), một số phổ biến trong phân loại học máy có giám sát là Logistic regression, K-nearest neighbors, Decision Tree, Support vector machine, Ensemble boosting, Ensemble bagging, Artificial neural network. Trong nghiên cứu này, tác giả sử dụng 06 mô hình để thực nghiệm bao gồm: Logistics Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine, Neural Network và Random Forest.

Logistic Regression (LG) là thuật toán phổ biến và được sử dụng rộng rãi trong bài toán phân loại (Tatsat và Lookabaugh, 2020). Kết quả đầu ra của mô hình là xác suất được tính toán theo hàm hồi quy tuyến tính theo x, trong đó biến y năm trong khoảng giá trị từ 0 đến 1.

K-Nearest Neighbors (KNN) được xem như là thuật toán học máy đơn giản nhất, việc xây dựng mô hình chỉ bao gồm dữ liệu trên tập huấn luyện. Để dự đoán cho quan sát mới, thuật toán tìm những điểm dữ liệu gần trong tập dữ liệu huấn luyện, vì vậy thuật toán này được gọi là "láng giềng gần nhất" (Müller & Guido, 2016).

Decision Tree (DT) là phương pháp học máy kết hợp, trong quá trình huấn luyện, mô hình xây dựng nhiều nhánh "cây quyết định" tại cùng thời điểm và cho kết quả phân loại của từng nhánh cây.

Support Vector Machine (SVM) là thuật toán có thể dùng để phân loại cho cả dữ liệu tuyến tính và phi tuyến tính (Müller & Guido, 2016), được ứng dụng trong một số lĩnh vực như phân loại, nhận dạng chữ viết tay, nhận dạng đối tượng hoặc dữ liệu chuỗi thời gian (Han, Pei, & Tong, 2022).

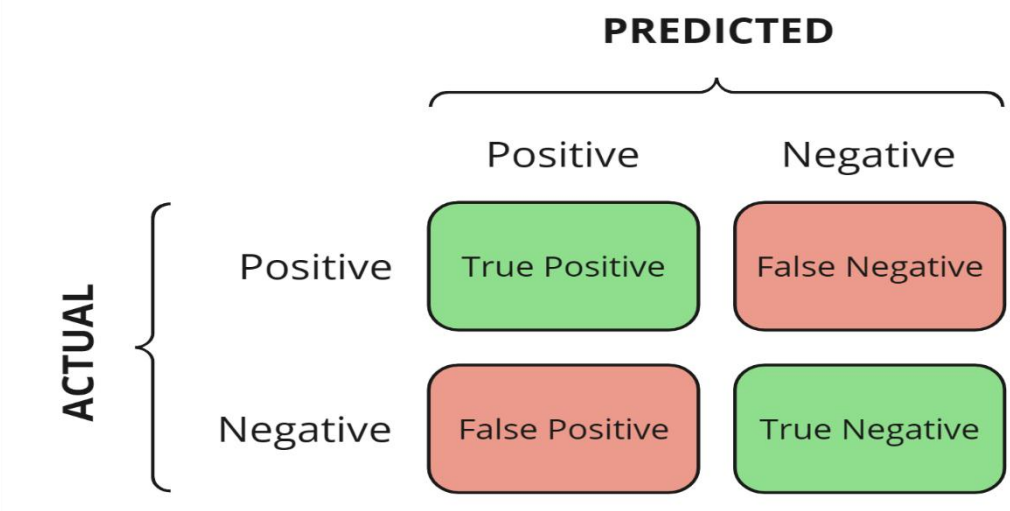
Mạng nơ-ron (Neural Network - NN) còn được gọi là Mạng thần kinh nhân tạo (Artificial Neural Network) là mô hình dự báo được mô tả hoạt động giống như bộ não của con người. Mạng nơ-ron bao gồm các nơ-ron nhân tạo, thực hiện tính toán dựa trên các dữ liệu đầu vào, được sử dụng nhiều để giải quyết các bài toán nhận dạng chữ viết tay, nhận diện khuôn mặt và được sử dụng nhiều trong học sâu (Grus, 2019).

Random Forest (RF) là một trong những phương pháp học kết hợp, mỗi một nhánh riêng lẻ được tạo ra bằng cách chọn ngẫu nhiên một thuộc tính tại mỗi nút để xác định sự phân chia (Agarwal, 2013). Trong quá trình phân loại, mỗi nhánh cây sẽ đưa ra kết quả dự đoán riêng, và kết quả nào dự báo chiếm phần lớn sẽ là kết quả cuối tuần của mô hình.

5. Kết quả nghiên cứu và thảo luận

Thông qua việc nghiên cứu và phân tích các công trình nghiên cứu và khảo sát các nghiên cứu thứ cấp, tác giả nhận thấy việc phân tích dự báo nhân viên nghỉ việc gặp khó khăn khi áp dụng các mô hình đối với bộ dữ liệu bị mất cân bằng. Từ đó, nhóm tác giả tiến hành thực nghiệm các mô hình để dự báo với 03 trường hợp như sau: sử dụng phương pháp máy học, sử dụng kết hợp RFE và phương pháp máy học, sử dụng RFE kết hợp SMOTE và phương pháp máy học để có thể so sánh, đánh giá kết quả đạt được.

Sau khi thực nghiệm các mô hình, tác giả sử dụng Ma trận nhầm lẫn (Confusion Matrix) để tìm ra mô hình phù hợp nhất để dự báo khả năng nghỉ việc của nhân viên.



Hình 4. Các chỉ số của Ma trận nhầm lẫn

Nguồn: Tác giả

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, cung cấp những thông tin hữu ích về điểm đúng và điểm bị lỗi về mô hình (Deng, Liu, Deng, & Mahadevan, 2016). Ma trận nhầm lẫn có cấu trúc dạng bảng, với 04 chỉ số đối với mỗi lớp phân loại. Trong khuôn khổ đề tài nghiên cứu, 04 chỉ số mang ý nghĩa như sau:

- 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

Từ 04 chỉ số trên, ta xác định được các chỉ số đánh giá quan trọng:

- **Accuracy**: là độ chính xác của mô hình. Độ chính xác càng cao càng tốt cho mô hình, tuy nhiên không phải trong tất cả các trường hợp thì độ chính xác càng cao thì chất lượng mô hình càng tốt.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP + TN + FP + FN}} \quad (1)$$

- **Precision**: Precision tỷ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive. Chỉ số này càng cao, số điểm mô hình dự đoán là Positive đều là Positive càng nhiều.

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \quad (2)$$

- **Recall**: Recall là tỷ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive. Recall càng cao, số điểm là positive bị bỏ sót càng ít; Recall = 1 tức là tất cả số điểm có nhãn là Positive đều được mô hình nhận ra.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \quad (3)$$

- **F1-Score**: F1-score là trung bình điều hòa của precision và recall.

$$\text{F1}_{\text{score}} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision + Recall}} \quad (4)$$

Bảng 2, Bảng 3 và Bảng 4 là kết quả so sánh về độ chính xác giữa các mô hình dựa trên accuracy, precision, recall và F1-Score khi thực nghiệm lần lượt các mô hình với tập dữ liệu gốc, tập dữ liệu sau khi chọn lựa đặc trưng bằng phương pháp RFE, tập dữ liệu sau khi chọn lựa đặc trưng và xử lý mất cân bằng dữ liệu bằng phương pháp RFE và SMOTE.

Bảng 2

Kết quả thực nghiệm từ dữ liệu gốc

Số thứ tự	***Model***	**Kết quả model từ dữ liệu gốc**			
		Accuracy	***Precision***	***Recall***	***F1-Score***
1	Logistic Regression	0.897	0.833	0.431	0.568
2	K-Nearest Neighbors	0.848	0.55	0.19	0.282
3	Decision Tree	0.802	0.358	0.328	0.342
4	SVM (Linear Kernel)	0.883	0.759	0.379	0.506
5	Neural Network	0.856	0.571	0.345	0.43
6	Random Forest	0.872	0.867	0.224	0.356

Nguồn: Tác giả

Bảng 3

Kết quả thực nghiệm sau khi áp dụng REF

Số thứ tự	***Model***	**Kết quả model sau khi sử dụng REF**			
		Accuracy	***Precision***	***Recall***	***F1-Score***
1	Logistic Regression	0.889	0.815	0.379	0.518
2	K-Nearest Neighbors	0.859	0.65	0.224	0.333
3	Decision Tree	0.777	0.3	0.31	0.305
4	SVM (Linear Kernel)	0.886	0.833	0.345	0.488
5	Neural Network	0.872	0.657	0.397	0.495
6	Random Forest	0.872	0.762	0.276	0.405

Nguồn: Tác giả

Bảng 4

Kết quả thực nghiệm sau khi áp dụng REF + SMOTE

Số thứ tự	***Model***	**Kết quả model sau khi sử dụng REF + SMOTE**			
		Accuracy	***Precision***	***Recall***	***F1-Score***
1	***Logistic Regression***	0.912	0.934	0.893	0.913
2	***K-Nearest Neighbors***	0.896	0.902	0.896	0.899
3	***Decision Tree***	0.838	0.852	0.83	0.841
4	***SVM (Linear Kernel)***	0.917	0.953	0.884	0.917
5	***Neural Network***	0.911	0.931	0.893	0.912
6	***Random Forest***	0.922	0.969	0.877	0.921

Nguồn: Tác giả

Đối với nguồn dữ liệu ban đầu, từ Bảng 2 có thể nhận thấy chỉ số Accuracy khá cao (các mô hình đều trên 80%), tuy nhiên chỉ số F1-score khá thấp dao động từ 0.28 đến 0.57. Do bộ dữ liệu ban đầu đang bị mất cân bằng giữa tỷ lệ nghỉ việc nên kết quả về độ chính xác mô hình với chỉ số F1_score chỉ nằm khoảng dưới 60%. Đồng thời, khi sử dụng phương pháp RFE để chọn ra những đặc trưng, tuy nhiên, độ chính xác vẫn không thể cải thiện. Vì vậy, để cải thiện kết quả thực nghiệm mô hình, tác giả sử dụng phương pháp RFE để chọn lọc đặc trưng, sau đó sử dựng phương pháp SMOTE để giải quyết vấn đề mất cân bằng dữ liệu, các mô hình được thực nghiệm tuần tự và kết quả như Bảng 4. Từ kết quả trên, ta có thể thấy kết quả mô hình được cải thiện đáng kể, cụ thể mô hình có chỉ số Accuracy cao nhất là RF, SVM, LG (92.2%, 91.7% và 91.2%); mô hình có chỉ số F1-score cao nhất là RF, SVM, LG (0.921, 0.917 và 0.913). Từ những kết quả trên, mô hình RF cho thấy khả năng dự báo tốt đối với tập dữ liệu mất cân bằng tương tự như bộ dữ liệu về nhân sự của IBM được sử dụng trong nghiên cứu này. Mặc dù các mô hình thử nghiệm đạt được độ chính xác cao, nhưng trong trường hợp của các bộ dữ liệu mất cân bằng như trong nghiên cứu này, F1-score là một chỉ số quan trọng cần được xem xét để đảm bảo mô hình đạt được độ chính xác cao hơn và mang lại kết quả dự báo tốt hơn (Chicco & Jurman, 2020). Do đó, việc phân tích và mô tả tình trạng của bộ dữ liệu trước khi quyết định sử dụng mô hình phân tích dự đoán nào là vô cùng quan trọng, vì nó có ảnh hưởng trực tiếp đến kết quả thực nghiệm của mô hình. Bằng cách hiểu rõ về tính chất của dữ liệu, như các đặc điểm thống kê, phân phối, và sự mất cân bằng giữa các nhóm dữ liệu, chúng ta có thể lựa chọn phương thức tiếp cận và xử lý phù hợp, đồng thời điều chỉnh và tối ưu hóa các tham số trong mô hình để nâng cao kết quả dự báo. Từ đó, hỗ trợ quyết định và thực hiện hiệu quả các chiến lược quản lý nhân sự trong môi trường doanh nghiệp.

6. Kết luận và hướng phát triển

Sự nghỉ việc của nhân viên là vấn đề nhức nhối của hầu hết các doanh nghiệp, bởi nó gây nhiều ảnh hưởng đến phát triển nguồn nhân lực và sự phát triển của doanh nghiệp nói chung, do đó việc có thể dự đoán nhân viên nào có khả năng nghỉ việc sẽ mang lại nhiều lợi ý cho các doanh nghiệp. Bên cạnh đó, nghiên cứu này đã thực nghiệm nhiều mô hình máy học để thực nghiệm dự báo nhân viên có khả năng nghỉ việc tại công ty công nghệ IBM dựa trên 06 thuật toán máy học bao gồm: LG, KNN, DT, SVM, NN, RF. Từ kết quả dự báo này, ban lãnh đạo sẽ có những đánh giá, phân tích để tìm ra đặc điểm của nhân sự nghỉ việc hoặc nhân sự trung thành; đối với đặc điểm chung của nhân sự nghỉ việc như thông tin cá nhân và kinh nghiệm làm việc thì ban lãnh đạo có thể tham khảo nhận định khả năng làm việc lâu dài với doanh nghiệp hay không; hoặc đối với những đặc điểm chung của những nhân sự trung thành thì ban lãnh đạo có thể tiếp tục duy trì chiến sách tốt hoặc cải thiện hơn nữa để nhân viên có điều kiện, môi trường làm việc và phúc lợi tốt nhất khi làm việc trong điều kiện ngân sách và phù hợp với chiến lược phát triển của doanh nghiệp.

Trong bối cảnh nghiên cứu đang thực hiện trên tập dữ liệu của một công ty công nghệ, cụ thể là tập dữ liệu của IBM, việc áp dụng kết quả của nghiên cứu này vào môi trường doanh nghiệp ở Việt Nam có thể gặp phải một số hạn chế và thách thức do sự khác biệt về bối cảnh, văn hóa, quy mô và đặc thù ngành, tuy nhiên, khi có tập dữ liệu khác thì nghiên cứu này có thể tùy chỉnh các mô hình hoặc chỉnh sửa các báo cáo đề phù hợp với tập dữ liệu thực tế và đảm bảo kết quả ban đầu của nghiên cứu. Ngoài ra, nghiên cứu có thể tiếp tục mở rộng ứng dụng thêm các thuật toán học máy hoặc học sâu như học kết hợp (Ensemble learning), LSTM, … để nâng cao độ chính xác của mô hình.

LỜI CÁM ƠN

Bài báo này là sản phẩm của đề tài nghiên cứu khoa học công nghệ cấp trường có mã số CTD-2023-08 được tài trợ bởi Đại học Kinh tế TP. Hồ Chí Minh.

Tài liệu tham khảo

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. In *2013 International conference on machine intelligence and research advancement* (pp. 203-207). Katra, India: IEEE.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal, 26*(1), 1-11.

Brownlee, J. (2020). *Data preparation for machine learning: Data cleaning, feature selection, and data transforms in Python*. Victoria, Australia: Machine Learning Mastery.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 1-13. doi:10.1186/s12864-019-6413-7

Davidescu, A. A., Apostu, S. A., Paul, A., & Casuneanu, I. (2020). Work flexibility, job satisfaction, and job performance among Romanian employees-implications for sustainable human resource management. *Sustainability, 12*(15), Article 6086.

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences, 340*, 250-261. doi:10.1016/j.ins.2016.01.033

Dutta, S., Bandyopadhyay, S. K., & Bandyopadhyay, S. K. (2020). Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management Research, 6*(3), 80-85.

Grus, J. (2019). *Data science from scratch: First principles with python*. Sebastopol, CA: O'Reilly Media.

Guest, D. E. (1997). Human resource management and performance: A review and research agenda. *International Journal of Human Resource Management, 8*(3), 263-276.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Burlington, MA: Morgan Kaufmann.

Hinkin, T. R., & Tracey, J. B. (2000). The cost of turnover: Putting a price on the learning curve. *Cornell Hotel and Restaurant Administration Quarterly, 41*(3), 14-21.

Kamath, D. R., Jamsandekar, D. S., & Naik, D. P. (2019). Machine learning approach for employee attrition analysis. *International Journal of Trend in Scientific Research and Development,* 62-67.

Kaur, B., & Dogra, A. (2022). A machine learning model for predicting employees retention: An initiative towards HR through machine. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 653-657). Solan, Himachal Pradesh, India: IEEE.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. *The International Journal of Human Resource Management, 28*(1), 3-26.

Marvin, G., Jackson, M., & Alam, M. G. (2021). A machine learning approach for employee retention prediction. In *2021 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-8). Jeju, Korea: IEEE.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. Sebastopol, CA: O'Reilly Media, Inc.

Pavansubhasht. (2017). *IBM HR analytics employee attrition & performance*. Truy cập ngày 03/01/2024 tại https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Ray, A. N., & Sanyal, J. (2019). Machine learning based attrition prediction. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-4). Bangalore, India: IEEE.

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. (2022). Predicting employee attrition using machine learning approaches. *Applied Sciences, 12*(13), Article 6424.

Setiawan, I., Suprihanto, S., Nugraha, A. C., & Hutahaean, J. (2020). HR analytics: Employee attrition analysis using logistic regression. *IOP Conference Series: Materials Science and Engineering, 830*(3), Article 032001. doi:10.1088/1757-899X/830/3/032001

Tatsat, H., Puri, S., & Lookabaugh, B. (2020). *Machine learning and data science blueprints for finance*. Sebastopol, CA: O'Reilly Media.

Tran, T. K. (2015). Nghiên cứu mối quan hệ giữa hoạt động quản trị nhân sự và hiệu quả kinh doanh trong các doanh nghiệp vừa và nhỏ Việt Nam [A study on the relationship between human resource management activities and business performance in Vietnamese small and medium enterprises]. *Tạp chí Kinh tế và Phát triển, 220*(3), 61-68.

Vadakattu, R., Panda, B., Narayan, S., & Godhia, H. (2015). Enterprise subscription churn prediction. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1317-1321). Santa Clara, CA: IEEE.

Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance, 4*(2), 157-178.

Wang, Y., & Ni, X. S. (2019). *A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization*. Ithaca, NY: Cornell University.

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade, 58*(2), 472-482.

Zhao, J., & Dang, X. H. (2008). Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer churn as the example. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing* (pp. 1-4). Dalian, China: IEEE.