| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **CÔNG NGHỆ KHOA HỌC DỮ LIỆU**  **PHÂN TÍCH HIỆU SUẤT NHÂN VIÊN**  **Sinh viên thực hiện: Huỳnh Quốc Khánh-2100007626**  **Lớp: 21DTH2D**  **Chuyên ngành: KHOA HỌC DỮ LIỆU**  **Khóa : 2021-2025**  **Tp.HCM, ngày 01 tháng 08 Năm 2024**  **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **CÔNG NGHỆ KHOA HỌC DỮ LIỆU**  **PHÂN TÍCH HIỆU SUẤT NHÂN VIÊN**  **Sinh viên thực hiện:**  **Huỳnh Quốc Khánh-2100007626**  **Trần Đình Hùng**  **Trần Hồng Kính**  **Nguyễn Quốc Tấn Lộc**  **Lớp: 21DTH2D**  **Chuyên ngành: KHOA HỌC DỮ LIỆU**  **Khóa : 2021-2025**  **Tp.HCM, ngày 01 tháng 08 Năm 2024** |
| --- |

**LỜI MỞ ĐẦU**

Hiệu suất làm việc của nhân viên là yếu tố then chốt quyết định sự thành công của mọi tổ chức và doanh nghiệp. Trong bối cảnh cạnh tranh ngày càng gay gắt, việc quản lý và nâng cao hiệu suất làm việc không chỉ là nhiệm vụ quan trọng mà còn là một thách thức lớn đối với các nhà quản lý. Sự phát triển vượt bậc của công nghệ thông tin và trí tuệ nhân tạo đã mở ra những cơ hội đột phá, giúp các doanh nghiệp tiếp cận và phân tích dữ liệu một cách hiệu quả hơn bao giờ hết, từ đó đưa ra các quyết định chiến lược về nhân sự một cách chính xác và kịp thời.

**LỜI CẢM ƠN**

Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý Thầy Cô ở Khoa Công Nghệ Thông Tin Trường Đại Học Nguyễn Tất Thành đã truyền đạt vốn kiến thức quý báu của quý thầy cô cho chúng em trong suốt thời gian học tập tại trường.

Bài báo cáo của  em còn hạn chế nên không tránh khỏi những thiếu sót, em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy,cô để kiến thức của em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao ý thức và trình độ của mình của mình.

Em xin chân thành cảm ơn các quý thầy cô rất nhiều!

Sinh viên thực hiện

Huỳnh Quốc Khánh

**NHẬN XÉT**

*Tp.HCM, Ngày 01 tháng 08 năm 2024*

**Giảng viên**

**MỤC LỤC**

[**DANH MỤC HÌNH** 5](#_heading=h.tyjcwt)

[**DANH MỤC BẢNG** 6](#_heading=h.3dy6vkm)

[CHƯƠNG 1: GIỚI THIỆU 7](#_heading=h.1t3h5sf)

[**1.1.**](#_heading=h.4d34og8) **Giới thiệu đề tài** 7

[**1.2.**](#_heading=h.26in1rg) **Lý do chọn đề tài** 7

[CHƯƠNG 2 LÝ THUYẾT 9](#_heading=h.1y810tw)

[CHƯƠNG 3 PHÂN TÍCH DỮ LIỆU 21](#_heading=h.49x2ik5)

[3.1.Tập dữ liệu 21](#_heading=h.2p2csry)

[3.1.Phân tích dữ liệu 22](#_heading=h.23ckvvd)

[CHƯƠNG 5 THỰC NGHIỆM MÔ HÌNH 48](#_heading=h.4k668n3)

[KẾT LUẬN 52](#_heading=h.3q5sasy)

[Kết quả đạt được. 52](#_heading=h.25b2l0r)

[Hạn chế và hướng phát triển 53](#_heading=h.kgcv8k)

[TÀI LIỆU THAM KHẢO 54](#_heading=h.34g0dwd)

**DANH MỤC HÌNH**

[Hình 2.1.Biểu diễn độ lệch thông qua đồ thị 11](#_heading=h.4i7ojhp)

[Hình 2.2 .IQR 12](#_heading=h.2xcytpi)

[Hình 2.3.Ý tưởng chính của PCA 14](#_heading=h.1ci93xb)

[Hình 2.4. Các bước thực hiện PCA. 15](#_heading=h.3whwml4)

[Hình 2.5. Minh họa thuật toán SVM 16](#_heading=h.2bn6wsx)

[Hình 2.6. Minh họa cách phân lớp 17](#_heading=h.qsh70q)

[Hình 2.7. Phân loại rừng ngẫu nhiên 18](#_heading=h.3as4poj)

[Hình 2.8. Một lớp ẩn MLP 19](#_heading=h.1pxezwc)

[Hình 3.1 Đọc dữ liệu 23](#_heading=h.ihv636)

[Hình 3.2 Thông tin dữ liệu 23](#_heading=h.32hioqz)

[Hình 3.3. Thống kế đặc trưng số 24](#_heading=h.1hmsyys)

[Hình 3.4.Thống kế đặc phân loại 24](#_heading=h.41mghml)

[Hình 3.5.Kiểm tra trùng lặp 24](#_heading=h.vx1227)

[Hình 3.6. Kiểm tra dữ liệu thiếu 25](#_heading=h.3fwokq0)

[Hình 3.7. Tỷ lệ hiệu suất của các phòng ban 40](#_heading=h.3tbugp1)

[Hình 3.8. Phân bố dữ liệu 41](#_heading=h.nmf14n)

[Hình 3.9.Độ lệch 42](#_heading=h.37m2jsg)

[Hình 3.10.Độ nhọn 42](#_heading=h.1mrcu09)

[Hình 3.10. Phân phối trung bình mỗi hàng trong dữ liệu 43](#_heading=h.46r0co2)

[Hình 3.10. Kết quả kiểm tra ngoại lệ 44](#_heading=h.2lwamvv)

[Hình 3.11. Kết quả kiểm tra sau khi sử lý ngoại lệ 45](#_heading=h.111kx3o)

[Hình 3.2. Kết quả kiểm tra độ lệch độ nhọn 45](#_heading=h.3l18frh)

[Hình 5.1 Hình hàm xử lý và dự đoán của phần tự nhập các đặc trưng 50](#_heading=h.2zbgiuw)

[Hình 5.2 Hàm xử lý dữ liệu của phần dự đoán bằng file csv 50](#_heading=h.1egqt2p)

[Hình 5.3 Hàm tải file kết quả dự đoán 51](#_heading=h.3ygebqi)

[Hình 5.4 Hàm xác định file csv 51](#_heading=h.sqyw64)

[Hình 5.5 Xây dựng giao diện 51](#_heading=h.3cqmetx)

[Hình 5.6 Kết quả 1 51](#_heading=h.1rvwp1q)

[Hình 5.7 Dự đoán bằng cách nhập số liệu 52](#_heading=h.4bvk7pj)

[Hình 5.8 Dự đoán bằng tải file csv 52](#_heading=h.2r0uhxc)

[Hình 5.9 Thông báo định dạng file csv 53](#_heading=h.1664s55)

**DANH MỤC BẢNG**

[Bảng.3.1 Các đặc trưng của dữ liệu 21](#_heading=h.147n2zr)

[Bảng.3.2 Gía trị của đặc trưng thứ tự 22](#_heading=h.3o7alnk)

[Bảng 3.3 Giải thích biểu đồ 1 27](#_heading=h.1v1yuxt)

[Bảng 3.4 Giải thích biểu đồ 2 28](#_heading=h.4f1mdlm)

[Bảng 3.5 Giải thích biểu đồ 3 32](#_heading=h.2u6wntf)

[Bảng 3.6 Giải thích biểu đồ 4 34](#_heading=h.19c6y18)

[Bảng 3.7 Giải thích biểu đồ 6 39](#_heading=h.28h4qwu)

[Bảng 3.9.Kết quả accuracy 47](#_heading=h.206ipza)

# CHƯƠNG 1: GIỚI THIỆU

* 1. **Giới thiệu đề tài**

Trong thời buổi kinh tế phát triển và mức độ cạnh tranh giữa các doanh nghiệp gay gắt thì hiệu suất làm việc của nhân viên đóng một vai trò quan trọng trong việc phát triển của doanh nghiệp

Hiệu suất làm việc của nhân viên là yếu tố quan trọng quyết định sự thành công của một tổ chức. Việc phân tích hiệu suất nhân viên không chỉ giúp nâng cao hiệu quả làm việc mà còn cải thiện môi trường làm việc và sự hài lòng của nhân viên.

Phân tích này để tìm ra các yếu tố ảnh hưởng đến hiệu suất của nhân viên và dựa vào các yếu tố trên để áp dụng mô hình học máy để dự đoán hiệu suất làm việc của nhân viên

* 1. **Lý do chọn đề tài**

**Tầm quan trọng chiến lược của hiệu suất nhân viên**

Hiệu suất làm việc của nhân viên không chỉ ảnh hưởng trực tiếp đến năng suất và lợi nhuận của doanh nghiệp mà còn định hình văn hóa tổ chức và động lực làm việc. Việc phân tích và nâng cao hiệu suất nhân viên là yếu tố then chốt quyết định sự thành công và phát triển bền vững của mọi tổ chức.

**Đối mặt với thách thức quản lý trong thời đại mới**

Trong kỷ nguyên công nghệ số và cạnh tranh toàn cầu, các nhà quản lý phải đối diện với nhiều thách thức phức tạp trong việc theo dõi, đánh giá và cải thiện hiệu suất nhân viên. Các phương pháp truyền thống không còn đủ khả năng để đáp ứng yêu cầu về tính khách quan, toàn diện và kịp thời.

**Nâng cao chất lượng quyết định chiến lược**

Phân tích này cho các nhà quản lý cái nhìn sâu sắc và toàn diện hơn về hiệu suất làm việc của từng cá nhân và cả tổ chức. Thông tin chi tiết này hỗ trợ việc đưa ra các quyết định chiến lược chính xác, từ đó tối ưu hóa nguồn lực và nâng cao hiệu quả hoạt động của doanh nghiệp.

* 1. **Mục tiêu đề tài.**

Phân tích hiệu suất của từng phòng ban để có biện pháp cải thiện hoặc nâng cao của các phòng ban

Phân tích ra được các yếu tố quan trọng ảnh hưởng đến hiệu suất

Đề xuất cải thiện hiệu suất của nhân viên

Đào tạo mô hình dự đoán hiệu suất nhân viên

# CHƯƠNG 2 LÝ THUYẾT

***2.1. Thống kê mô tả***

Một nhiệm vụ cơ bản trong nhiều phân tích thống kê là xác định đặc điểm của vị trí và sự biến đổi của một tập dữ liệu. Tập dữ liệu có thể được phân phối theo nhiều cách, như trải rộng hơn ở bên trái hoặc bên phải hoặc trải đều. Độ lệch và độ nhọn là hai đại lượng số thể hiện đặc trưng của hình dáng phân phối và cung cấp nhiều thông tin hơn để đánh giá rủi ro hơn là chỉ sử dụng độ lệch chuẩn. Trong đó, biểu đồ Histogram là một kỹ thuật đồ họa hiệu quả để biểu diễn 2 đại lượng này.

*2.1.1. Độ lệch (skewness)*

Độ lệch (skewness) là thước đo mức độ đối xứng, bất đối xứng của một phân phối. Một phân phối, hoặc tập dữ liệu được xem là đối xứng nếu nó giống nhau ở bên trái và bên phải điểm trung tâm.

* Độ lệch có nhiều cách tính khác nhau:
* Đối với dữ liệu đơn biến x1, x2, ..., xN, công thức dưới đây được gọi là độ lệch Fisher-Pearson (Fisher-Pearson coefficient of skewness)

là giá trị trung bình,s là độ lệch chuẩn

N là số điểm dữ liệu

* Công thức tính độ lệch Galton (còn được gọi là độ lệch của Bowley) theo các giá trị của tứ phân vị như sau:

Q1 là tứ phân vị thứ nhất (hay phân vị dưới)

  Q3 là tứ phân vị thứ 3 (hay phân vị trên)

Q2 là trung vị.

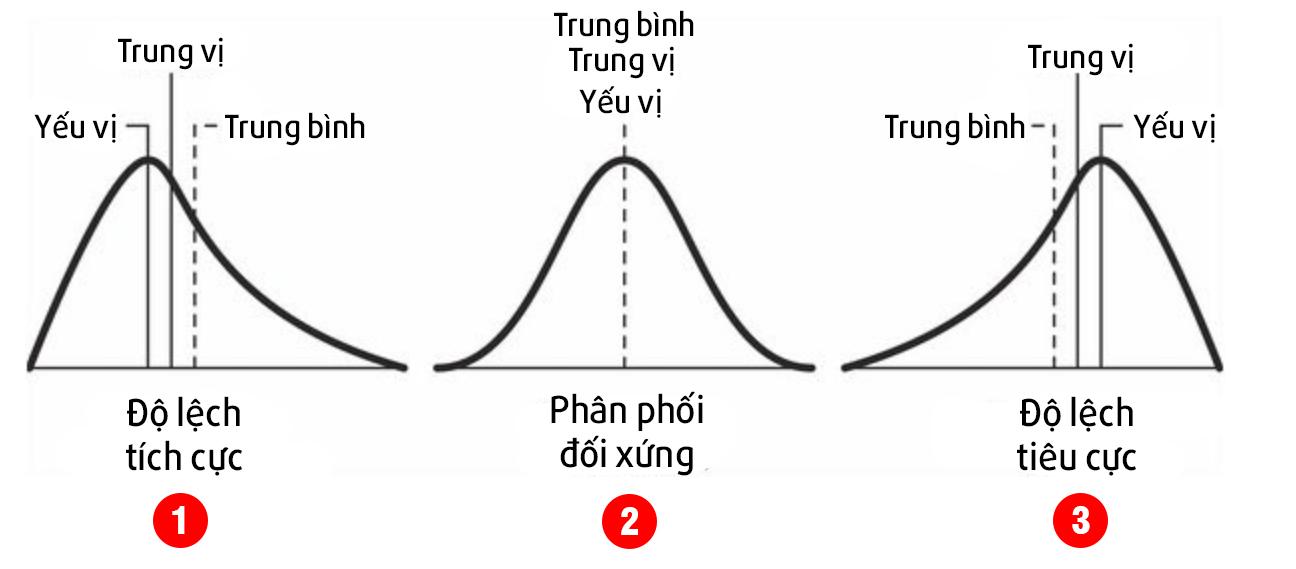
* Công thức tính độ lệch đơn giản do Karl Pearson đề xuất theo yếu vị (SPearson 1) và trung vị (SPearson 2) :

x ̅   là giá trị trung bình

x ̂   là yếu vị,

x ̃   là trung vị

s là độ lệch chuẩn



*Hình 2.1.Biểu diễn độ lệch thông qua đồ thị*

*2.1.2.Độ nhọn*

Độ nhọn (kurtosis) là là một đại lượng thống kê mô tả đo mức độ tập trung của các quan sát ở phần đuôi hoặc đỉnh của phân phân phối. Đỉnh là phần cao nhất của phân phối và đuôi là phần cuối của phân phối. Các tập dữ liệu có hệ số kurtosis cao có xu hướng dữ liệu tập trung về phần đuôi hoặc phần dữ liệu ngoại lệ, các tập dữ liệu có hệ số kurtosis thấp có xu

hướng dữ liệu tập trung quanh vị trí trung tâm và có thể không có dữ liệu ngoại lệ.

Công thức tính độ nhọn như sau:

x ̅   là giá trị trung bình

 s là độ lệch chuẩn

N là số điểm dữ liệu

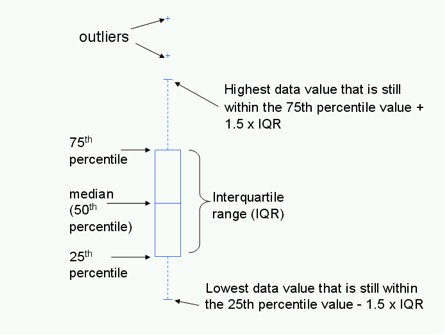
Lưu ý: khi tính toán hệ số kurtosis, s được tính bằng N ở mẫu số thay vì N-1.

Hệ số kurtosis cho phân phối chuẩn nếu tính theo công thức trên sẽ là 3

***2.2. Kiểm tra xử lý ngoại lệ(IQR)***

Tứ phân vị là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất Q1 (25th), thứ hai Q2 (50th) hay median, và thứ ba Q3 (75th). Ba giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau. Tứ phân vị được xác định như sau:

* + - * Sắp xếp các số theo thứ tự tăng dần
      * Cắt dãy số thành 4 phàn bằng nhau
      * Tứ phân vị là các giá trị tại vị trí cắt



*Hình 2.2 .IQR*

IQR là sự khác biệt giữa tứ phân vị thứ nhất Q1 và tứ phân vị thứ ba Q3:

IQR=Q3−Q1

Giá trị IQR có thể sử dụng để xác định outliers bằng cách thiết lập các giá trị biên Upper/Lower giống với phương pháp STD như sau: Nếu chúng ta trừ đi kxIQR từ tứ phân vị đầu tiên Q1, bất kỳ giá trị dữ liệu nào nhỏ hơn con số này được coi là giá trị outliers. Tương tự như vậy, nếu chúng ta thêm kxIQR đến tứ phân vị thứ ba Q3, bất kỳ giá trị dữ liệu nào lớn hơn con số này được coi là outliers. Giá trị k thường được chọn là 1.5. Trong trường hợp xác định các extreme outliers có thể dùng giá trị k = 3.

Các giá trị ngoại lệ thường được định nghĩa là những giá trị nằm ngoài khoảng [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR].

***2.3. Chuẩn hóa dữ liệu(StandardScaler)***

StandardScaler là một công cụ trong thư viện scikit-learn của Python dùng để chuẩn hóa các đặc trưng (features) trong dữ liệu. Nó giúp biến đổi dữ liệu sao cho chúng có phân phối chuẩn (mean = 0 và độ lệch chuẩn = 1).

x là giá trị của đặc trưng

μ là giá trị trung bình của đặc trưng

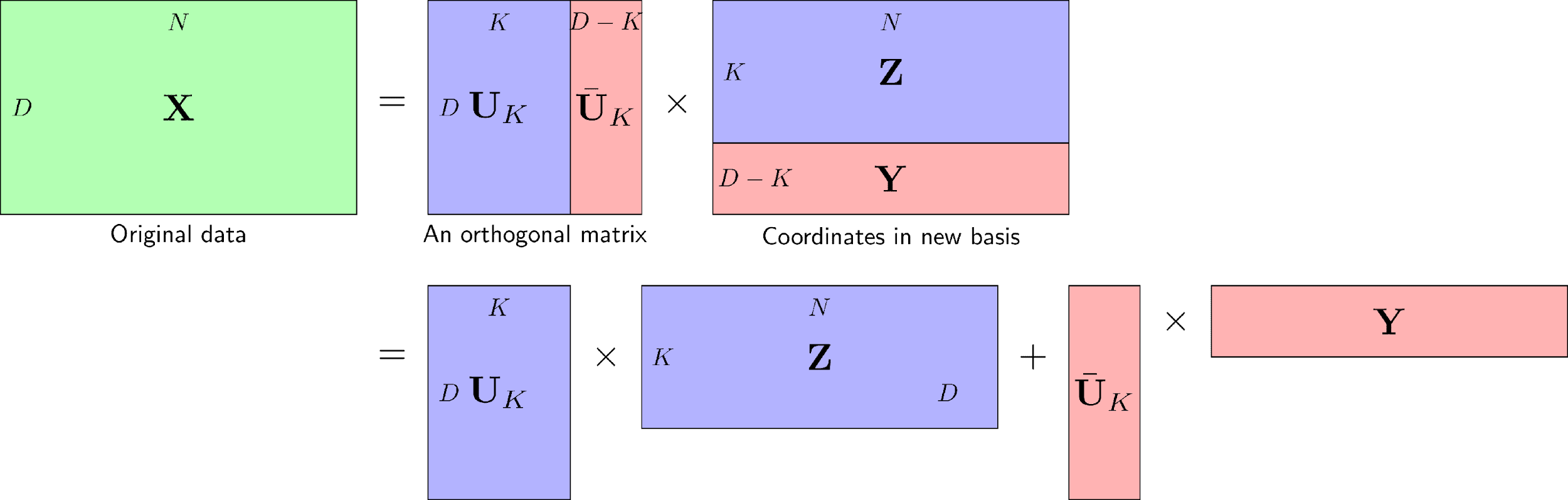
σ là độ lệch chuẩn của đặc trưng.

Việc chuẩn hóa giúp cải thiện hiệu suất của nhiều mô hình học máy, đặc biệt là những mô hình nhạy cảm với tỉ lệ của các đặc trưng như hồi quy logistic, SVM, và các mạng nơ-ron.

***2.4. Giảm chiều dữ liệu (PCA)***

Phân tích thành phần chính (PCA) là một kỹ thuật phổ biến để phân tích các tập dữ liệu lớn chứa nhiều chiều/tính năng trên mỗi quan sát, tăng khả năng diễn giải dữ liệu trong khi vẫn bảo toàn lượng thông tin tối đa và cho phép trực quan hóa dữ liệu đa chiều. Về mặt hình thức, PCA là một kỹ thuật thống kê để giảm chiều của một tập dữ liệu

PCA chính là phương pháp đi tìm một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin. Và để cho đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới.



*Hình 2.3.Ý tưởng chính của PCA*

Quan sát hình vẽ trên với cơ sở mới U=[UK,¯UK]𝑈=[𝑈𝐾,𝑈¯𝐾] là một hệ trực chuẩn với UK𝑈𝐾 là ma trận con tạo bởi K𝐾 cột đầu tiên của U𝑈. Với cơ sở mới này, ma trận dữ liệu có thể được viết thành:

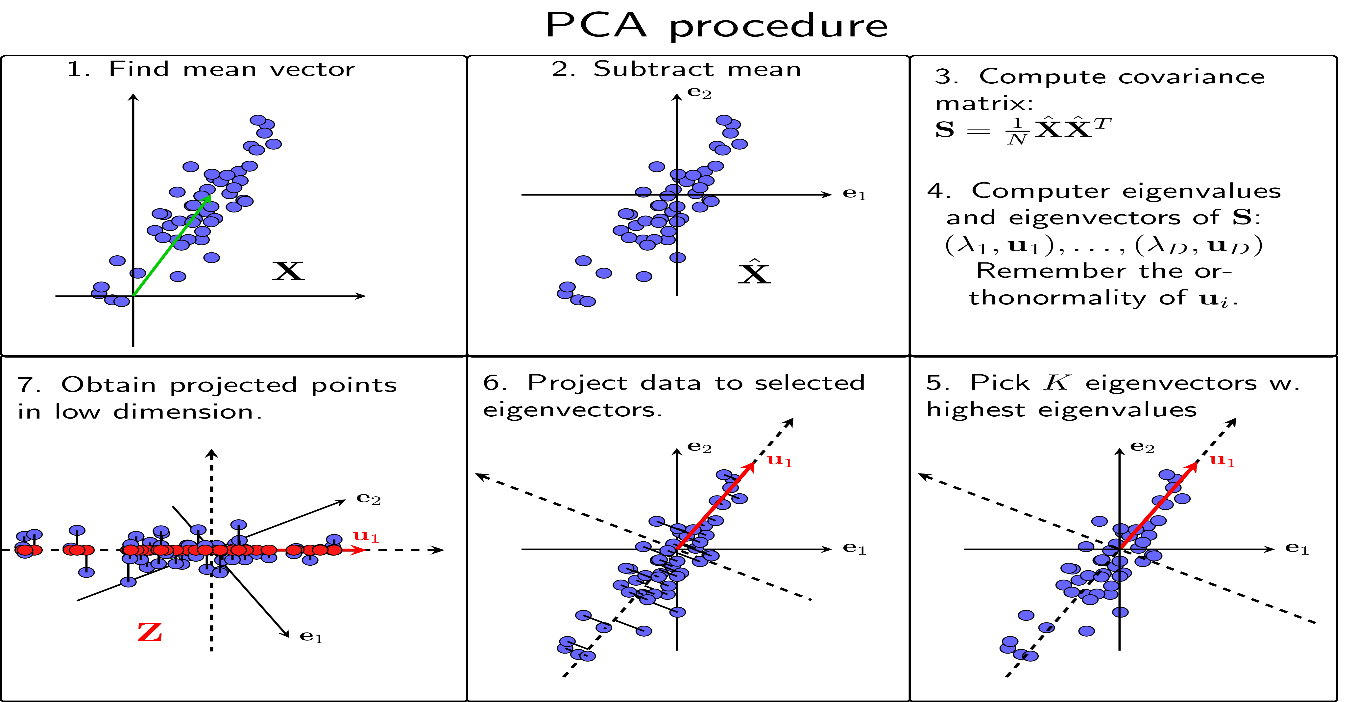
X=UKZ+¯UKY

Từ đây ta cũng suy ra:

Mục đích của PCA là đi tìm ma trận trực giao U𝑈 sao cho phần lớn thông tin được giữ lại ở phần màu xanh  và phần màu đỏ  sẽ được lược bỏ và thay bằng một ma trận không phụ thuộc vào từng điểm dữ liệu

-Các bước thực hiện PCA:

1. Tính vector kỳ vọng của toàn bộ dữ liệu
2. Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu
3. Tính ma trận hiệp phương sai
4. Tính các trị riêng và vector riêng có norm bằng 1 của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.
5. Chọn 𝐾 vector riêng ứng với 𝐾 trị riêng lớn nhất để xây dựng ma trận 𝑈𝐾 có các cột tạo thành một hệ trực giao. 𝐾 vectors này, còn được gọi là các thành phần chính, tạo thành một không gian con *gần* với phân bố của dữ liệu ban đầu đã chuẩn hoá.
6. Chiếu dữ liệu ban đầu đã chuẩn hoá xuống không gian con tìm được.
7. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới



*Hình 2.4. Các bước thực hiện PCA.*

***2.5. Tăng cường dữ liệu(SMOTE)***

SMOTE (Synthetic Minority Over-sampling Technique) là một phương pháp phổ biến được sử dụng để giải quyết vấn đề mất cân bằng dữ liệu trong các bài toán học máy. Khi dữ liệu mất cân bằng, lớp thiểu số thường bị bỏ qua, dẫn đến hiệu suất phân loại kém. SMOTE giúp cân bằng dữ liệu bằng cách tạo ra các mẫu mới từ lớp thiểu số.

-Cách hoạt động của smote:

1. SMOTE tạo ra các mẫu mới từ lớp thiểu số bằng cách tạo ra các điểm dữ liệu tổng hợp giữa các điểm hiện có. Các bước thực hiện như sau:
2. Chọn ngẫu nhiên một mẫu từ lớp thiểu số: Giả sử bạn chọn một điểm từ lớp thiểu số.
3. Tìm k-láng giềng gần nhất: Tìm k điểm láng giềng gần nhất của trong lớp thiểu số (thường k=5).
4. Tạo mẫu mới: Chọn ngẫu nhiên một trong các láng giềng gần nhất sau đó tạo một điểm mới nằm giữa và . Điểm mới được tính theo công thức:

trong đó λ là một giá trị ngẫu nhiên trong khoảng [0, 1].

***2.6. Mô hình***

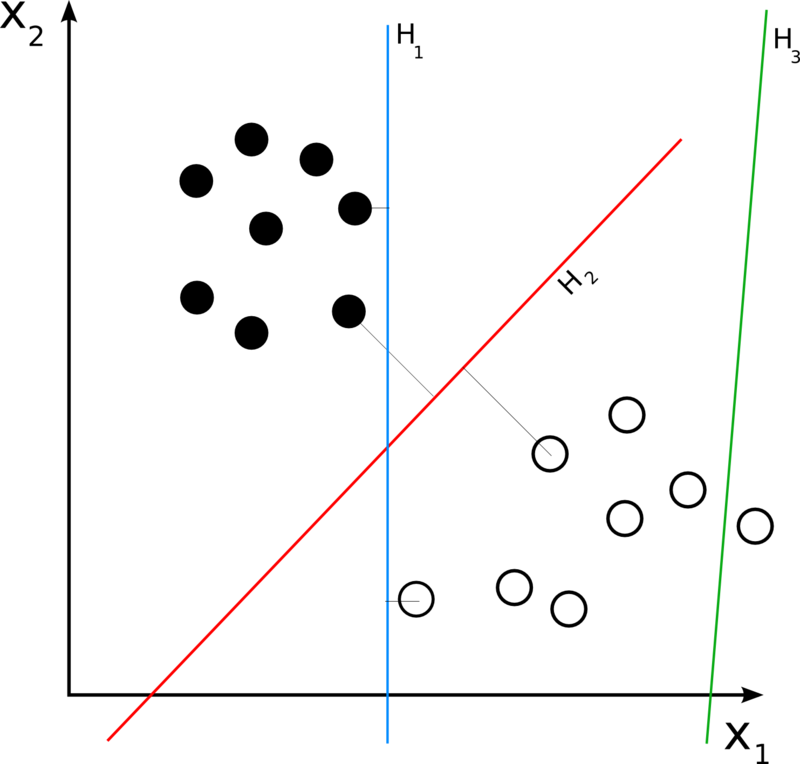
*2.6.1 Support Vector Machine(SVM)*

Mô hình Support Vector Machine - SVM là một mô hình máy học thuộc nhóm Supervised Learning được sử dụng cho các bài toán Classification (Phân lớp) và Regression (Hồi quy).

Ta còn có thể phân loại mô hình này vào loại mô hình tuyến tính (Linear Model), loại này bao gồm các thuật toán có chung một dạng hypothesis function (có thể gọi là hàm giả định, tóm lại là cách nó đưa ra một dự đoán - prediction) có dạng là những đường thẳng (trong không gian 2 chiều), mặt phẳng, siêu phẳng (trong không gian nhiều chiều).

-Nền tảng mô hình SVM

Mô hình SVM đặt ra vấn đề giải quyết 1 bài toán đơn giản nhất là phân lớp 2 lớp dữ liệu



*Hình 2.5. Minh họa thuật toán SVM*

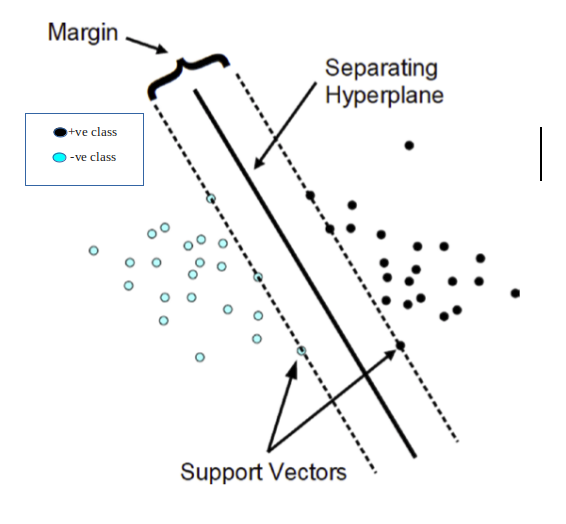
Đường thẳng phân tách 2 lớp dữ liệu cách đều lớn nhất sẽ cho ra khả năng tổng quát tốt nhất. Chúng ta gọi khoảng cách đều đó là margin (lề). Vậy nên mô hình SVM chỉ có ý tưởng cơ bản như vậy tìm đường thẳng margin lớn nhất.

-Tập Hypothesis của SVM

Tập Hypothesis của SVM có dạng:

Với  𝑥∈𝑅,𝑤∈𝑅,𝑏∈𝑅, ta có 𝑥 là bộ dữ liệu trong không gian 𝑑 chiều và 𝑤,𝑏 là các tham số về siêu phẳng trong không gian để phân tách.

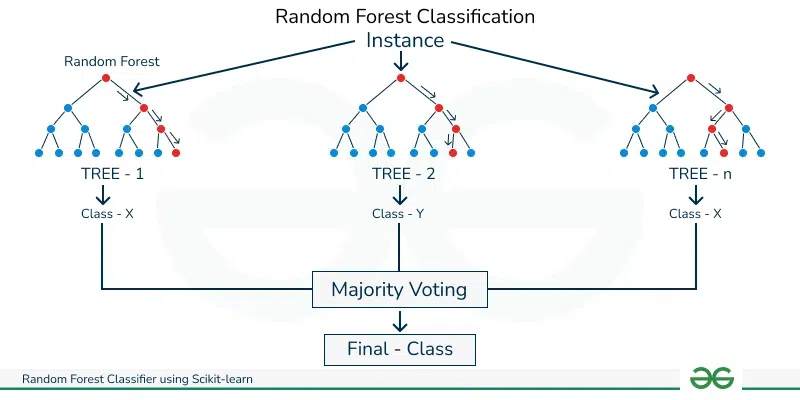
Ngoài ra sign(x) là một xác định dấu, nếu x>=0 sign(x)=1) ngược lại nếu x< sign(x)=−1.Đây là cách phân lớp của từng điểm dữ liệu y € =1,+1



*Hình 2.6. Minh họa cách phân lớp*

*2.6.2.Random Forest Classifier*

Rừng ngẫu nhiên hoặc rừng quyết định ngẫu nhiên là một thuật toán học máy có giám sát được sử dụng để phân loại, hồi quy và các tác vụ khác bằng cách sử dụng cây quyết định. Rừng ngẫu nhiên đặc biệt phù hợp để xử lý các tập dữ liệu lớn và phức tạp, xử lý các không gian đặc điểm có nhiều chiều và cung cấp thông tin chi tiết về tầm quan trọng của đặc điểm. Thuật toán duy trì độ chính xác dự đoán cao trong khi giảm thiểu tình trạng quá khớp



*Hình 2.7. Phân loại rừng ngẫu nhiên*

Bộ phân loại rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định từ một tập hợp con được chọn ngẫu nhiên của tập huấn luyện. Đây là một tập hợp các cây quyết định (DT) từ một tập hợp con được chọn ngẫu nhiên của tập huấn luyện và sau đó thu thập các phiếu bầu từ các cây quyết định khác nhau để quyết định dự đoán cuối cùng

-*Cách thức hoạt động:*

Thuật toán xây dựng nhiều cây quyết định trong quá trình đào tạo và đưa ra lớp là chế độ của các lớp phân loại. Mỗi cây quyết định trong rừng ngẫu nhiên được xây dựng bằng cách sử dụng một tập hợp con của dữ liệu đào tạo và một tập hợp con ngẫu nhiên của các tính năng giới thiệu sự đa dạng giữa các cây, làm cho mô hình mạnh mẽ hơn và ít bị quá khớp hơn.

Thuật toán rừng ngẫu nhiên sử dụng một kỹ thuật gọi là bagging (Bootstrap Aggregating) để tạo ra các tập hợp con đa dạng

Tại mỗi lần chia tách, thuật toán sẽ chọn tính năng tốt nhất từ ​​tập hợp con ngẫu nhiên, tối ưu hóa để đạt được thông tin hoặc độ tạp chất Gini

Sau khi rừng ngẫu nhiên được đào tạo, nó có thể đưa ra dự đoán bằng cách sử dụng "phiếu bầu" của mỗi cây cho một lớp và lớp có nhiều phiếu bầu nhất sẽ trở thành lớp được dự đoán cho dữ liệu đầu vào.

-Lựa chọn tính năng :

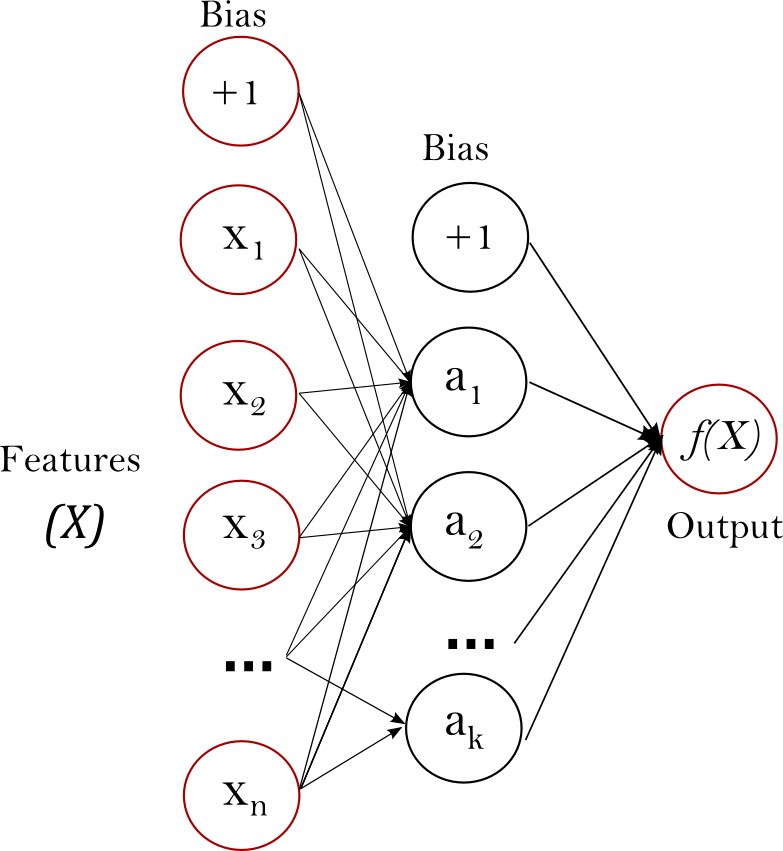
Việc lựa chọn tính năng trong rừng ngẫu nhiên vốn được nhúng vào quá trình xây dựng từng cây quyết định riêng lẻ và quá trình tổng hợp.

*2.6.3. Artificial Neural Network [MLP Classifier]*

Multi-layer Perceptron (MLP) là một thuật toán học có giám sát học một hàm bằng cách đào tạo trên một tập dữ liệu,trong đó m là số chiều đầu vào và o là số chiều cho đầu ra

Cho một tập hợp các tính năng X=x1,x2,…,xn và mục tiêu y mó có thể học một hàm xấp xỉ phi tuyến cho cả phân loại hoặc hồi quy.

Nó khác với hồi quy logistic ở chỗ giữa lớp đầu vào và lớp đầu ra, có thể có một hoặc nhiều lớp phi tuyến tính, được gọi là lớp ẩn



*Hình 2.8. Một lớp ẩn MLP*

Lớp ngoài cùng bên trái, được gọi là lớp đầu vào, bao gồm một tập hợp các tế bào thần kinh đại diện cho các tính năng đầu vào. Mỗi nơ-ron trong lớp ẩn chuyển đổi các giá trị từ lớp trước đó bằng phép tổng tuyến tính có trọng số theo sau là một hàm kích hoạt phi tuyến tính

giống như hàm tan hyperbolic Lớp đầu ra nhận các giá trị từ lớp ẩn cuối cùng và chuyển đổi chúng thành các giá trị đầu ra.

MLP đào tạo sử dụng Stochastic Gradient Descent , Adam hoặc L-BFGS . Stochastic Gradient Descent (SGD) cập nhật các tham số bằng cách sử dụng gradient của hàm mất mát đối với một tham số cần điều chỉnh, tức là

là tốc độ học tập kiểm soát kích thước bước trong tìm kiếm không gian tham số.

𝐿𝑜𝑠𝑠là hàm mất mát được sử dụng cho mạng.

-Độ phức tạp

Giả sử có 𝑛mẫu đào tạo,𝑚 đặc trưng,𝑘 các lớp ẩn, mỗi lớp chứa ℎtế bào thần kinh - để đơn giản, và𝑜 đầu ra của tế bào thần kinh. Độ phức tạp thời gian của sự lan truyền ngược là  
𝑂(𝑖⋅𝑛⋅(𝑚⋅ℎ+(𝑘−1)⋅ℎ⋅ℎ+ℎ⋅𝑜)) 𝑖 là số lần lặp lại. Vì truyền ngược có độ phức tạp về thời gian cao nên nên bắt đầu với số lượng nơ-ron ẩn nhỏ hơn và một vài lớp ẩn để đào tạo

***2.7. Đánh giá mô hình***

*2.7.1.* Accuracy\_score

Accuracy\_score là một chỉ số đo lường hiệu suất của mô hình phân loại, thể hiện tỷ lệ dự đoán đúng trên tổng số dự đoán. Nó là một trong những chỉ số đơn giản và phổ biến nhất để đánh giá các mô hình học máy.

Công thức tính Accuracy

Với ma trận confusion matrix(ma trận hỗn độn) accuracy được tính như sau:

*TP* = Kết quả dương tính thật, *TN* = Kết quả âm tính thật, *FP* = Kết quả dương tính giả và *FN* = Kết quả âm tính giả.

*2.7.2.* Classification\_report

classification\_report trong thư viện scikit-learn là một công cụ hữu ích để đánh giá hiệu suất của mô hình phân loại. Nó cung cấp một bản tóm tắt chi tiết về các chỉ số chính như precision, recall, f1-score, và support cho từng lớp trong tập dữ liệu.

Các chỉ số trong classification\_report:

Precision: Tỷ lệ của các dự đoán đúng trong số tất cả các dự đoán cho một lớp nhất định.

Recall: Tỷ lệ của các dự đoán đúng trong số tất cả các mẫu thực sự thuộc về lớp đó.

F1-Score: Trung bình điều hòa của precision và recall. F1-score là một chỉ số tốt khi bạn cần cân bằng giữa precision và recall.

Support: Số lượng thực tế của các mẫu thuộc về mỗi lớp trong tập kiểm tra.

# CHƯƠNG 3 PHÂN TÍCH DỮ LIỆU

## 3.1.Tập dữ liệu

Bộ dữ liệu [Employee\_Performance](https://drive.google.com/drive/folders/1R5rQXSxxC7b7hG50jLqt6XMeP2ljJEd1?usp=drive_link) bao gồm 1200 hàng. Các đặc trưng có trong dữ liệu là 28 cột. Đặc trưng được phân loại thành định lượng và định tính trong đó 19 đặc trưng là định lượng (11 cột bao gồm dữ liệu số và 8 cột bao gồm dữ liệu thứ tự) và 8 đặc trưng là định tính. EmpNumber bao gồm dữ liệu chữ số (các giá trị riêng biệt) không đóng vai trò là tính năng có liên quan để đánh giá hiệu suất.

| Đặc trưng doanh mục | Đặc trưng số | Đặc trưng thứ tự |
| --- | --- | --- |
| EmpNumber  Gender  EducationBackground  MaritalStatus  EmpDepartment  EmpJobRole  BusinessTravelFrequency  OverTime  Attrition | Age  DistanceFromHome  EmpHourlyRate  NumCompaniesWorked  EmpLastSalaryHikePercent  TotalWorkExperienceInYears  TrainingTimesLastYear  ExperienceYearsAtThisCompany  ExperienceYearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager | EmpEducationLevel  EmpEnvironmentSatisfaction  EmpJobInvolvement  EmpJobLevel  EmpJobSatisfaction  EmpRelationshipSatisfaction  EmpWorkLifeBalance  PerformanceRating |

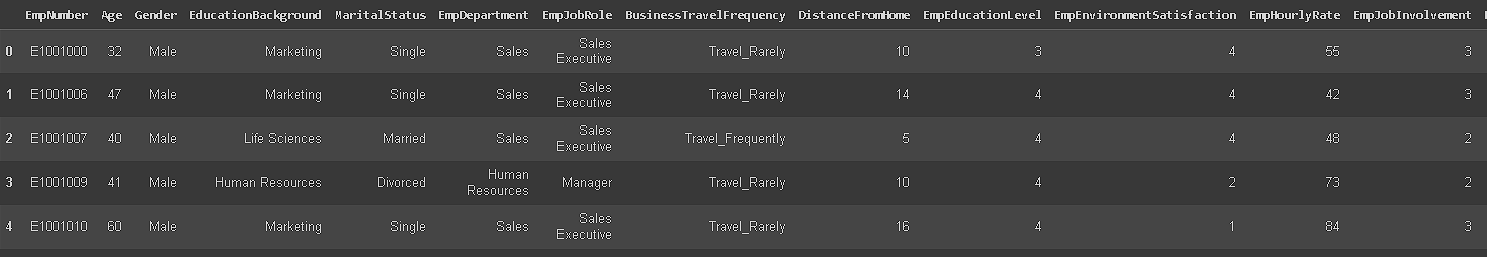
*Bảng.3.1 Các đặc trưng của dữ liệu*

| EmpEducationLevel | | 1 'Below College' | | --- | | 2 'College' | | 3 'Bachelor' | | 4 'Master' | | 5 'Doctor' | |
| --- | --- | --- | --- | --- | --- | --- |
| EmpEnvironmentSatisfaction  EmpJobInvolvement  EmpJobSatisfaction  RelationshipSatisfaction | | 1 'Low' | | --- | | 2 'Medium' | | 3 'High' | | 4 'Very High' | |
| PerformanceRating | | 1 'Low' | | --- | | 2 'Good' | | 3 'Excellent' | | 4 'Outstanding' | |
| EmpWorkLifeBalance | | 1 'Bad' | | --- | | 2 'Good' | | 3 'Better' | | 4 'Best' | |
| EmpJobLevel | 1 ‘Basic’  2 ‘Intermediate’  3 ‘Advanced’  4 ‘Expert’  5 ‘Specialist’ |

*Bảng.3.2 Gía trị của đặc trưng thứ tự*

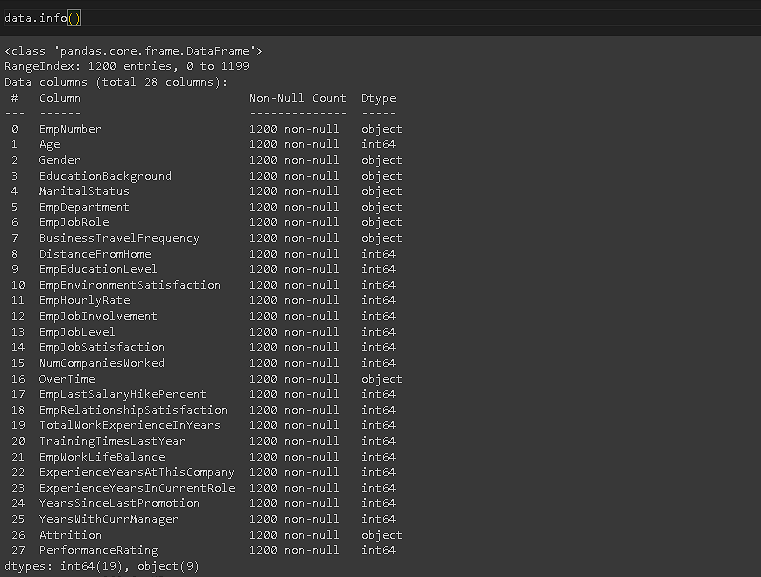
## 3.1.Phân tích dữ liệu

*-Đọc dữ liệu*



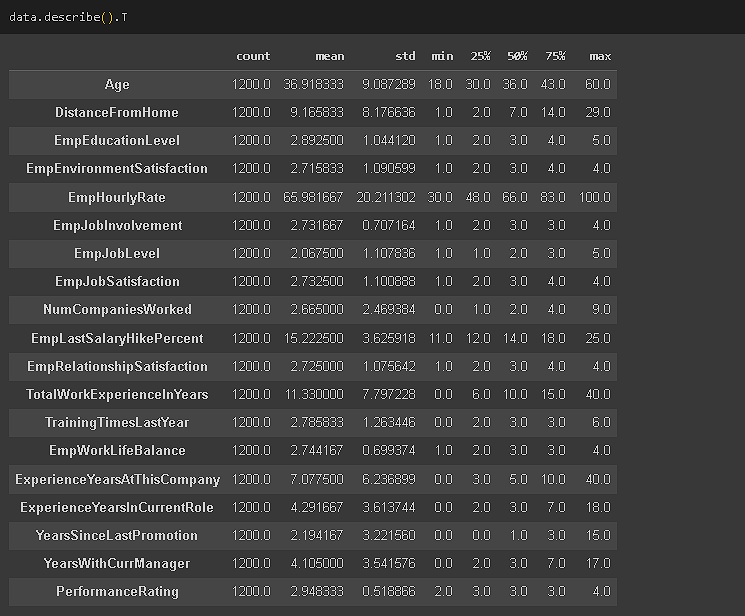
*Hình 3.1 Đọc dữ liệu*

*-Thông tin của dữ liệu*



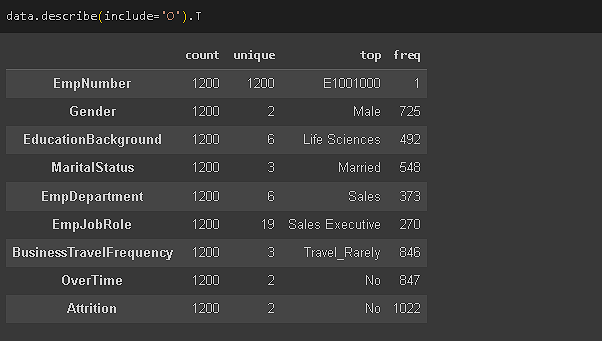
*Hình 3.2 Thông tin dữ liệu*

*-Thống kê đặc trưng số*



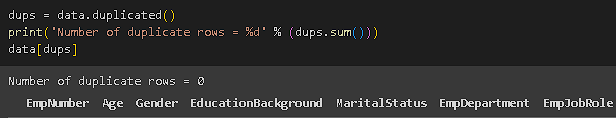
*Hình 3.3. Thống kế đặc trưng số*

*-Thống kê đặc trưng phân loại*



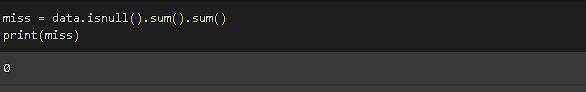
*Hình 3.4.Thống kế đặc phân loại*

*-Kiểm tra trùng lặp*

**

*Hình 3.5.Kiểm tra trùng lặp*

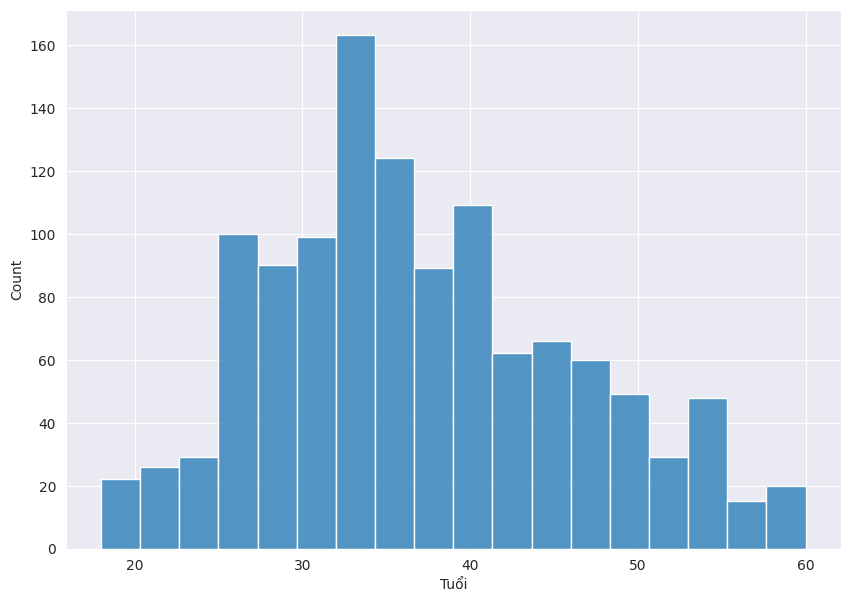
*-Kiểm tra dữ liệu thiếu*



*Hình 3.6. Kiểm tra dữ liệu thiếu*

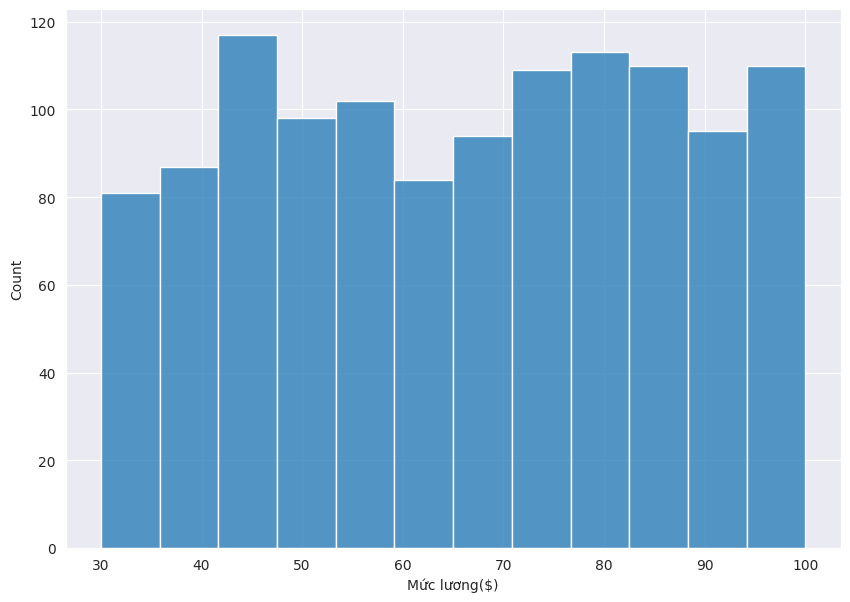
* **Phân tích đơn biến**

*1.Tuổi*



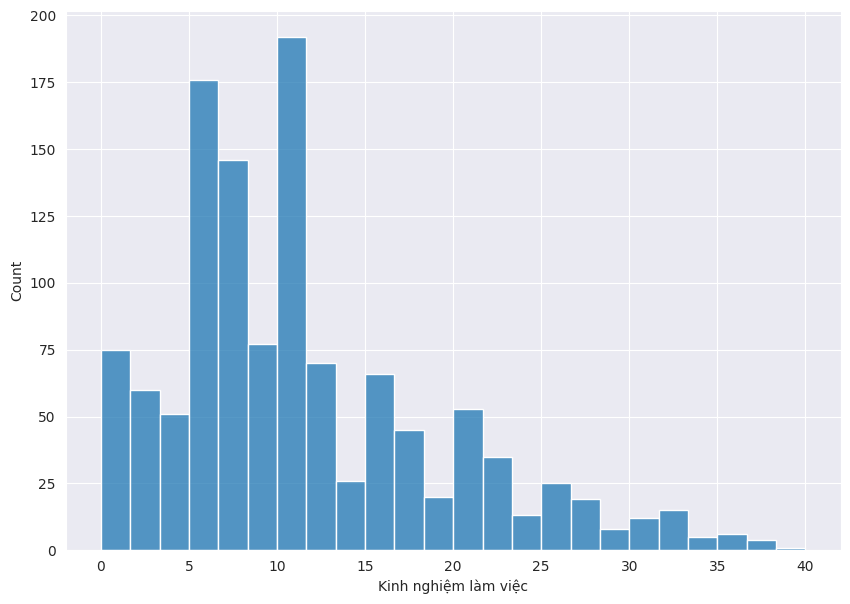
Độ tuổi từ 18 đến 60, phần lớn nhân viên có độ tuổi từ 25 đến 40.

*2.Mức lương theo giờ*



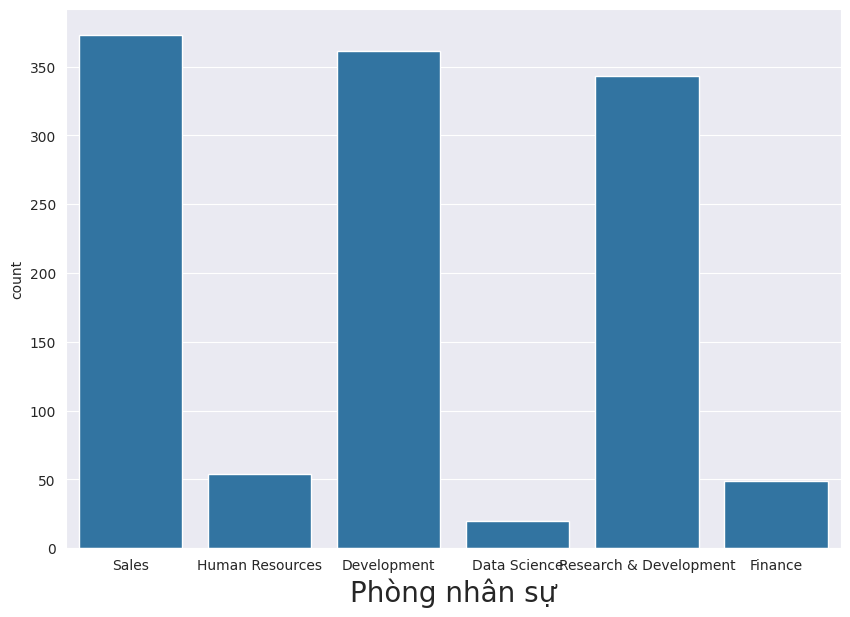
Mức lương theo giờ của nhân viên dao động từ 30 đến 100, phần lớn mức lương theo giờ của nhân viên là 45.

*3.Tổng kinh nghiệm làm việc*



Kinh nghiệm làm việc từ 0 đến 40, hầu hết nhân viên có kinh nghiệm từ 5 đến 10

*4.Phòng nhân sự*

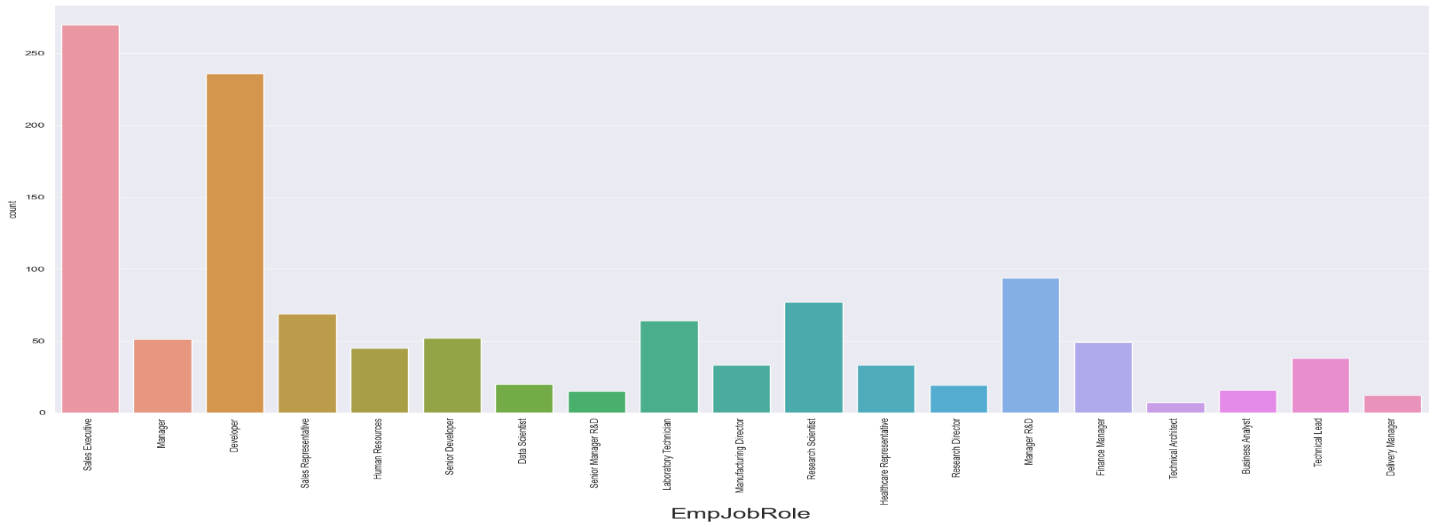


Hầu hết nhân viên từ phòng Sales cũng như phòng Development và phòng Research & Development

Nhân viên phòng Human Resources và Finace gần như bằng nhau.

Nhân viên data Science có số lượng ít nhất trong công ty.

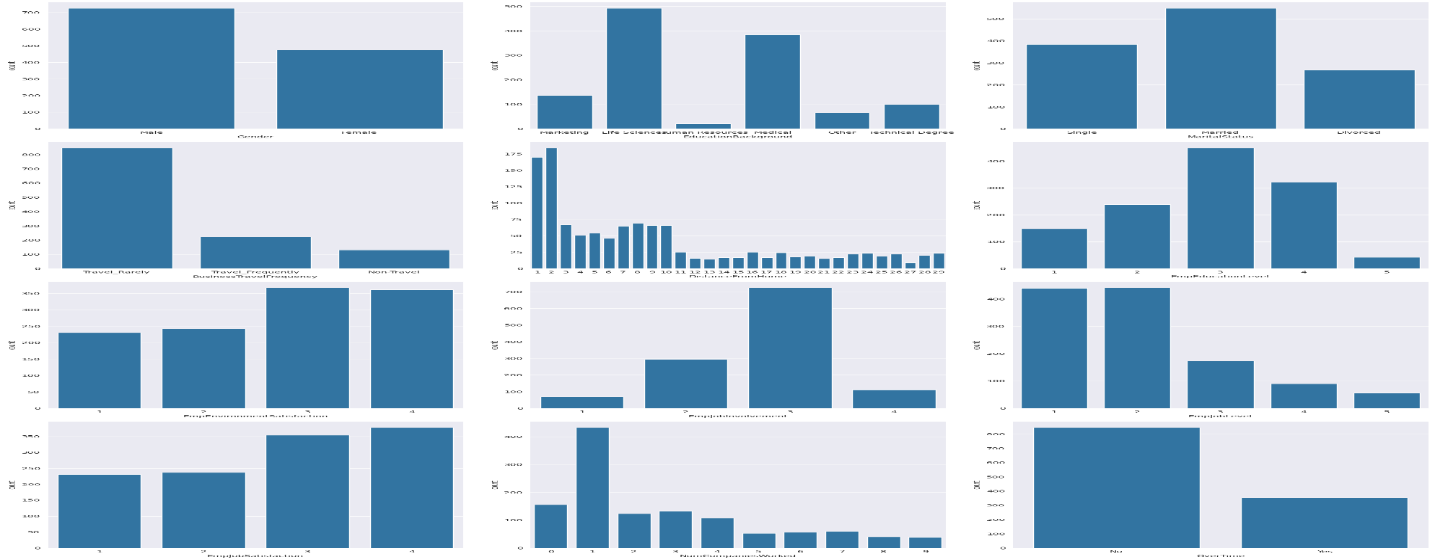
*5.Vai trò công việc*



Vai trò công việc của Sales Executive là tối đa cũng như vai trò công việc của Development.

Manager R&D và Research Scientist là ít hơn 100 và tất cả các vai trò công việc còn lại là ít hơn 65.

*-Các đặc trưng khác*



| *6. Gender:*  Hầu hết nam giới có mặt trong công ty,nữ giới chiếm khoảng 2/3 nam giới  *7. EducationBackground:*  Hầu hết trình độ học vấn của nhân viên là Life Sciences,Medical và Maketting là trên 100  Còn lại là dưới 100.  *8. MaritalStatus:*  Hầu hết nhân viên đã kết hôn và dưới 400 nhân viên còn độc thân.  Dưới 300 nhân viên tình trạng hôn nhân là đã ly hôn.  *9. BusinessTravelFrequency:*  Hầu hết là hiếm khi đi công tác và khoảng 200 nhân viên là đi công tác  còn lại dưới 120 nhân viên không đi công tác  *10. DistanceFromHome:*  Khoảng cách từ nhà đến công ty là từ 1km đến 29km.  Khoảng cách từ nhà đến công ty của nhân viên nhiều nhất là dưới 10km.  *11. EmpEducationLevel:*  Tổng phạm vi trình độ học vấn là từ 1 đến 5, Số lượng nhân viên trình độ học vấn tối đa là 3 & 4.  Nhân viên trình độ học vấn 2 là dưới 250 và nhân viên trình độ học vấn 1 & 5 còn lại là dưới 150. | *12. EmpEnvironmentSatisfaction:*  Phần lớn sự hài lòng của nhân viên là 3 & 4  Sự hài lòng về môi trường 1 & 2 là dưới 250  *13. EmpJobInvolvement:*  Phần lớn sự tham gia công việc của nhân viên là 3 cũng như sự tham gia công việc của nhân viên là 2  Phần còn lại là 1 và 4 công việc nhân viên tham gia ít hơn 100.  *14. EmpJobLevel:*  Phạm vi cấp bậc công việc của nhân viên là từ 1 đến 5, hầu hết cấp bậc công việc của nhân viên là 1 & 2.  Số lượng cấp bậc công việc tối thiểu là 5, cũng như nhân viên cấp bậc công việc 3 & 4 là ít hơn 180.  *15. EmpJobSatisfaction:*  Hầu hết sự hài lòng về công việc của nhân viên là điểm 3 & 4 cũng như điểm 1 & 2 là ít hơn 250.  *16. NumCompaniesWorked:*  Số lượng nhân viên làm việc tối đa chỉ với một công ty.  Số lượng nhân viên làm việc ít hơn ở hơn 5 công ty.  *17. OverTime:*  Hầu hết các nhân viên không làm thêm giờ và số nhân viên làm thêm giờ khoảng 350 nhân viên |
| --- | --- |

*Bảng 3.3 Giải thích biểu đồ 1*

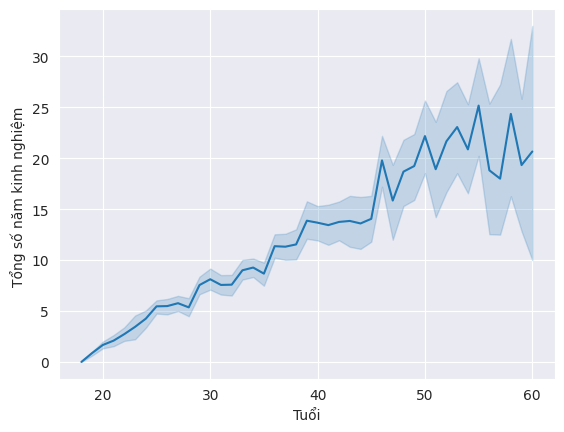


| 18. *EmpLastSalaryHikePercent:*  Phần lớn số lượng phần trăm tăng lương của nhân viên nằm trong khoảng từ 11 đến 14%. Tổng phạm vi tăng lương là từ 11 đến 25.  Dưới 80 nhân viên tăng lương gần đây nhất trong khoảng từ 15 đến 19.  19. *EmpRelationshipSatisfaction:*  Phạm vi sự hài lòng về mối quan hệ của nhân viên là từ 1 đến 4, phần lớn sự hài lòng về mối quan hệ là từ 3 đến 4 cũng như dưới 250 nhân viên có sự hài lòng về mối quan hệ là từ 1 đến 2.  20. *TrainingTimesLastYear:*  Khoảng thời gian đào tạo trong năm ngoái từ 0 đến 6, Hầu hết nhân viên tham gia đào tạo lần 2 và lần 3 và thời gian còn lại đều dưới 100.  21. *EmpWorkLifeBalance:*  Hầu hết cân bằng công việc và cuộc sống là 3, dưới 300 cân bằng công việc và cuộc sống là 2.  Cân bằng công việc và cuộc sống tối thiểu là 1, dưới 100 cân bằng công việc và cuộc sống là 4.  22. *ExperienceYearsInCurrentRole:*  Kinh nghiệm vai trò hiện tại tối đa của nhân viên là 18.  Hầu hết nhân viên có 2 năm kinh nghiệm trong vai trò hiện tại, sau đó nhân viên có 7 năm kinh nghiệm là dưới 180  Dưới 200 nhân viên không có kinh nghiệm trong vai trò hiện tại. | 23. *YearsSinceLastPromotion:*  Khoảng thời gian từ năm kể từ lần thăng chức gần nhất là từ 0 đến 15.  Hầu hết nhân viên không được thăng chức trong năm ngoái, ít hơn 300 nhân viên đã được thăng chức ít nhất 1 năm.  24. *YearsWithCurrManager:*  Hầu hết nhân viên có 2 năm làm việc với quản lý hiện tại và ít nhất 16 năm làm việc với quản lý hiện tại.  Khoảng thời gian từ 0 đến 17 và ít hơn 180 nhân viên có 7 năm làm việc với quản lý hiện tại.  25. *Attrition:*  Số lượng nhân viên tối đa chưa rời khỏi công ty và ít hơn 200 nhân viên sẽ rời khỏi công ty.  26. *PerformanceRating:*  Khoảng thời gian xếp hạng hiệu suất là từ 2 đến 4, hầu hết nhân viên có hiệu suất làm việc là 3 và ít hơn là 4 lần xếp hạng hiệu suất.  Dưới 200 nhân viên có 2 lần xếp hạng hiệu suất. |
| --- | --- |

*Bảng 3.4 Giải thích biểu đồ 2*

* **Phân tích hai biến**

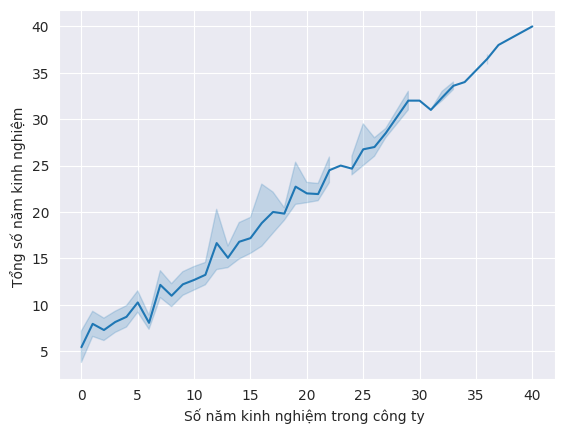
*1.Mối quan hệ giữa tuổi tác và số năm kinh nghiệm*



Nếu độ tuổi từ 18 đến 25 thì tổng số năm kinh nghiệm ít hơn 5 năm.

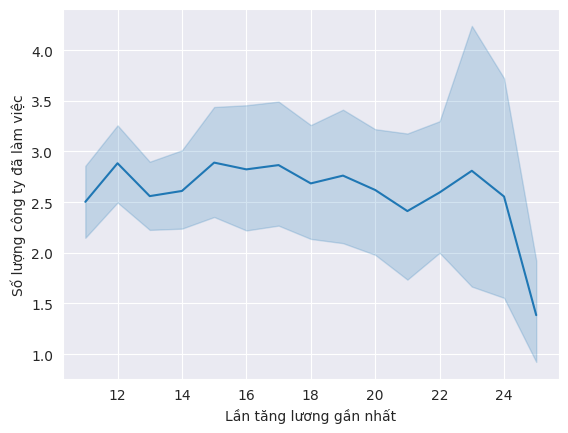
Độ tuổi tăng thì thời gian kinh nghiệm cũng tăng theo.

*2.Mối quan hệ giữa số năm kinh nghiệm và tổng số kinh nghiệm làm việc*



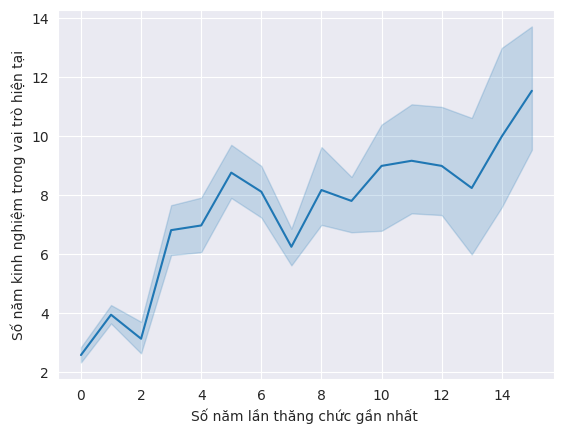
Nếu có kinh nghiệm làm việc tại công ty này một năm thì tổng kinh nghiệm làm việc cũng tăng lên.

*3.Mối quan hệ giữa lần tăng lương gần nhất của nhân viên và số lượng công ty đã làm việc*



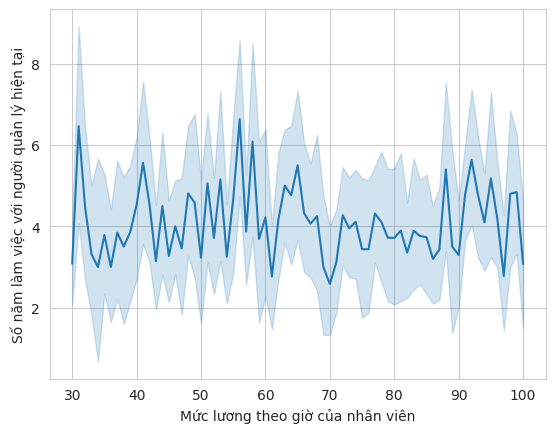
Nếu nhân viên làm việc cho 1 hoặc 2 công ty thì lần tăng lương gần nhất của nhân viên đó sẽ được tăng phần trăm.

*4.Mối quan hệ giữa số năm kể từ lần thăng chức gần nhất và số năm kinh nghiệm trong vai trò hiện tại*



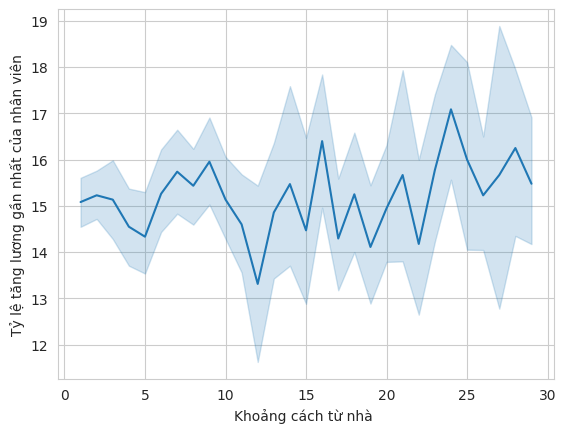
Trong biểu đồ trên, có thể thấy rõ năm kể từ lần thăng chức cuối cùng tăng lên khi nhân viên có nhiều kinh nghiệm hơn với vai trò hiện tại

*5.Mối quan hệ giữa mức lương theo giờ của nhân viên và số năm làm việc với người quản lý hiện tại*



Hầu hết mức lương theo giờ của nhân viên với người quản lý hiện tại đều không đổi, nhưng một số nhân viên có mức lương theo giờ tối đa với số năm ít hơn với người quản lý hiện tại.

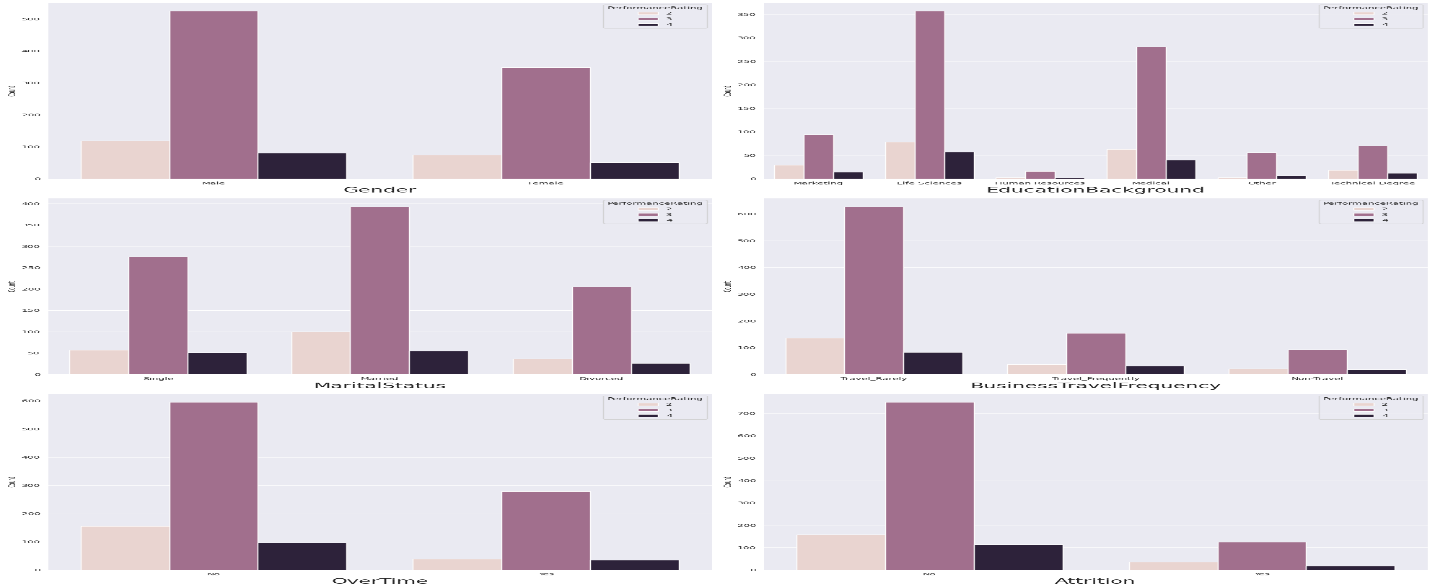
*6.Mối quan hệ giữa khoảng cách từ nhà và tỷ lệ tăng lương gần nhất của nhân viên*



Nếu khoảng cách từ nhà xa hơn thì mức tăng lương của nhân viên sẽ tăng nhẹ.

Khoảng cách từ nhà từ 10 đến 13 giờ thì mức tăng lương của nhân viên sẽ giảm.

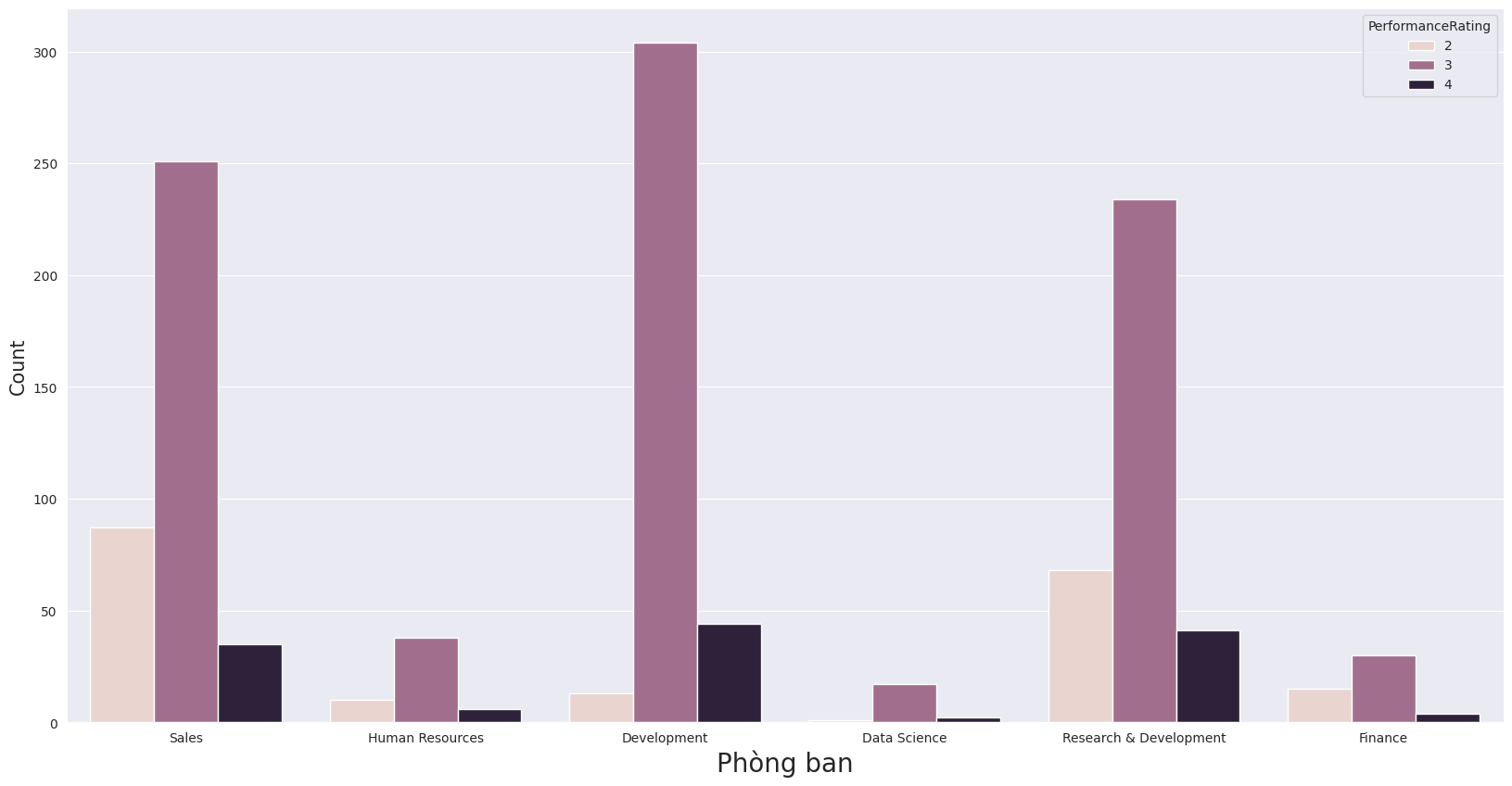
*-Phân tích đặc trưng phân loại với đánh giá hiệu suất*



| *7.Mối quan hệ giữa Gender và PerformanceRating*  Số lượng nam giới tối đa được đánh giá với 3 đánh giá hiệu suất cũng như ở nữ giới, hầu hết phụ nữ được đánh giá với 3 đánh giá hiệu suất.  Bốn đánh giá hiệu suất ở cả nam và nữ là ít hơn  *8. Mối quan hệ giữa EducationBackground và PerformanceRating*  Nhân viên có trình độ khoa học sự sống và giáo dục y tế tối đa được đánh giá với 3 đánh giá hiệu suất  Nhân viên có trình độ khoa học sự sống và giáo dục y tế cũng được đánh giá với 2 và 4 đánh giá hiệu suất hầu hết thời gian  *9. Mối quan hệ giữa MaritalStatusvà PerformanceRating*  Hầu hết nhân viên đã kết hôn được đánh giá với 3 đánh giá hiệu suất cũng như nhân viên có tình trạng hôn nhân độc thân và đã ly hôn được đánh giá với 3 đánh giá hiệu suất hầu hết thời gian.  Trong tình trạng hôn nhân đã kết hôn, nhân viên có 2 đánh giá hiệu suất cao hơn so với nhân viên độc thân hoặc đã ly hôn. | *10. Mối quan hệ giữa BusinessTravelFrequency và PerformanceRating*  Số lượng công tác tối đa hiếm khi nhân viên đạt được 3 đánh giá hiệu suất  Những nhân viên thường xuyên công tác này cũng được đánh giá 2, 3 và 4  *11. Mối quan hệ giữa OverTime và PerformanceRating*  Số lượng nhân viên không làm thêm giờ tối đa được đánh giá 3 đánh giá hiệu suất  Những nhân viên làm thêm giờ mà thời gian 2 & 4 đánh giá hiệu suất là không đổi.  *12. Mối quan hệ giữa Attrition và PerformanceRating*  Hầu hết nhân viên được đánh giá 3 đánh giá hiệu suất không có nghĩa là (Họ không rời công ty)  Những nhân viên có 3 đánh giá hiệu suất cũng rời công ty |
| --- | --- |

*Bảng 3.5 Giải thích biểu đồ 3*

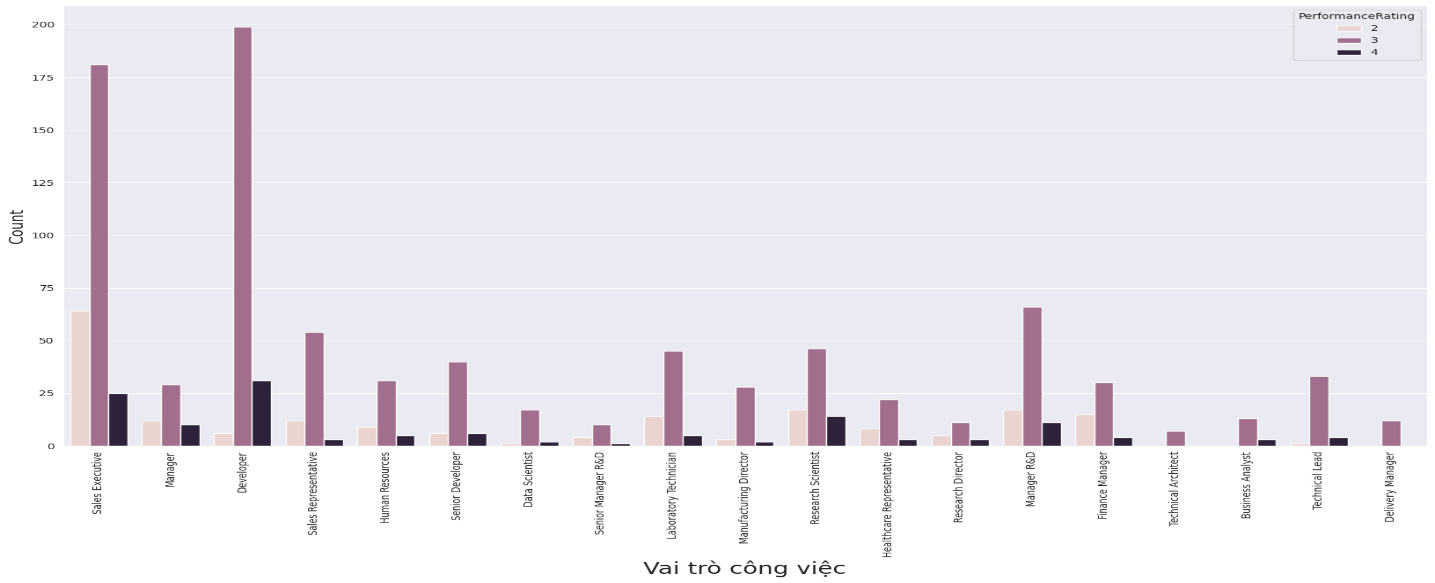
*13.Mối quan hệ giữa phòng ban nhân viên và đánh giá hiệu suất*



Hầu hết các nhân viên phòng Sales,Development,Research&Development có đánh giá hiệu suất là 3

Nhân viên phòng Data Science có đánh giá hiệu suất thấp.

*14.Mối quan hệ giữa vai trò công việc của nhân viên và đánh giá hiệu suất*



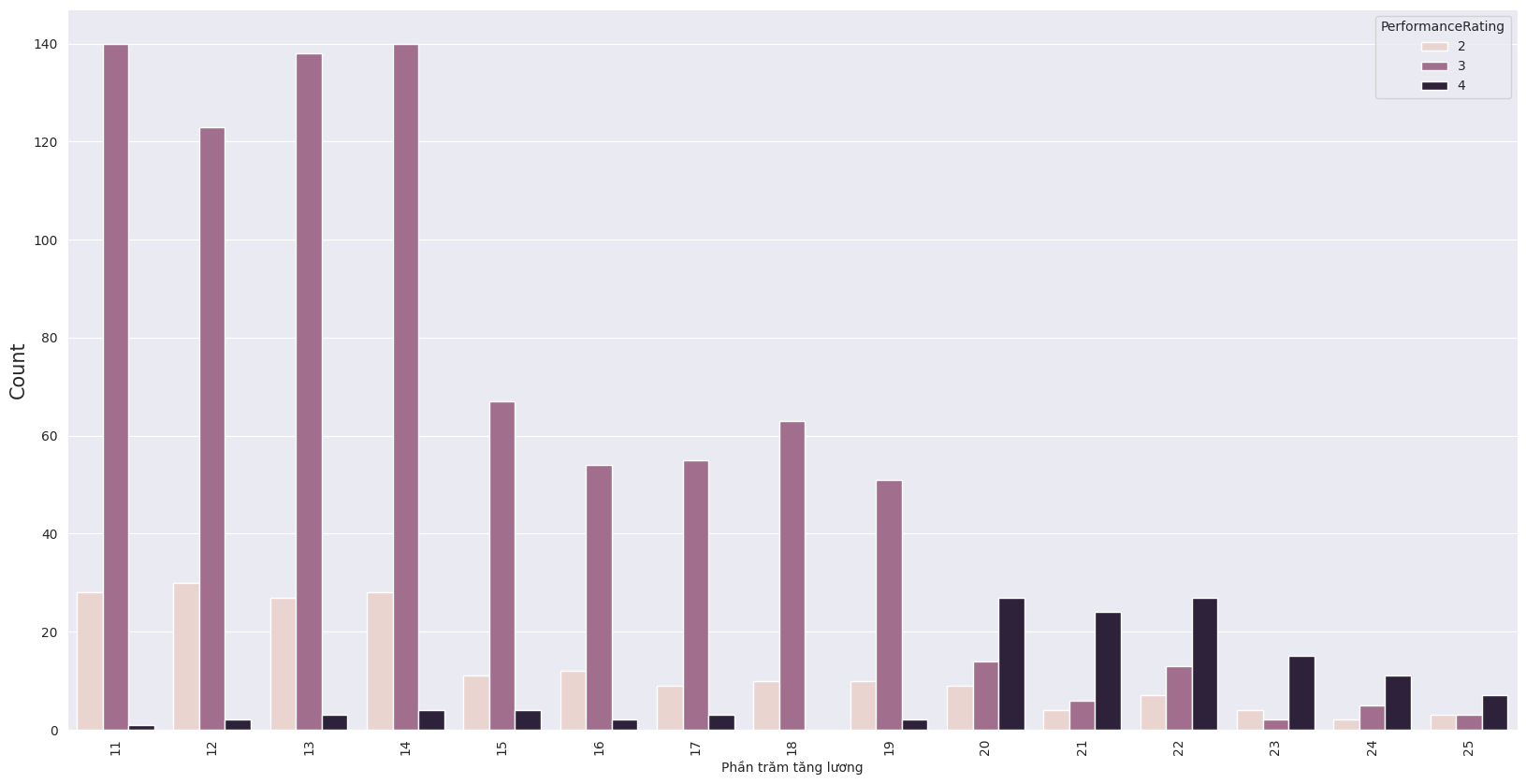
Xếp hạng hiệu suất của nhân viên trong vai trò quản lý bán hàng và phát triển phần lớn thời gian là 3.

Trong vai trò quản lý kiến ​​trúc kỹ thuật và giao hàng, nhân viên không được xếp hạng với xếp hạng hiệu suất là 2 & 4.

Trong vai trò quản lý bán hàng, cũng có ít hơn 60 nhân viên được xếp hạng với xếp hạng hiệu suất là 2.

Tất cả các vai trò công việc còn lại, nhân viên có xếp hạng hiệu suất là 3 hầu hết thời gian.

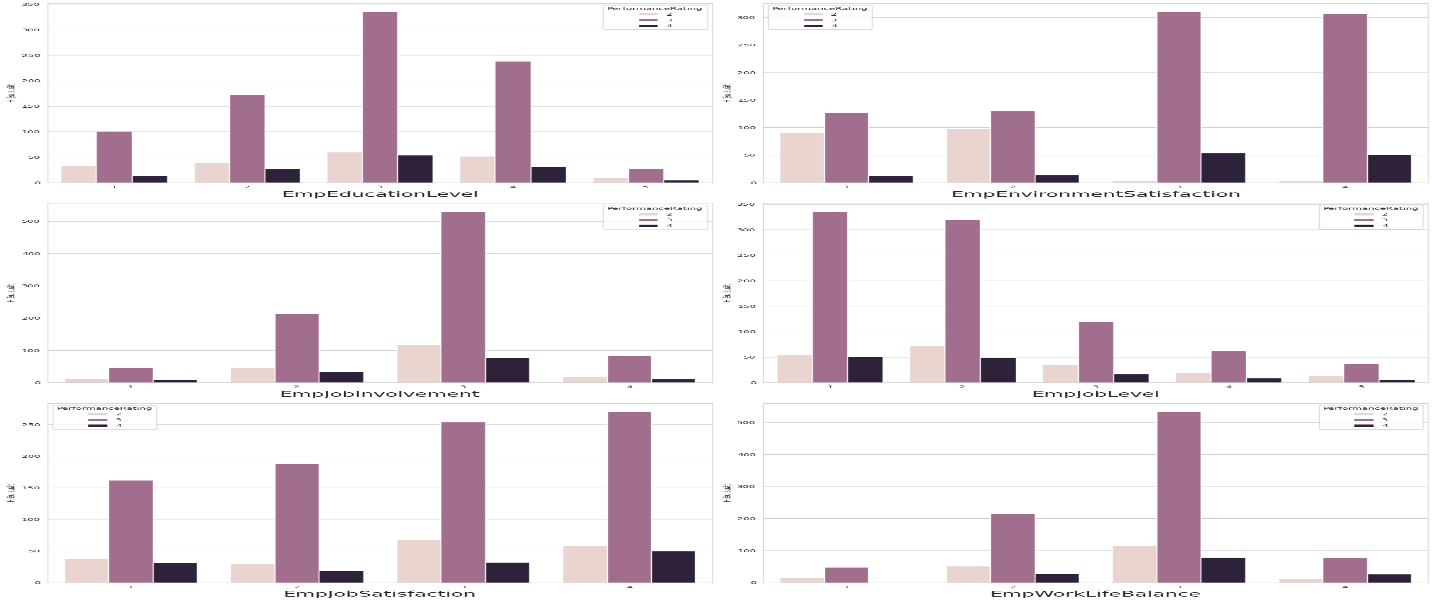
*15.Mối quan hệ giữa % tăng lương và đánh giá hiệu suất*



Phần trăm tăng lương nằm từ khoảng 11 đến 25% đa số hiệu suất làm việc ở mức 3

Khi ở mức 20% trở lên thì hiệu suất làm việc ở mức 4 có tăng lên

*-Phân tích đặc trưng số với đánh giá hiệu suất*

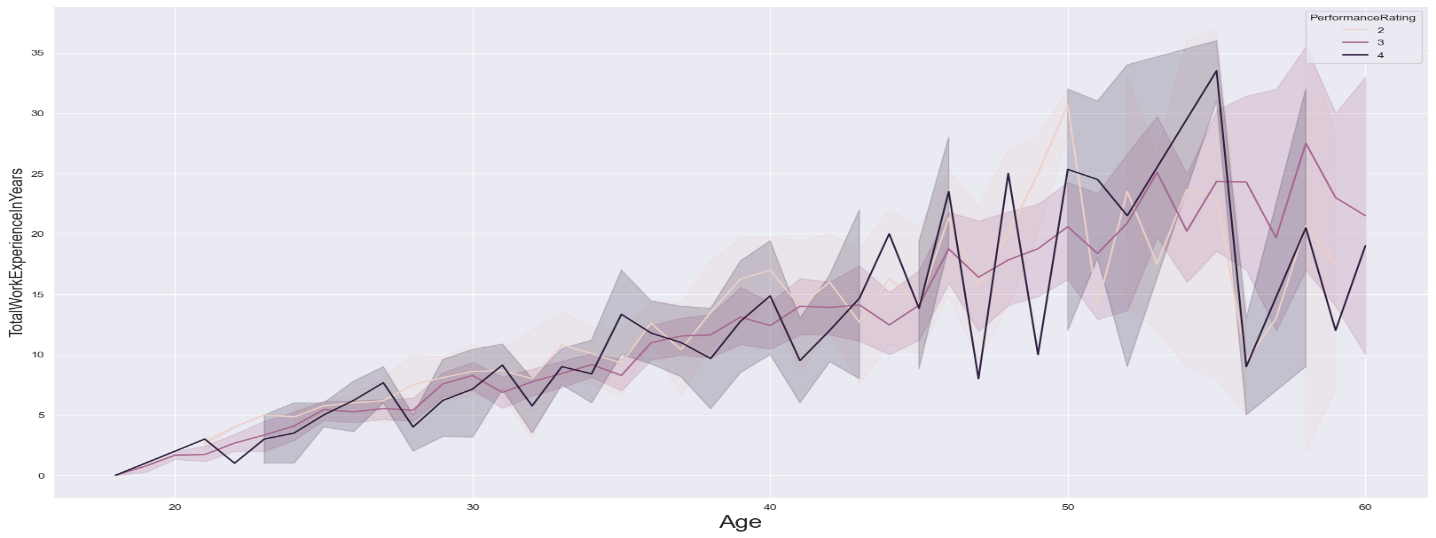


| *16. Mối quan hệ giữa EmpEducationLevel và PerformanceRating*  Trình độ học vấn của nhân viên là [3,4 & 2], đa số hiệu suất là 3  Nếu đánh giá hiệu suất là 2 & 4 thì tất cả nhân viên có trình độ học vấn đều dưới 60.  *17. Mối quan hệ giữa EmpEnvironmentSatisfaction và PerformanceRating*  Số lượng tối đa của thống kê môi trường nhân viên là 3 và 4 với 3 đánh giá hiệu suất.  Nếu sự hài lòng về môi trường nhân viên là 1 & 2 thì lúc đó nhân viên tối đa có 2 đánh giá hiệu suất.  Nếu sự hài lòng về môi trường nhân viên tăng thì lúc đó 4 đánh giá hiệu suất cũng tăng.  *18. Mối quan hệ giữa EmpJobInvolvement và PerformanceRating*  Những nhân viên có mức độ tham gia công việc là 2 & 3 thì nhân viên này thường có mức đánh giá hiệu suất là 3  Nhân viên có mức đánh giá hiệu suất là 2 & 4 thì ít hơn trong mọi mức độ tham gia công việc. | *19. Mối quan hệ giữa EmpJobLevel và PerformanceRating*  Nếu mức độ công việc của nhân viên là 1, 2 & 3 thì hầu hết nhân viên đều có mức đánh giá hiệu suất là 3  Ở mọi mức độ công việc, mức đánh giá hiệu suất là 2 & 4 của tất cả nhân viên đều thấp hơn 60.  *20. Mối quan hệ giữa EmpJobSatisfaction và PerformanceRating*  Mức độ hài lòng với công việc tối đa là 4 & 3 thì hầu hết nhân viên đều có mức đánh giá hiệu suất là 3.  Trong tất cả các mức độ hài lòng công việc 2 & 4, xếp hạng hiệu suất đều thấp hơn 55 nhân viên  *21. Mối quan hệ giữa EmpWorkLifeBalance và PerformanceRating*  Nếu sự cân bằng giữa công việc và cuộc sống của nhân viên là 3 thì số lượng nhân viên có xếp hạng hiệu suất là 3 là tối đa.  Nếu sự cân bằng giữa công việc và cuộc sống là 1 thì không ai có xếp hạng hiệu suất là 4. |
| --- | --- |

*Bảng 3.6 Giải thích biểu đồ 4*

* **Phân tích đa biến**

1.*Kiểm tra mối quan hệ giữa Age và TotalWorkExperienceInYears làm việc đối với PerformanceRating*

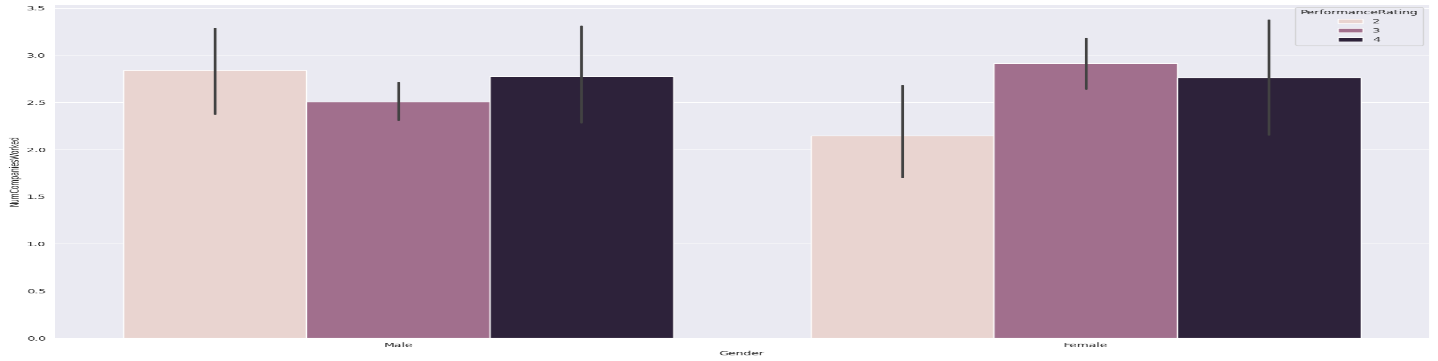


Nếu độ tuổi và tổng số kinh nghiệm tăng lên thì hầu hết đánh giá hiệu suất của nhân viên là 4.

Nhân viên có độ tuổi thấp và số năm kinh nghiệm tối thiểu cũng được xếp hạng 4 về hiệu suất.

Ở độ tuổi 50 và 30 năm kinh nghiệm, nhân viên có 2 mức đánh giá thành tích.

*2.Kiểm tra mối quan hệ giữa Gender & NumCompaniesWorked đã làm việc đối với PerformanceRating*

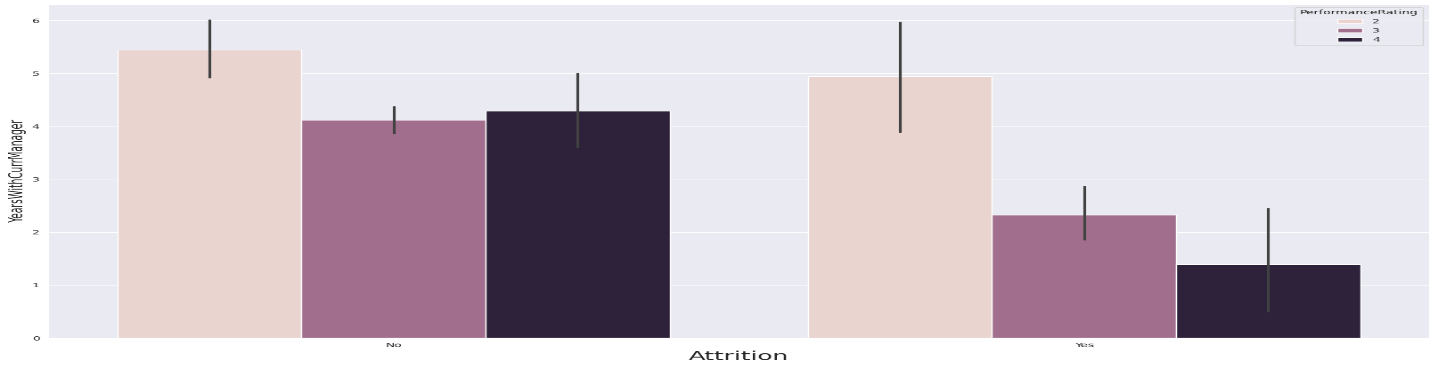


Nếu giới tính là nam và không có công ty nào làm việc thì thời điểm đó hầu hết xếp hạng hiệu suất là 2 & 4.

Nếu nữ làm việc ở ít công ty hơn thì thời gian đánh giá hiệu suất là 2.

Xếp hạng hiệu suất 3 & 4 ở phụ nữ tăng nếu số lượng công ty làm việc tăng lên.

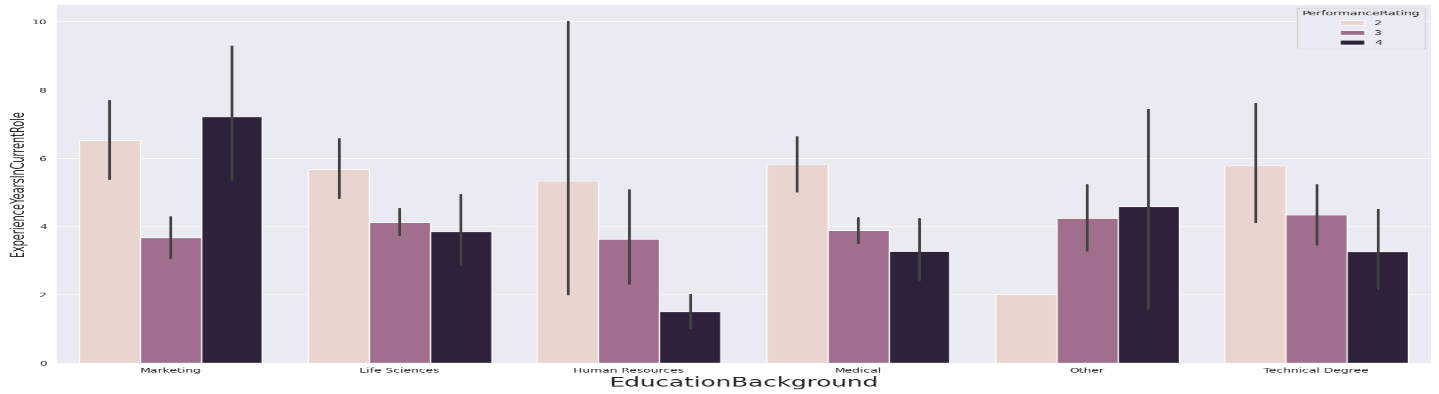
*3.Kiểm tra mối quan hệ giữa Attrition & NămWithCurrManager đối với PerformanceRating*



Nếu năm quản lý hiện tại của nhân viên tăng lên thì xếp hạng hiệu suất theo thời gian là 2 và 50-50% khả năng nhân viên rời công ty.

xếp hạng hiệu suất làm việc của nhân viên đó là 4 và năm quản lý hiện tại tăng lên vào thời điểm hầu hết nhân viên không rời công ty.

*4.Kiểm tra mối quan hệ giữa EducationBackground và ExperienceYearsInCurrentRole đối với PerformanceRating*

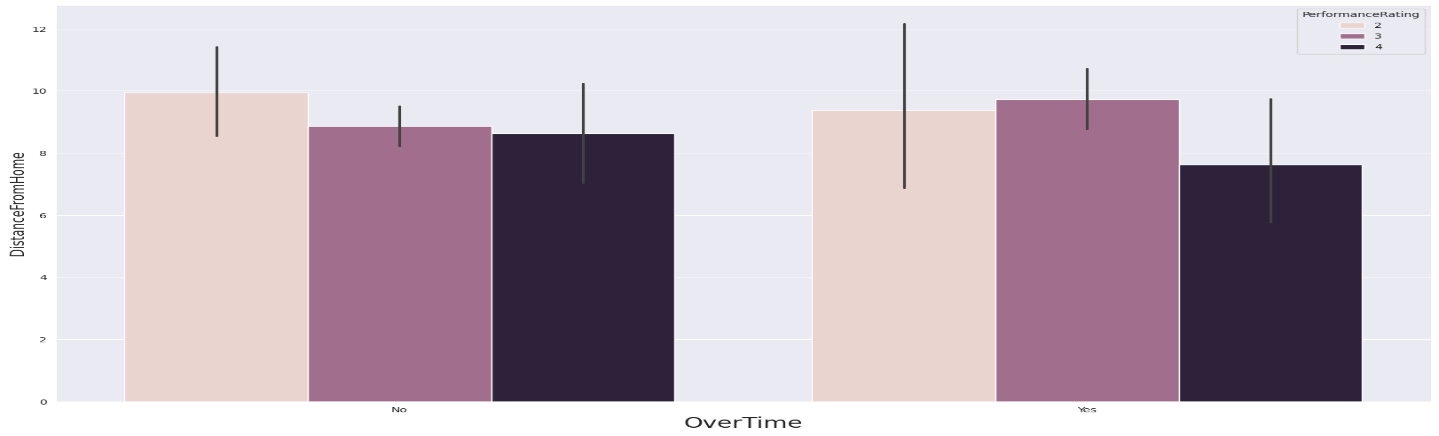


Nếu kinh nghiệm của vai trò hiện tại là nền tảng giáo dục tiếp thị, nhân viên có 4 xếp hạng hiệu suất hầu hết thời gian

Ở tất cả các trình độ học vấn có kinh nghiệm làm việc hiện tại lớn hơn 5 thì đánh giá hiệu suất của nhân viên là 2.

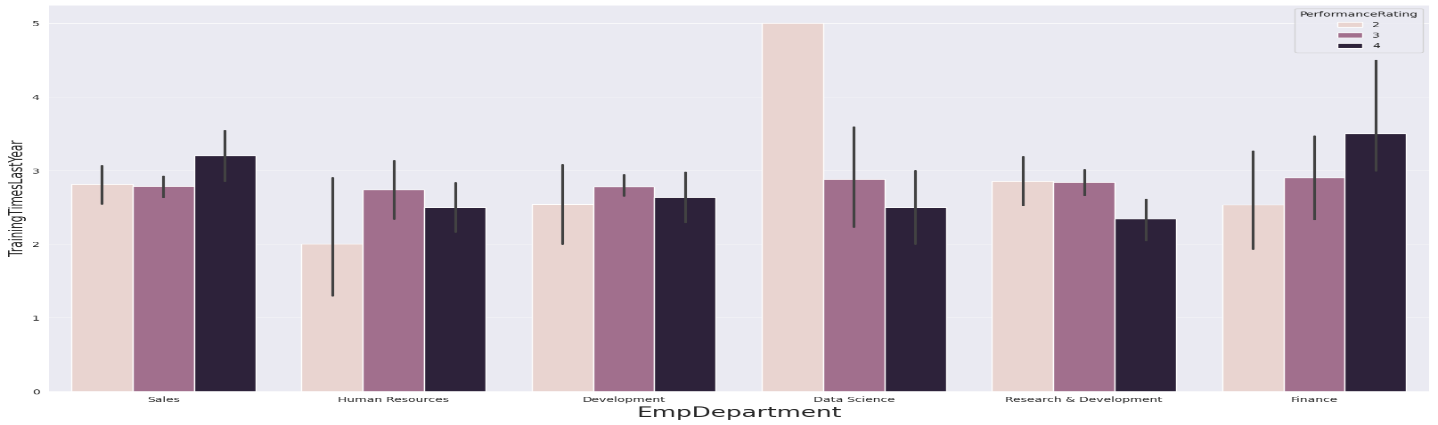
Nhân viên có trình độ học vấn về khoa học đời sống, y tế, khác và trình độ kỹ thuật có dưới 4 năm kinh nghiệm ở vai trò hiện tại và xếp hạng hiệu suất thời gian là 3 & 4.

*5.Kiểm tra mối quan hệ giữa OverTime và DistanceFromHome đối với PerformanceRating*



Nếu khoảng cách từ nhà tăng lên thì đánh giá hiệu suất thời gian sẽ không đổi với việc nhân viên làm thêm giờ cũng như không làm thêm giờ.

*6.Kiểm tra mối quan hệ giữa EmpDepartment & TrainingTimesLastYear đối với PerformanceRating*



Nếu nhân viên từ bộ phận khoa học dữ liệu và có thời gian đào tạo nhiều nhất trong năm ngoái thì xếp hạng hiệu suất là 2 thời gian tối đa.

Các bộ phận khác đều là nhân viên được đào tạo dưới 3 lần trong năm ngoái thì xếp hạng hiệu suất là không đổi.

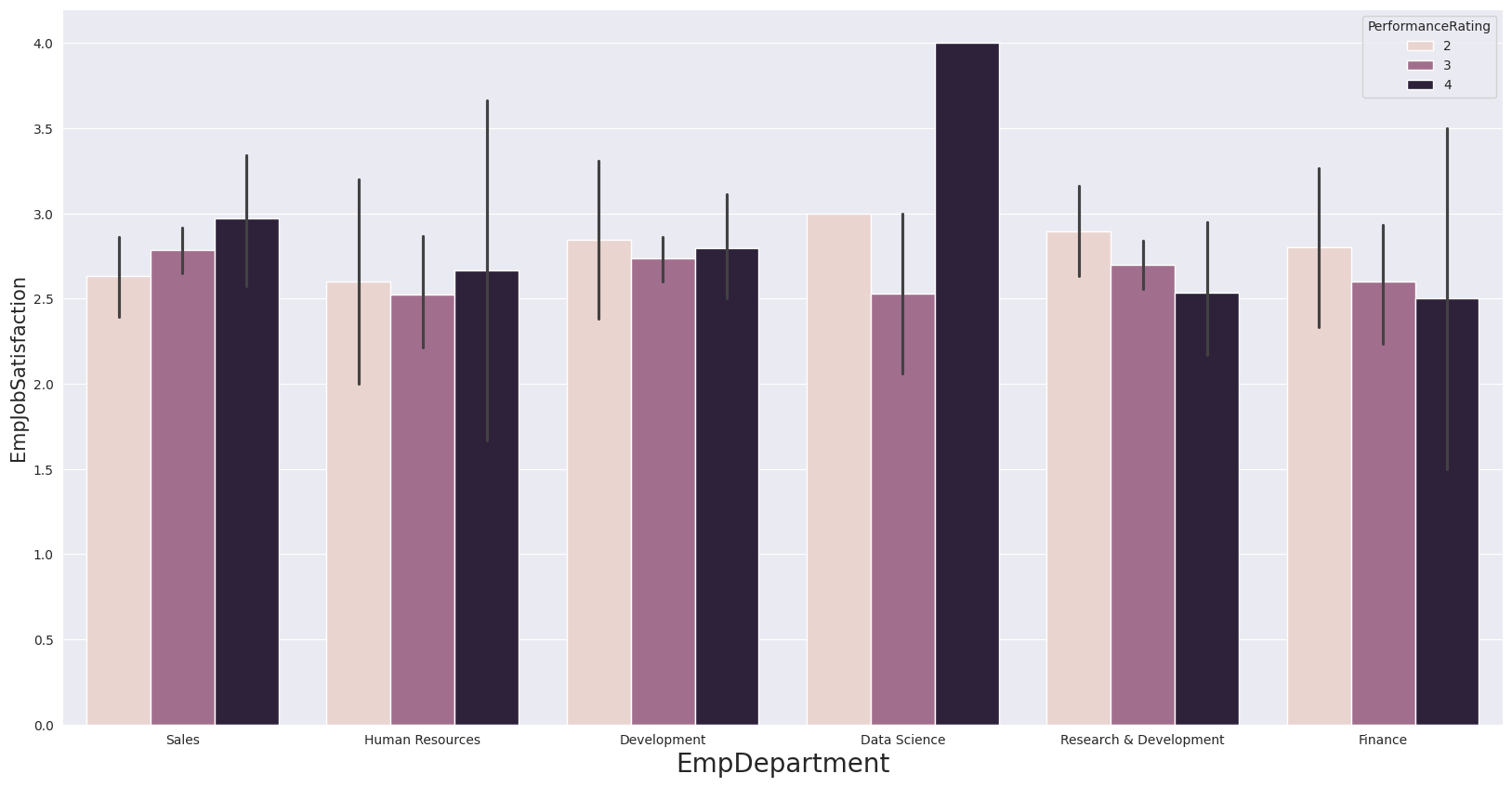
*7.Kiểm tra mối quan hệ giữa Attrition & NumCompaniesWorked đối với PerformanceRating*



Nếu số lượng nhân viên của công ty đã làm việc tăng lên thì thời điểm đó hầu hết nhân viên rời công ty và xếp hạng hiệu suất của họ là 2

Những nhân viên đó đã làm việc ở ít công ty hơn và họ không rời công ty và xếp hạng hiệu suất của họ là không đổi.

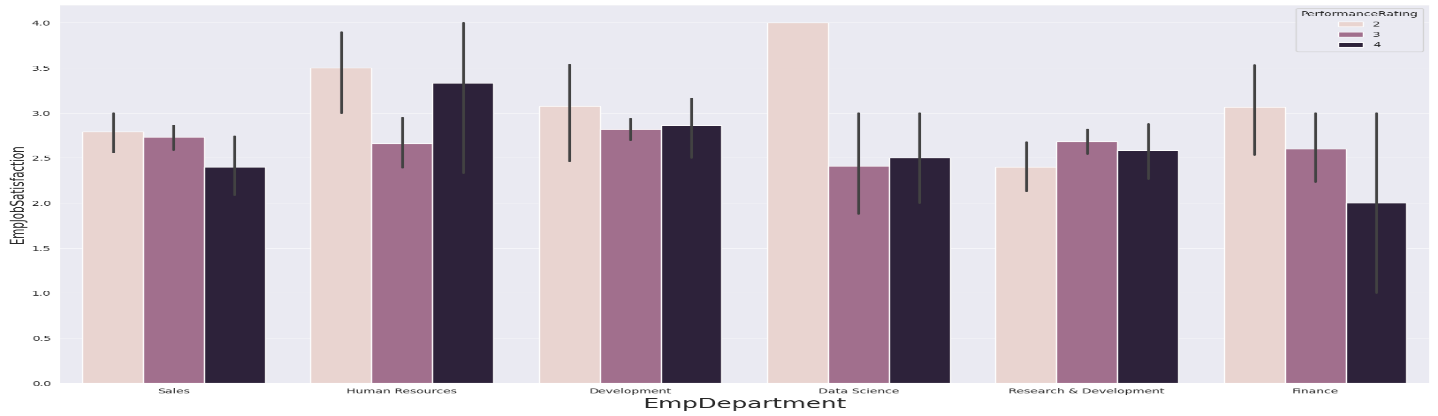
*8.Kiểm tra mối quan hệ giữa EmpJobSatisfaction & EmpDepartment đối với PerformanceRating*



Hầu hết mức độ hài lòng về công việc của các phòng ban nằm ở khoảng 1 đến 3

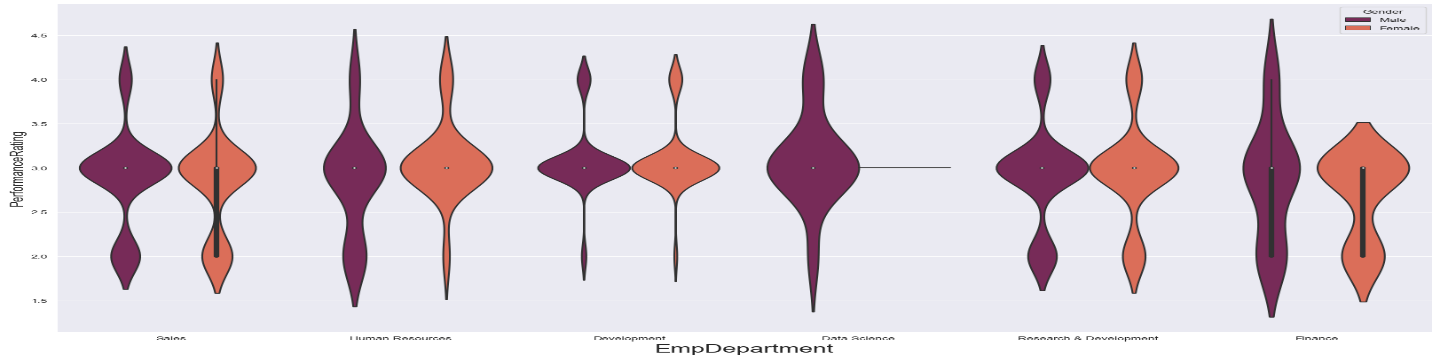
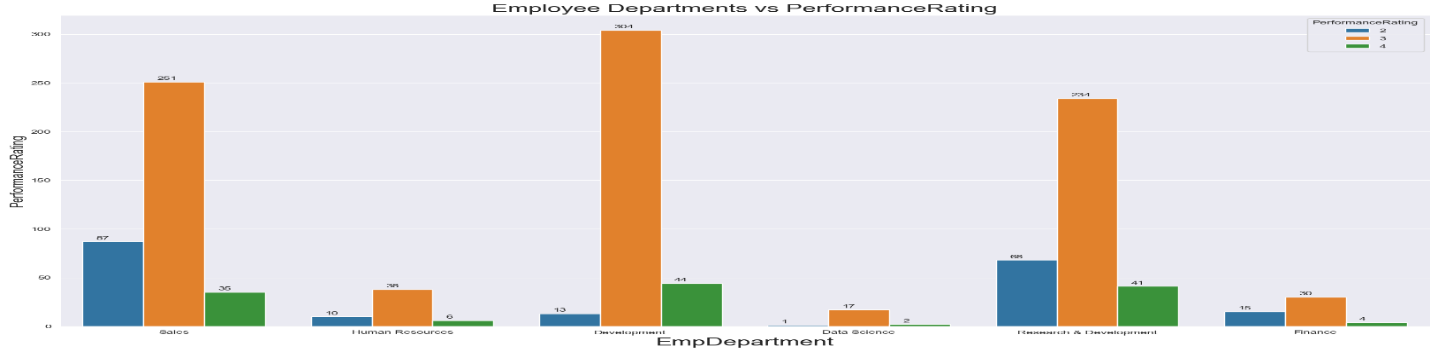
Phòng ban khoa học dữ liệu nổi trội hơn mức độ hài lòng về công việc là 4

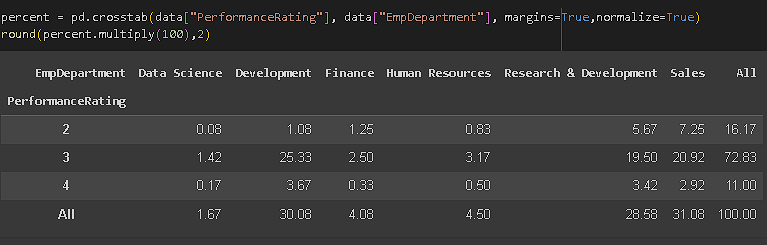
*9.Kiểm tra mối quan hệ giữa EmpRelationshipSatisfaction, & EmpDepartment đối với PerformanceRating*



Hầu hết các phòng ban đều nằm trong khoảng từ 1 đến 3

*-Đánh giá hiệu suất phòng ban*



*Hình 3.7. Tỷ lệ hiệu suất của các phòng ban*

| Từ 2 biểu đồ trên và bảng tỷ lệ ở trên, chúng em có thể kết luận rằng bộ phận **Developmen**t đang được xếp hạng hiệu suất cao  **Sales**: Xếp hạng hiệu suất cấp 3 cao hơn ở bộ phận bán hàng. Hiệu suất của nam được đánh giá cao hơn một chút so với nữ.  **Human Resources**:Phần lớn nhân viên nằm dưới năng suất cấp 3. Các nữ nhân viên phòng nhân sự thực hiện rất tốt công việc của mình.  **Development:** Số lượng nhân viên tối đa là người thực hiện cấp 3. Hiệu suất dựa trên giới tính gần như giống nhau đối với cả hai. | **Data Science**: Hiệu suất cấp 3 trung bình cao nhất là ở bộ phận khoa học dữ liệu. Nhân viên nam đang làm tốt ở bộ phận này.  **Research & Development**: Bộ phận R&D có những nhân viên nữ giỏi trong công việc.  **Finance:**  Các nhân viên nữ đang làm việc tốt |
| --- | --- |

*Bảng 3.7 Giải thích biểu đồ 6*

Từ phân tích trên:

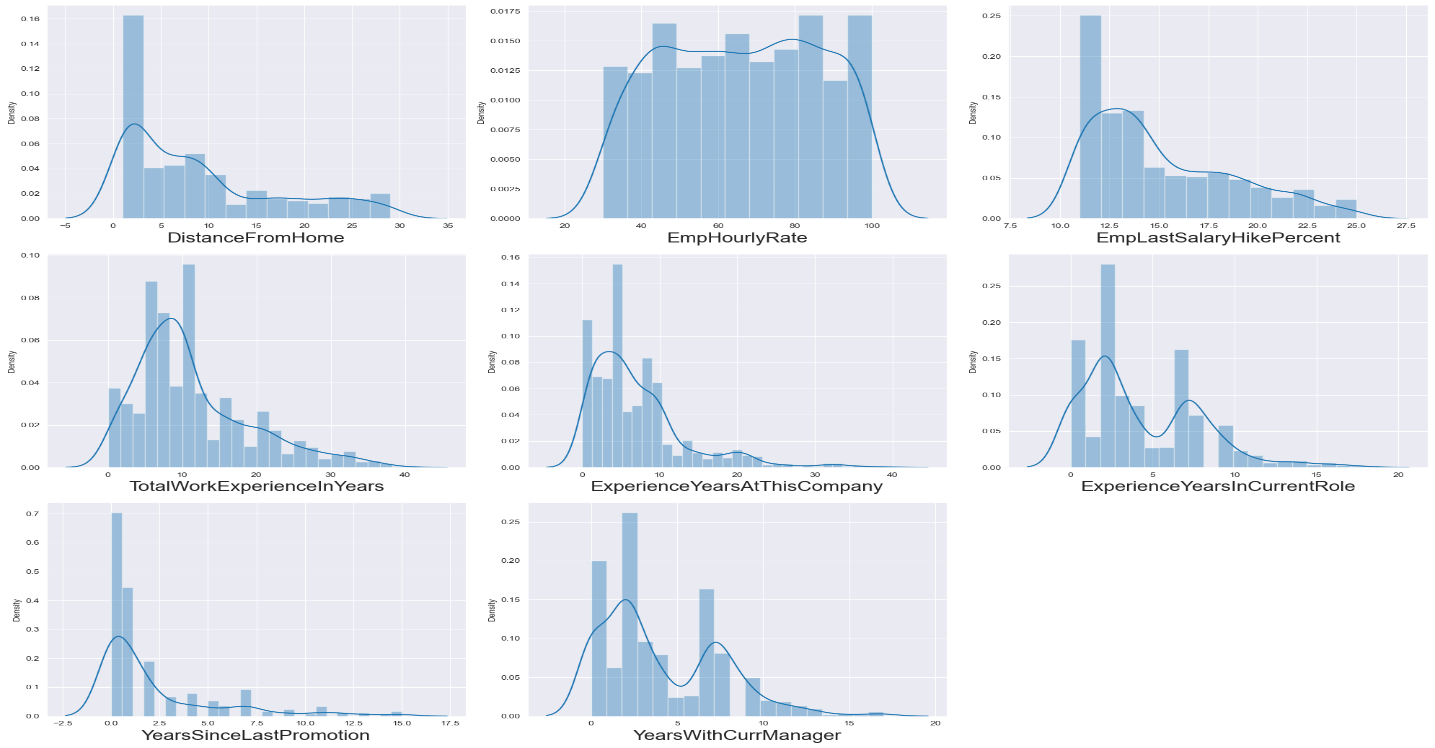
***Em thấy được các yếu tố ảnh hưởng đến hiệu suất làm việc của nhân viên là:***

1. Emp Environment Satisfaction
2. Emp Last Salary Hike Percent
3. Emp Work Life Balance

***-Đề xuất cải thiện hiệu suất của nhân viên:***

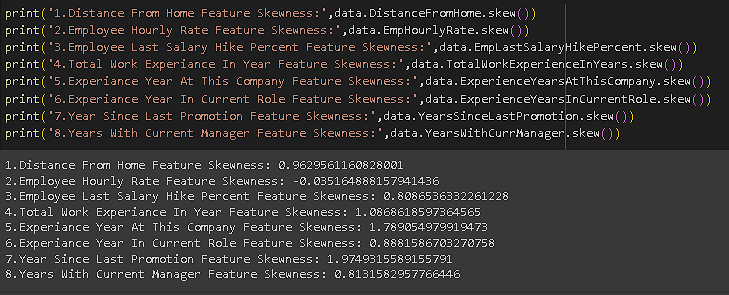
* Hiệu suất chung của nhân viên có thể đạt được thông qua sự hài lòng về môi trường của nhân viên. Công ty cần tập trung nhiều hơn vào sự hài lòng về môi trường của nhân viên.
* Việc tăng lương sẽ thúc đẩy nhân viên làm việc tốt.
* Thăng chức cho nhân viên sau mỗi 6 tháng
* Cải thiện sự cân bằng giữa công việc và cuộc sống của nhân viên, điều này ảnh hưởng đến xếp hạng hiệu suất.
* Khi tuyển dụng cho bộ phận nhân sự, hãy xem xét các ứng viên nữ có hiệu suất tốt hơn so với nam giới.
* Bộ phận Development và Sales có hiệu suất chung cao hơn so với các bộ phận còn lại. Trong khi một số nhân viên đưa ra phản hồi như thấp & trung bình từ đặc trưng EmpJobSatisfaction & EmpRelationshipSatisfaction, những nhân viên như vậy có hiệu suất cao nhiều hơn về số lượng. Vì vậy, công ty nên tập trung vào họ.

*-Kiểm tra phân bố của dữ liệu*



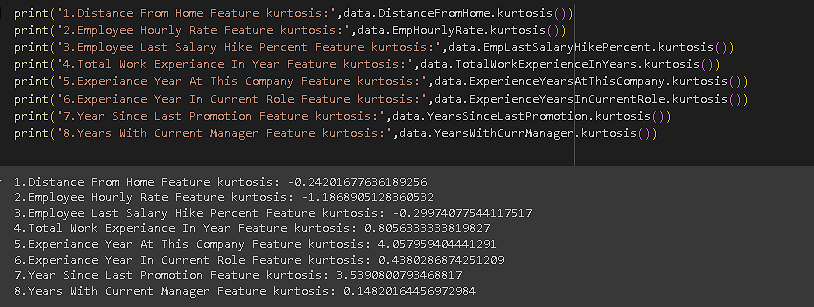
*Hình 3.8. Phân bố dữ liệu*

*-Độ lệch*



*Hình 3.9.Độ lệch*

Ta quan sát ở hình trên ta thấy có 2 đặc trưng có độ lệch dương khá cao so với phân phối chuẩn là Experiance Year At This Company và Year Since Last Promotion Feature em sẽ xử lý 2 đặc trưng này ở phần xử lý dữ liệu

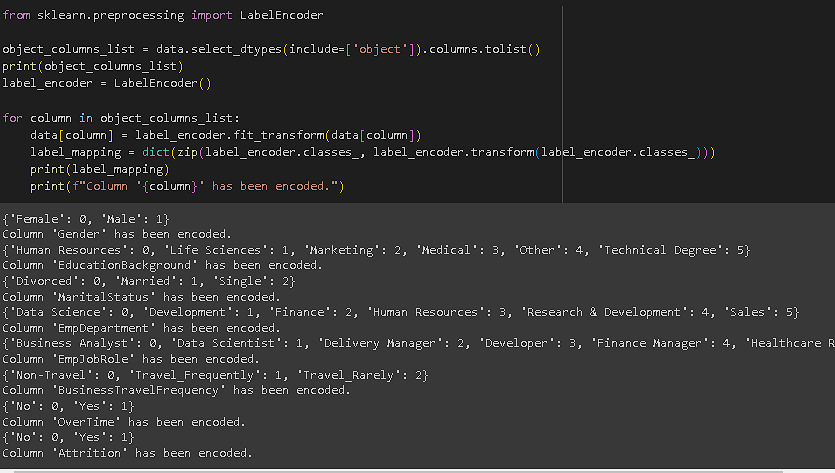


*Hình 3.10.Độ nhọn*

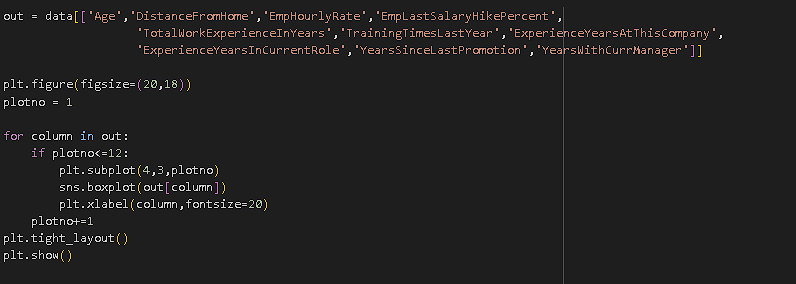
Ta quan sát ở hình trên ta thấy có 2 đặc trưng có phân phối đuôi rất dày so với phân phối chuẩn là Experiance Year At This Company và Year Since Last Promotion Feature em sẽ xử lý 2 đặc trưng này ở phần xử lý dữ liệu

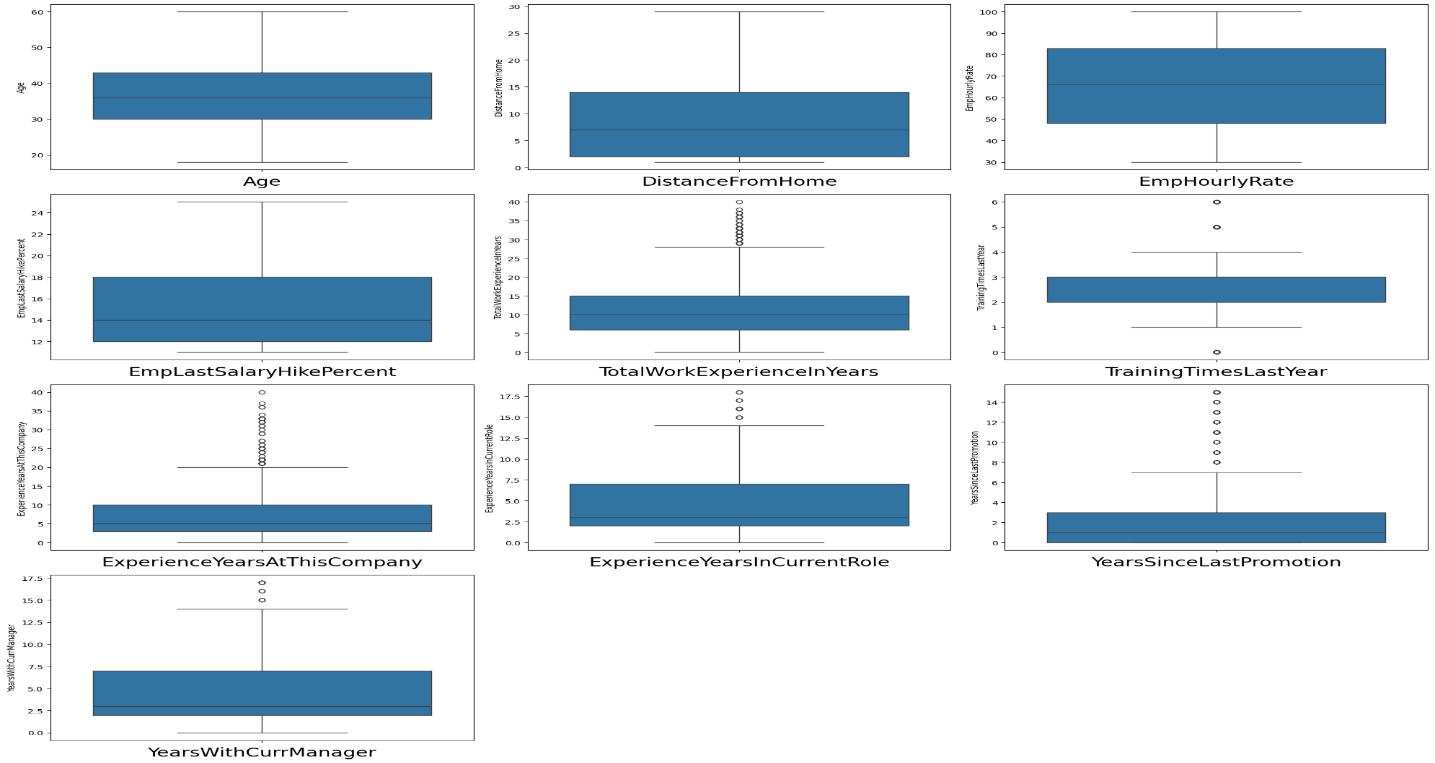
* 1. **Xử lý dữ liệu ,xây dựng mô hình**

-*Chuyển đổi đặc trưng phân loại thành số*



*-Kiểm tra giá trị ngoại lệ và sai số giá trị ngoại lệ*

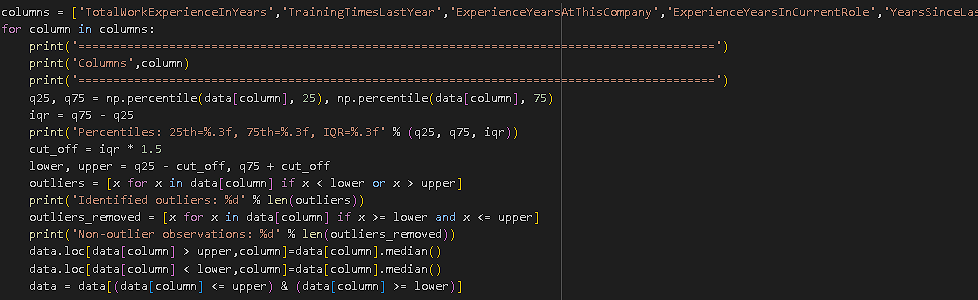


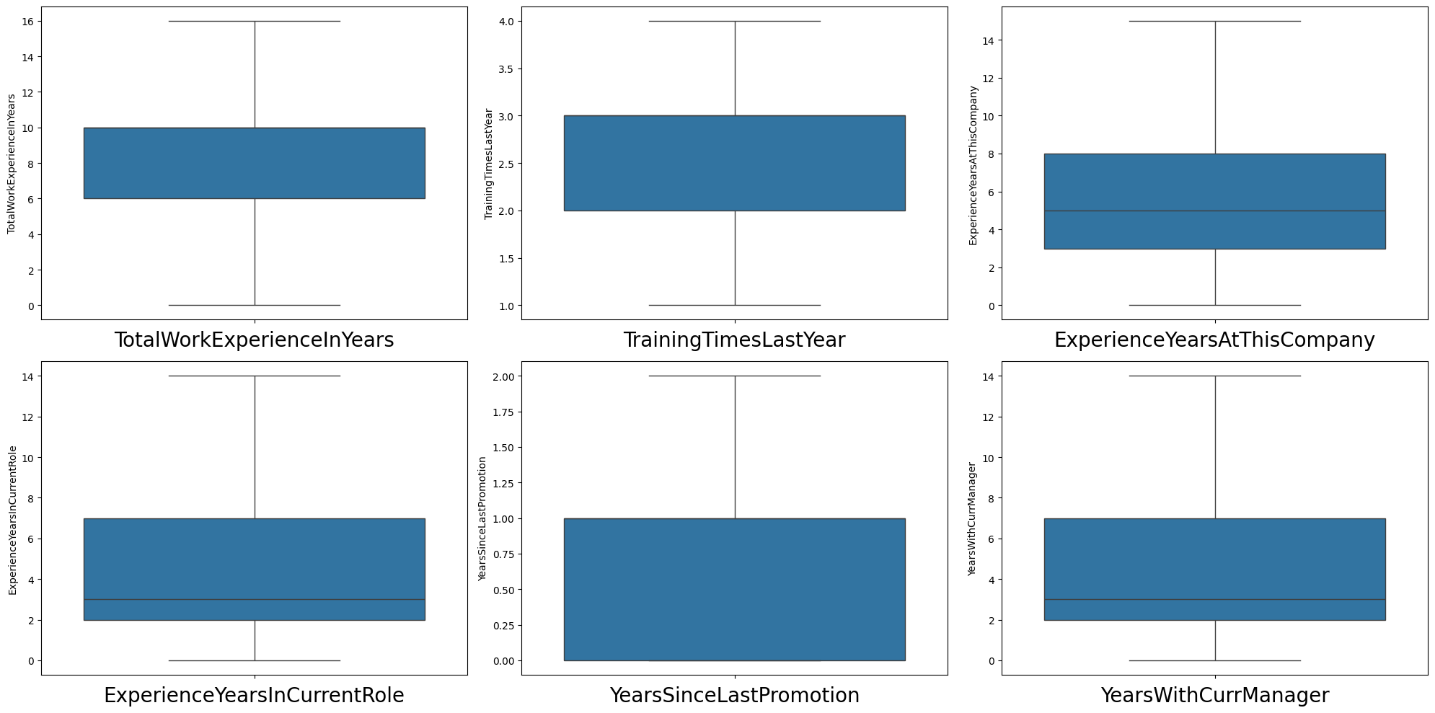


*Hình 3.10. Kết quả kiểm tra ngoại lệ*

*-Xử lý giá trị ngoại lệ*

Em sử dụng IQR để loại bỏ dữ liệu ngoại lai





*Hình 3.11. Kết quả kiểm tra sau khi sử lý ngoại lệ*

Sau khi loại bỏ ngoại lai em có có kiểm tra độ lệch và độ nhọn của 2 đặc trưng ExperienceYearsAtThisCompany và YearsSinceLastPromotion

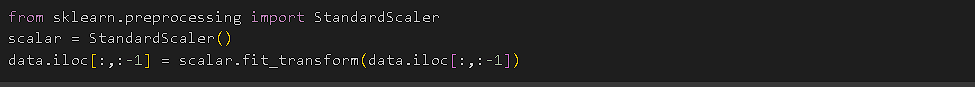
Thì em thấy cho ra kết quả trong khoản chấp nhận được nên em sẽ không chuyển đổi đặc trưng lại mà sẽ tiến hành vào xử lý dữ liệu để train mô hình phân loại năng suất nhân viên



*Hình 3.2. Kết quả kiểm tra độ lệch độ nhọn*

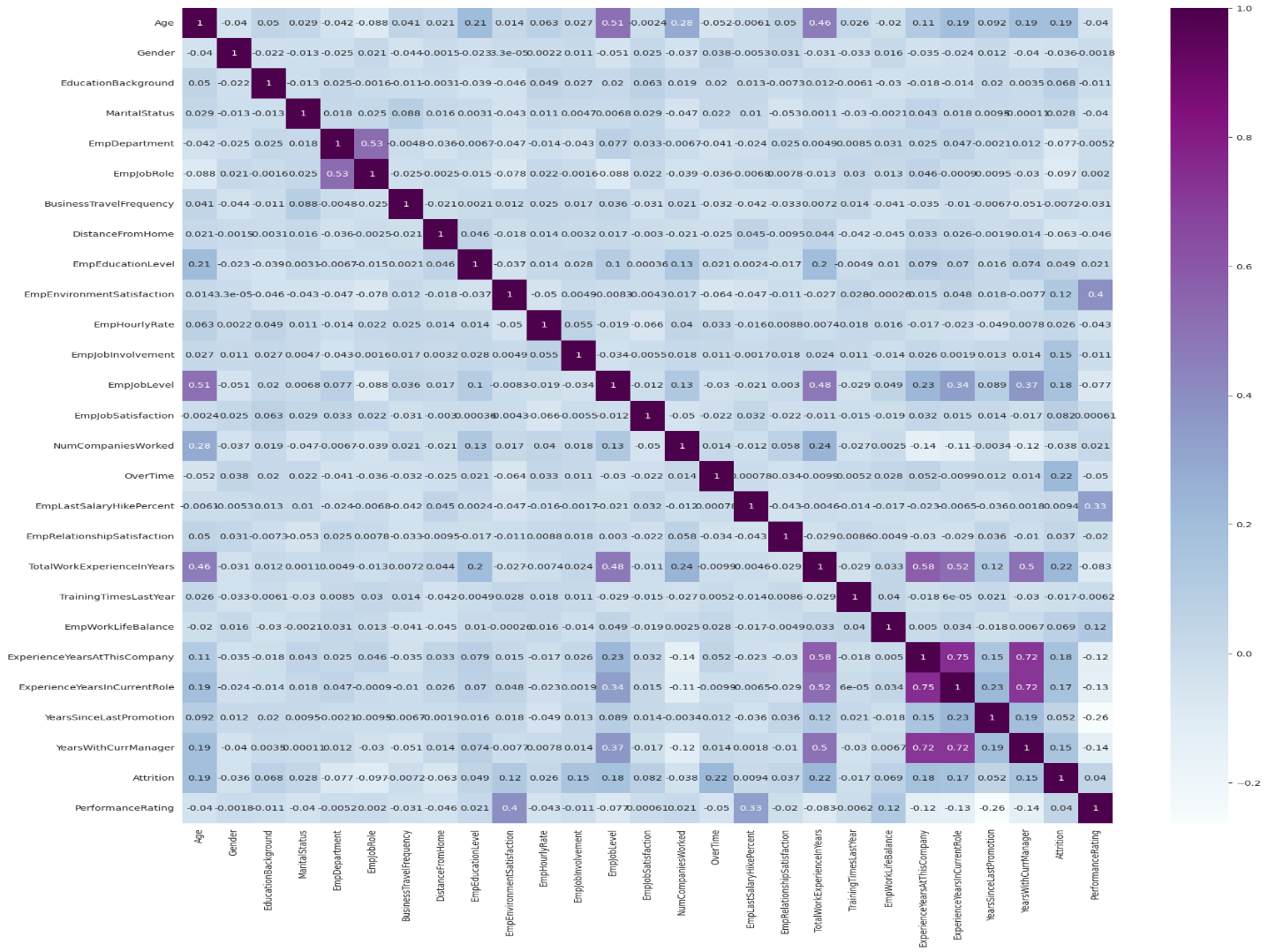
*-Chuẩn hóa dữ liệu*

Sau khi thử chuẩn hóa bằng nhiều cách thì em sử dụng StandardScaler để chuẩn hóa dữ liệu số về một khoảng với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1để cải thiện chất lượng đầu ra của mô hình



*-Lựa chọn đặc trưng*

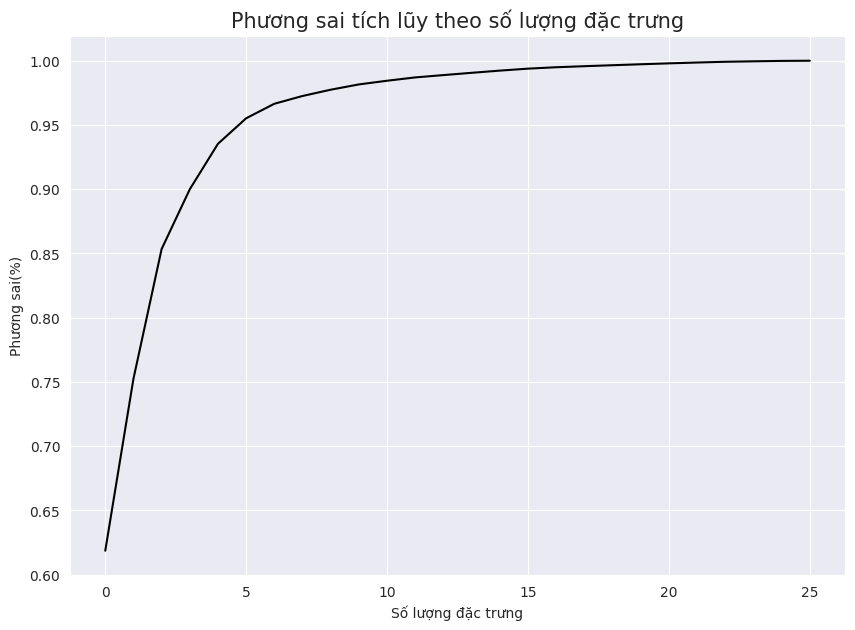
Ma trận tương quan



Ma trận tương quan cho thấy mức tương quan của các đặc trưng không có sự tương quan nhiều đến biến mục tiêu,em sẽ sử dụng thuật toán PCA để giảm chiều dữ liệu bởi vì dữ liệu của em có tới 26 đặc trưng,nên em nghĩ PCA sẽ thích hợp làm điều này

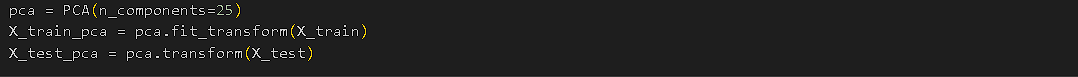
*-Phân tích thành phần chính*

Xác định số tính năng



Nhìn vào biểu đồ trên ta thấy số lượng đặc trưng mà thuật toán PCA cho thấy là 25 đặc trưng

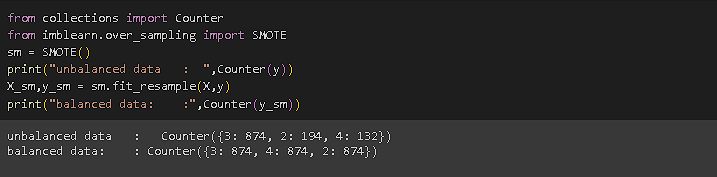
Chuyển đổi giảm chiều dữ liệu



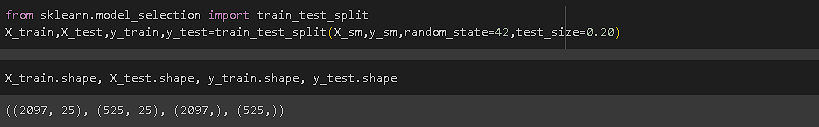
Cắt đặc trưng



Cân bằng đặc trưng mục tiêu

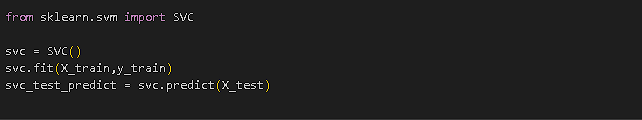


Cắt dữ liệu để train



*-Model*

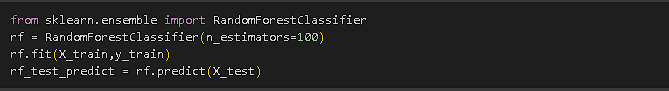
Support Vector Machine



Kết quả

|  | precision | recall | f1-score |
| --- | --- | --- | --- |
| 2 | 0.97 | 0.96 | 0.96 |
| 3 | 0.9 | 0.97 | 0.93 |
| 4 | 1.0 | 0.94 | 0.97 |

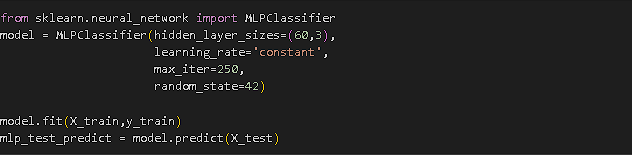
Random Forest



Kết quả

|  | precision | recall | f1-score |
| --- | --- | --- | --- |
| 2 | 0.95 | 0.91 | 0.93 |
| 3 | 0.87 | 0.93 | 0.90 |
| 4 | 0.98 | 0.95 | 0.96 |

Artificial Neural Network [MLP Classifier]



Kết quả

|  | precision | recall | f1-score |
| --- | --- | --- | --- |
| 2 | 0.97 | 0.97 | 0.97 |
| 3 | 0.94 | 0.96 | 0.95 |
| 4 | 0.99 | 0.97 | 0.98 |
|  |  |  |  |

Kết quả trực quan hóa



| Model | Testing accuracy |
| --- | --- |
| Support Vector Machine | 95.61% |
| Random Forest Classifier | 93.14% |
| MLP Classifier | 96.76% |

*Bảng 3.9.Kết quả accuracy*

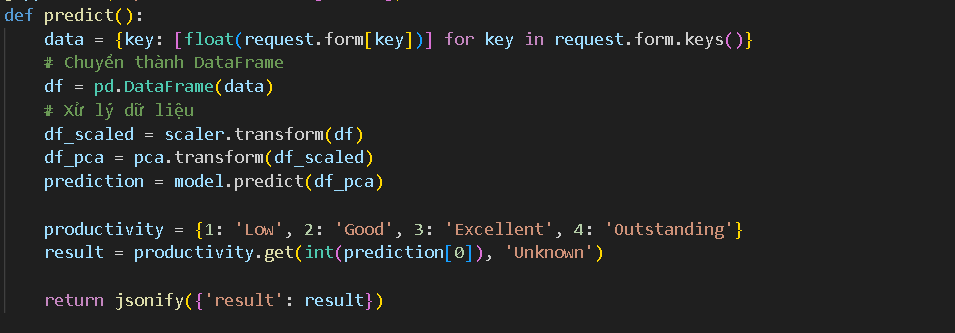
Em sẽ chọn mô hình **MLP Classifier** để làm mô hình dự đoán hiệu suất nhân viên để phục vụ trong quá trình quản lý hiệu suất nhân viên và tuyển dụng nhân viên

# CHƯƠNG 5 THỰC NGHIỆM MÔ HÌNH

Em sẽ chọn mô hình **MLP Classifier** để thực hiện xây dựng chương trình dự đoán năng suất nhân viên bằng Flask

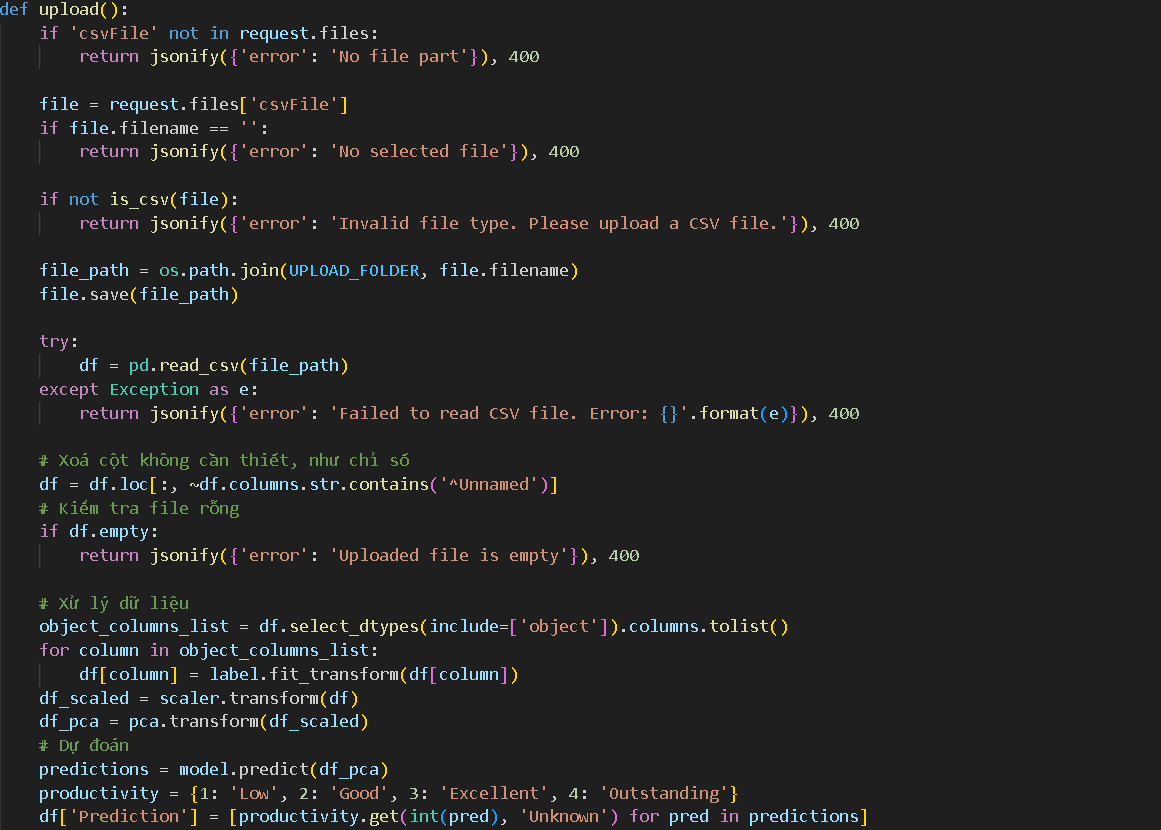
**-Xây phần backend(app.py)**

*Hàm xử lý và dự đoán của phần tự nhập các đặc trưng*



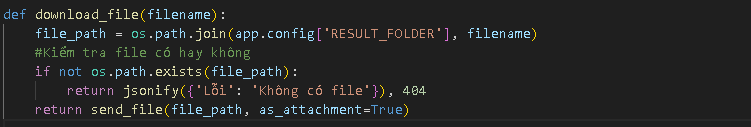
*Hình 5.1 Hình hàm xử lý và dự đoán của phần tự nhập các đặc trưng*

*Hàm xử lý dữ liệu của phần dự đoán bằng file csv*



*Hình 5.2 Hàm xử lý dữ liệu của phần dự đoán bằng file csv*

*Hàm tải file kết quả dự đoán*



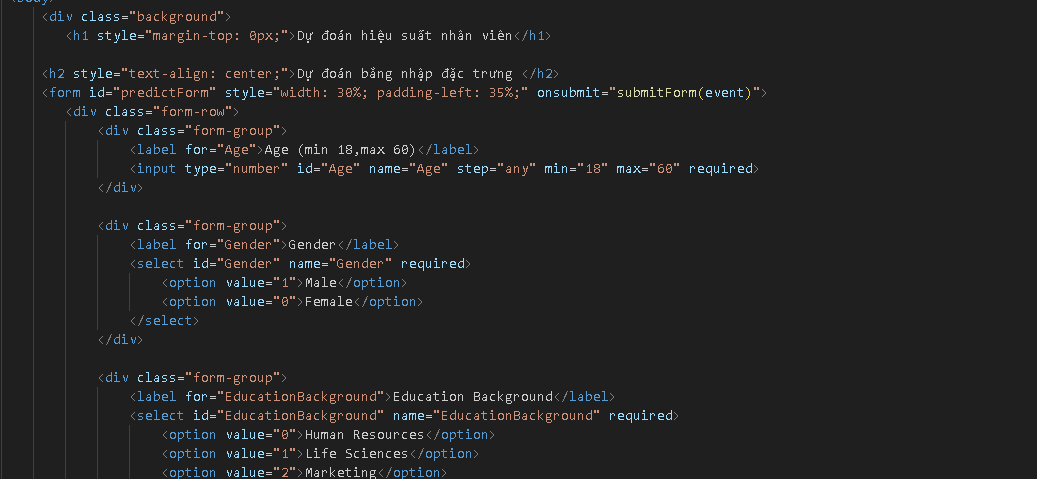
*Hình 5.3 Hàm tải file kết quả dự đoán*

*Hàm xác định file csv*



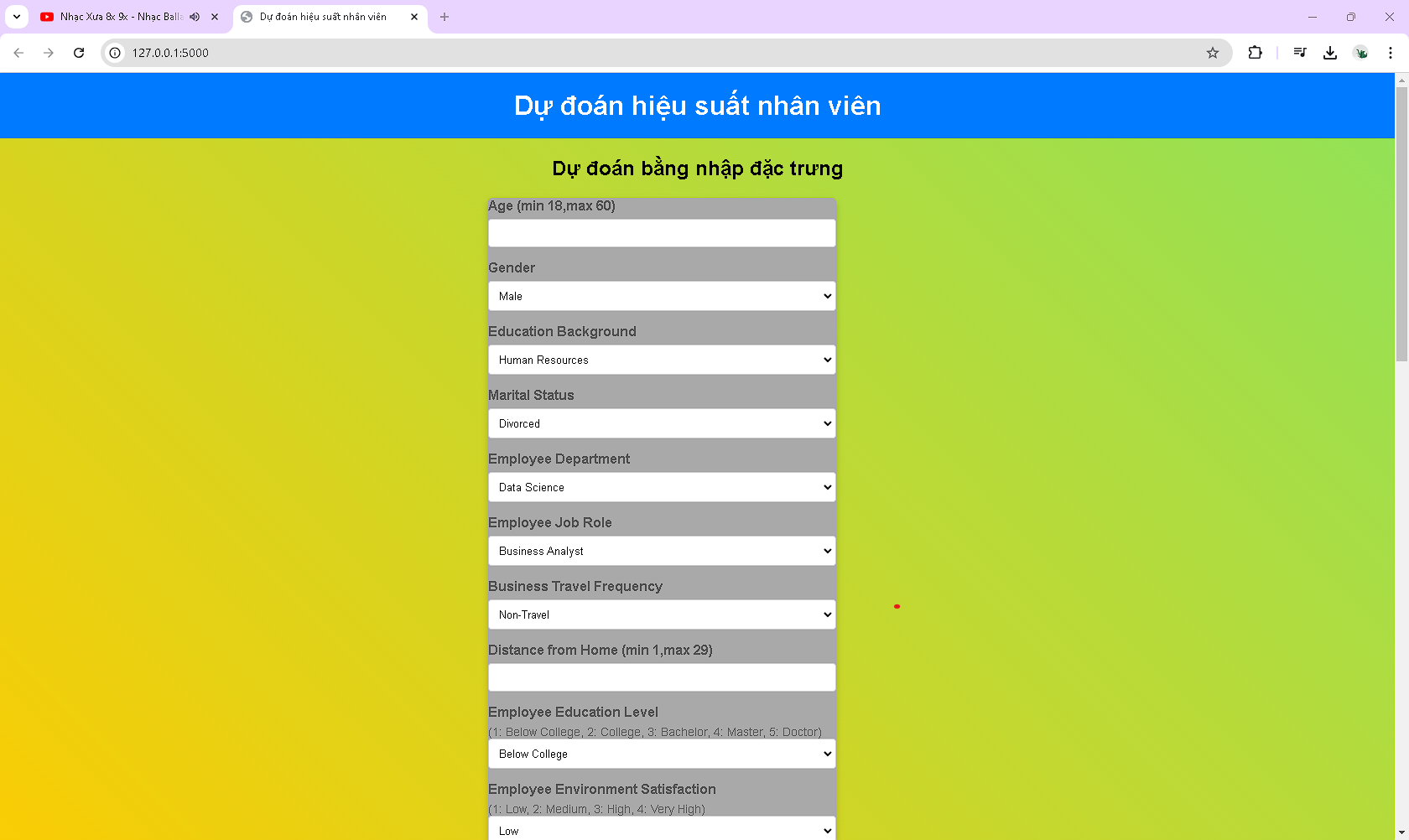
*Hình 5.4 Hàm xác định file csv*

-**Xây dựng giao diện(index.html)**



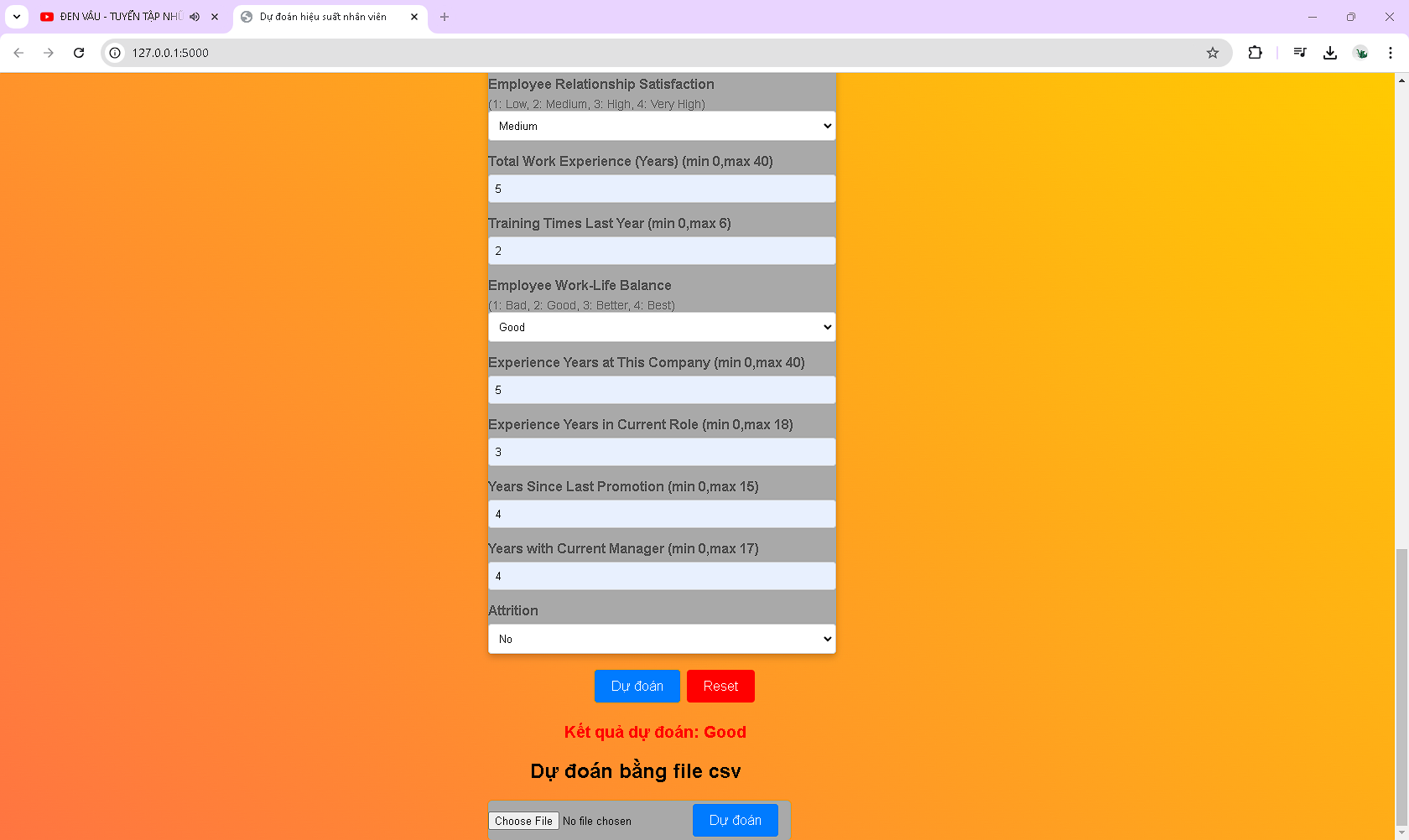
*Hình 5.5 Xây dựng giao diện*

**-Kết quả**

****

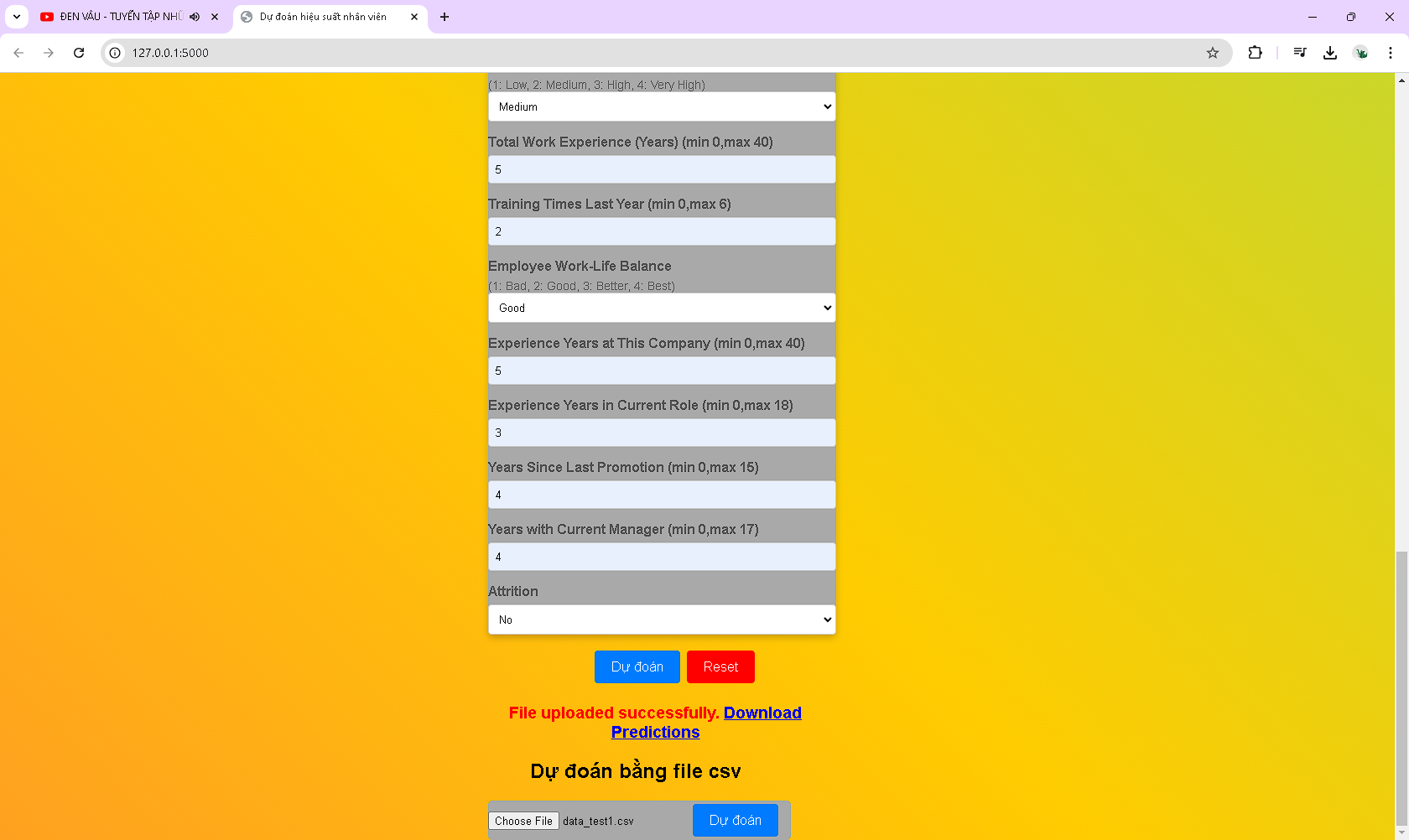
*Hình 5.6 Kết quả 1*

*Dự đoán bằng cách nhập số liệu*

****

*Hình 5.7* Dự đoán bằng cách nhập số liệu

*Dự đoán bằng tải file csv*

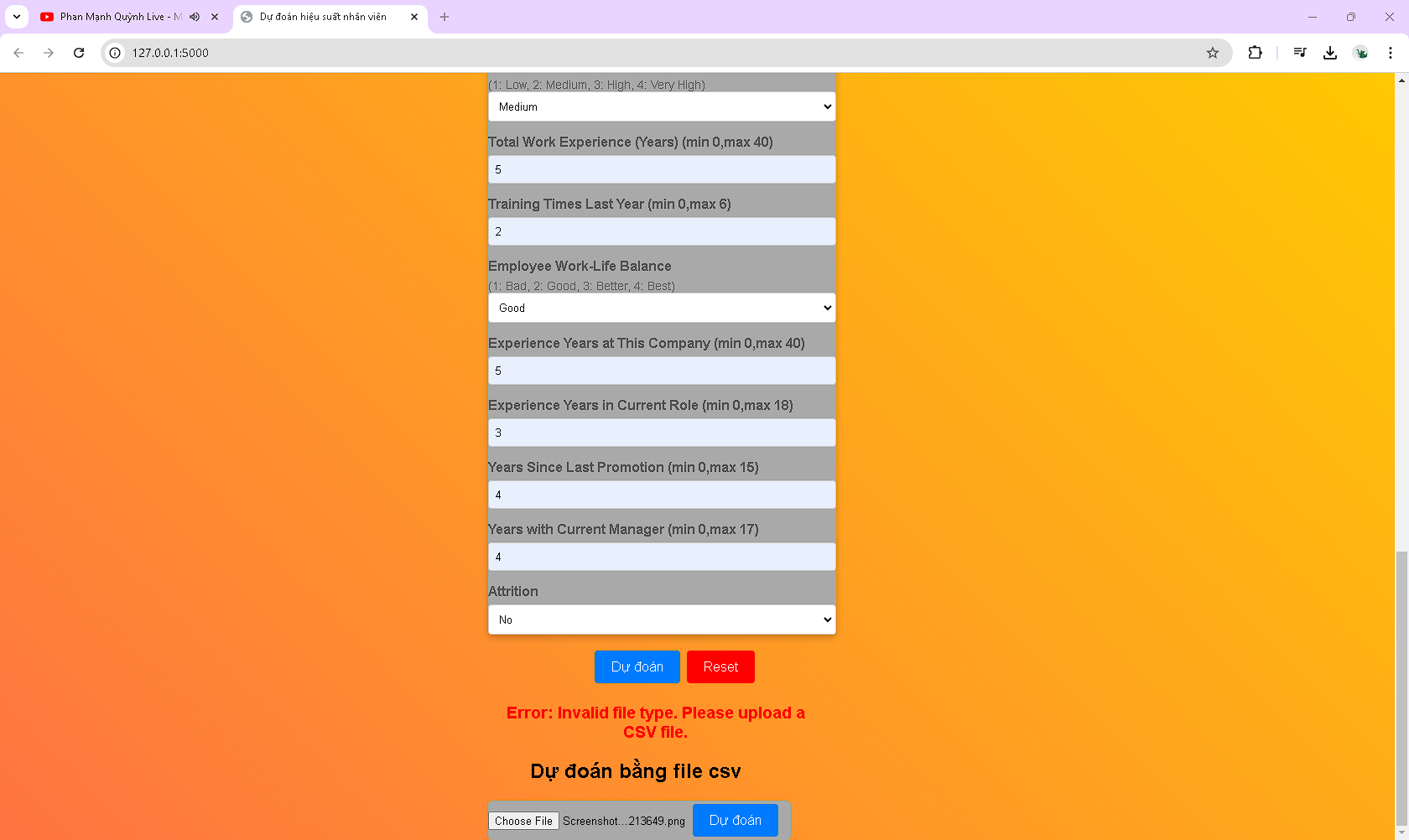


*Hình 5.8 Dự đoán bằng tải file csv*

Chúng ta tải một file csv cần dự đoán năng suất nhân viên lên và máy học sẽ dự đoán và trả về cho chúng ta một file csv kết quả,ta tải về để xem kết quả dự đoán

Data kết quả: [data\_submit](https://drive.google.com/drive/folders/1R5rQXSxxC7b7hG50jLqt6XMeP2ljJEd1?hl=vi)

*Khi tải một file không phải là file csv*



*Hình 5.9 Thông báo định dạng file csv*

Chương trình sẽ in ra giao diện dòng thông báo công phải kiểu file CSV,vui lòng tải file định dạng CSV

# KẾT LUẬN

## Kết quả đạt được.

Trên phân tích trên em đã phân tích dữ liệu và biết được các yếu tố ảnh hưởng đến hiệu suất của nhân viên và biết được năng suất của từng phòng ban để có kế hoạch quản lý và nâng cấp hiệu suất nhân viên của công ty

Qua bài bài phân tích trên em kết luận được

* ***Các yếu tố ảnh hưởng đến hiệu suất làm việc của nhân viên là:***

Emp Environment Satisfaction

Emp Last Salary Hike Percent

Emp Work Life Balance

* ***Đề xuất biện pháp cải thiện hiệu suất của nhân viên:***

1. Hiệu suất chung của nhân viên có thể đạt được thông qua sự hài lòng về môi trường của nhân viên. Công ty cần tập trung nhiều hơn vào sự hài lòng về môi trường của nhân viên.
2. Việc tăng lương sẽ thúc đẩy nhân viên làm việc tốt.
3. Thăng chức cho nhân viên sau mỗi 6 tháng
4. Cải thiện sự cân bằng giữa công việc và cuộc sống của nhân viên, điều này ảnh hưởng đến xếp hạng hiệu suất.
5. Khi tuyển dụng cho bộ phận nhân sự, hãy xem xét các ứng viên nữ có hiệu suất tốt hơn so với nam giới.
6. Bộ phận Development và Sales có hiệu suất chung cao hơn so với các bộ phận còn lại. Trong khi một số nhân viên đưa ra phản hồi như thấp & trung bình từ đặc trưng EmpJobSatisfaction & EmpRelationshipSatisfaction, những nhân viên như vậy có hiệu suất cao nhiều hơn về số lượng. Vì vậy, công ty nên tập trung vào họ

* **Thực hiện huấn luyện 3 mô hình đạt được hiệu suất khá tốt trên tập dữ liệu test**

Support Vector Machine 95.61%

Random Forest Classifier 93.14%

MLP Classifier 96.76%

Xây dựng được ứng dựng dự đoán hiệu suất bằng Flask có thể dự đoán bằng nhập liệu hoặc đọc trên file cvs để dự đoán kết quả hàng loạt

## Hạn chế và hướng phát triển

* **Hạn chế:**

Mặc dù kết quả accuracy trên tập test cho kết quả khá cao nhưng mô hình dự đoán cũng có nhiều trường hợp đưa ra kết quả bị sai lệch

Bộ dữ liệu đưa vào huấn luyện còn ít

Hệ thống dự đoán xây dựng còn chưa được hoàn chỉnh và đẹp mắt

* **Hướng phát triển:**

Thu thập dữ liệu nhiều hơn để huấn luyện mô hình được tốt hơn

Tích hợp mô hình vào ứng dụng mobile

Xây dựng thêm nhiều tiện ích và hỗ trợ thêm các đuôi file khác nhau

# TÀI LIỆU THAM KHẢO

| 1 | [Thống kê mô tả (truy cập ngày 20/07/2024)](https://nghiencuugiaoduc.com.vn/bai-7-thong-ke-mo-ta-descriptive-statistics/) |
| --- | --- |
| 2 | [Thống kê mô tả (truy cập ngày 20/07/2024)](https://nghiencuugiaoduc.com.vn/bai-7-thong-ke-mo-ta-descriptive-statistics/) |
| 3 | [Sử dụng thống kê để xác định và loại bỏ ngoại lai (truy cập ngày 20/07/2024)](https://svcuong.github.io/post/remove-outliers/) |
| 4 | [Giới thiệu về mô hình SVM(truy cập ngày 20/07/2024)](https://hllj.github.io/knowledge/machine%20learning/SVM-Introduction-Part-1/) |
| 5 | [Random Forest Classifier using Scikit-learn(truy cập ngày 20/07/2024)](https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/) |
| 6 | [Neural network models(truy cập ngày 20/07/2024)](https://scikit-learn.org/stable/modules/neural_networks_supervised.html#classification) |
| 7 | [Classification:Accuracy(truy cập ngày 20/07/2024)](https://developers.google.com/machine-learning/crash-course/classification/accuracy) |
| 8 | [scikit-learn](https://scikit-learn.org/stable/index.html) |
| 9 | [Flask](https://flask.palletsprojects.com/en/3.0.x/) |
| 10 | Dataset [Employee\_Performance](https://drive.google.com/drive/folders/1R5rQXSxxC7b7hG50jLqt6XMeP2ljJEd1?usp=drive_link) |
| 11 | Chat GPT |