

PHÂN TÍCH VÀ DỰ ĐOÁN

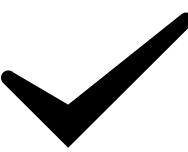
TỈ LỆ NGHỈ VIỆC CỦA NHÂN VIÊN

Thành viên:

Tô Huỳnh Ngọc Ngân
Trần ngọc Minh Khanh
Nguyễn Thị Hồng Phúc



MỤC LỤC



I. Bối cảnh và tổng quan về dữ liệu

II. Phân tích EDA + Biểu đồ

III. Tiền xử lý dữ liệu

IV. Mô hình hóa và đánh giá mô hình

V. Giải pháp và kết luận, đề xuất giải pháp

BỐI CẢNH



“Tỷ lệ nghỉ việc cao ảnh hưởng đến chi phí tuyển dụng, hiệu suất đội nhóm và tinh thần nhân viên. Do đó, việc dự đoán và can thiệp sớm là cần thiết.”



TÌNH TRẠNG NHÂN VIÊN NGHỈ VIỆC CÓ THỂ GÂY RA HỆ LỤY:



- Tình trạng nghỉ việc gây tăng chi phí tuyển dụng – đào tạo
- Gián đoạn công việc, giảm hiệu suất đội nhóm
- Mất kiến thức nội bộ, ảnh hưởng tinh thần nhân viên còn lại
- Lý do nghỉ việc khó quan sát, đánh giá thủ công thiếu chính xác, khó mở rộng

MỤC TIÊU



XÁC ĐỊNH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN NGHỈ VIỆC

DỰ ĐOÁN KHẢ NĂNG NGHỈ VIỆC CỦA NHÂN VIÊN

ĐỀ XUẤT CHIẾN LƯỢC GIỮ CHÂN HIỆU QUẢ DỰA
TRÊN DỮ LIỆU

TỔNG QUAN VỀ DỮ LIỆU

- Nguồn: Kaggle.com
- Dataset: Employee Dataset
- Số lượng bảng: 1
- Số lượng cột: 9
- Số lượng hàng: 4653
- No null values

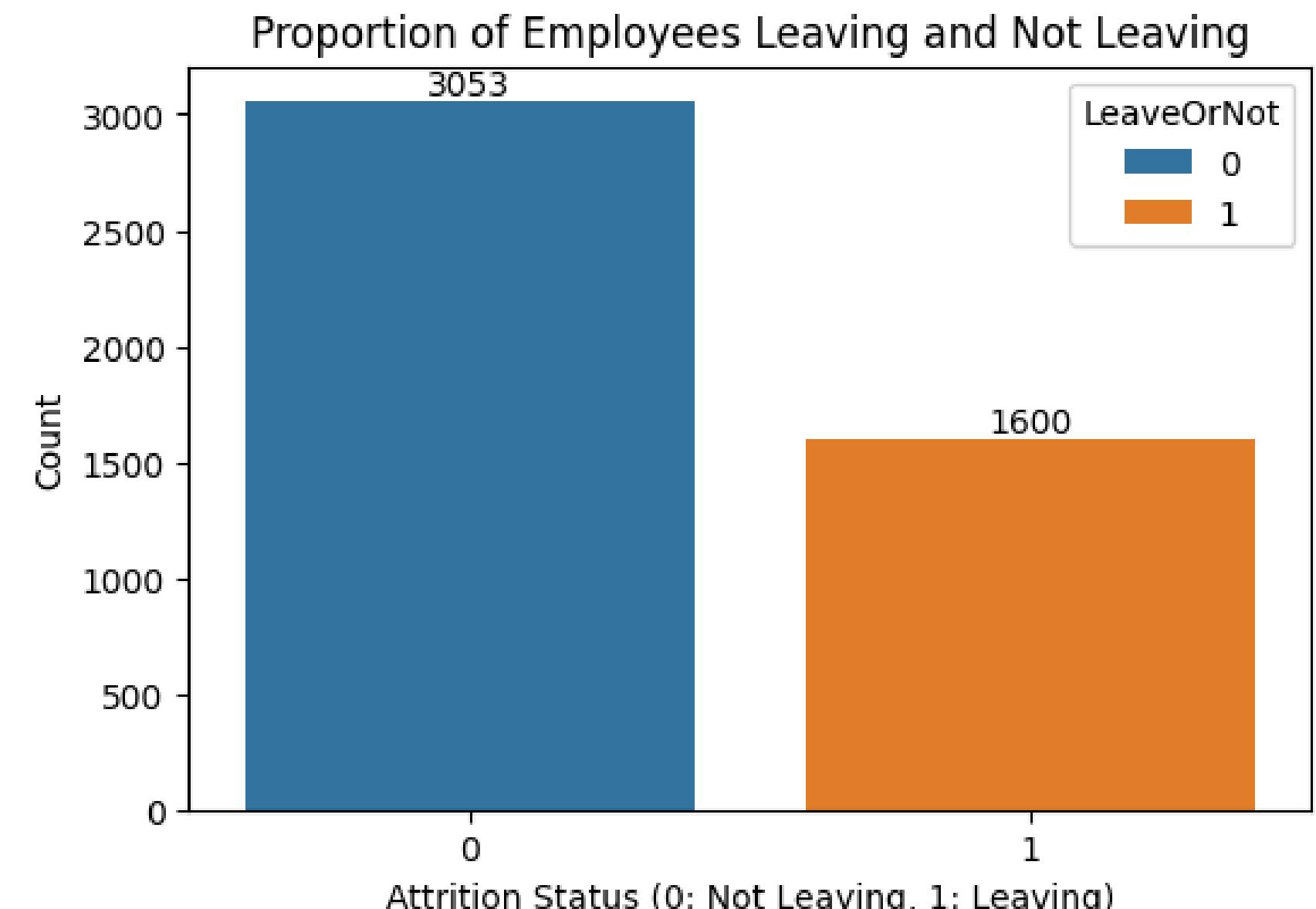
```
RangeIndex: 4653 entries, 0 to 4652
Data columns (total 9 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   Education        4653 non-null   object  
 1   JoiningYear     4653 non-null   int64  
 2   City             4653 non-null   object  
 3   PaymentTier      4653 non-null   int64  
 4   Age              4653 non-null   int64  
 5   Gender            4653 non-null   object  
 6   EverBenched      4653 non-null   object  
 7   ExperienceInCurrentDomain 4653 non-null   int64  
 8   LeaveOrNot       4653 non-null   int64  
dtypes: int64(5), object(4)
```

Cột dữ liệu	Mô tả
Education	Trình độ học vấn, bao gồm bằng cấp, ngành học và cơ sở đào tạo.
Joining Year	Năm gia nhập công ty – dùng để tính thời gian làm việc.
City	Thành phố nơi nhân viên đang sống hoặc làm việc.
Payment Tier	Bậc thanh toán – phân loại mức lương của nhân viên theo cấp bậc.
Age	Tuổi của nhân viên – cung cấp thông tin nhân khẩu học.
Gender	Giới tính – phục vụ phân tích về tính đa dạng trong tổ chức.
Ever Benched	Nhân viên từng bị "benched" (không được phân công công việc) hay chưa.
Experience in Current Domain	Số năm kinh nghiệm trong lĩnh vực hiện tại của nhân viên.
Leave or Not	Cho biết nhân viên có rời công ty hay không.

LABEL BIẾN MỤC TIÊU

Dữ liệu có:

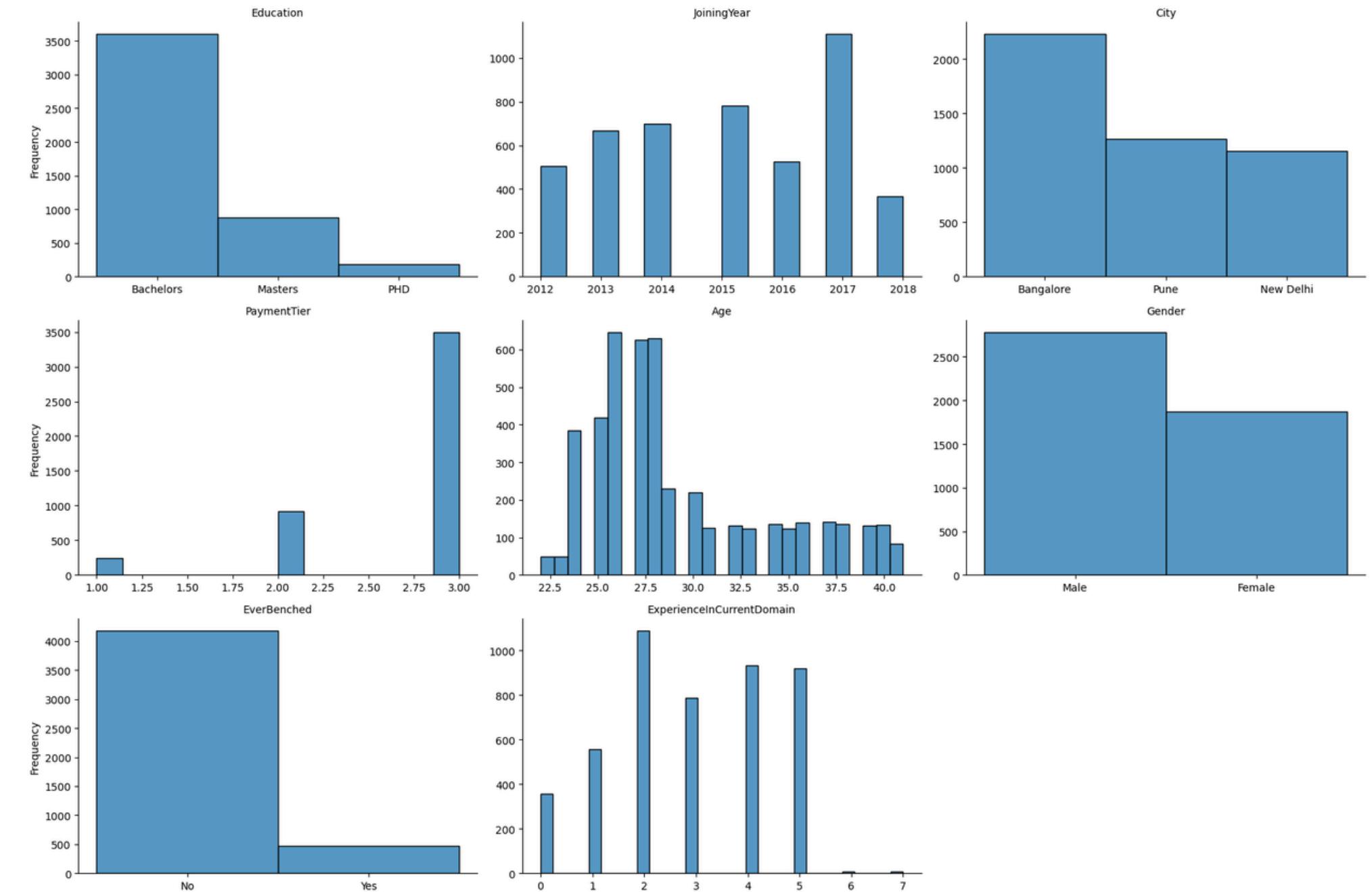
- 3053 cột chứa giá trị 0, nghĩa là những nhân viên ở lại, chiếm khoảng 65.6%.
- 1600 cột chứa giá trị 1, nghĩa là những nhân viên đã nghỉ việc, chiếm khoảng 34.4%.



PHÂN PHỐI DỮ LIỆU

- Dữ liệu có sự lệch đáng kể ở nhiều biến: học vấn, lương, tuổi, giới tính.

=> Insight quan trọng: Nhiều nhóm nhân viên trẻ, ít kinh nghiệm, lương thấp → dễ biến động nhân sự → cần chính sách giữ chân cụ thể.



MỤC LỤC



I. Bối cảnh và tổng quan về dữ liệu



II. Phân tích EDA + Biểu đồ

III. Tiền xử lý dữ liệu

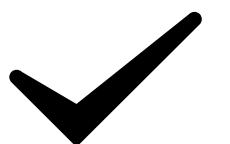
IV. Mô hình hóa và đánh giá mô hình

V. Giải pháp và kết luận, đề xuất giải pháp

PHÂN TÍCH DỮ LIỆU

BIẾN SỐ CÁ NHÂN

(Tuổi, Năm kinh nghiệm, Giới tính)



BIẾN SỐ TỔ CHỨC

(Thành phố, Bậc lương,
Everbenched)

HỌC VẤN VÀ NĂM GIA NHẬP

(Học vấn, Năm gia nhập cty)



PHÂN TÍCH THEO ĐỘ TUỔI

Nhận xét:

- 18 - 25: Tỷ lệ nghỉ việc tương đối cao => cải thiện trải nghiệm
- 26 - 35: Đóng nhất, tỷ lệ nghỉ việc cao nhất => cần giữ chân
- >35 tuổi: Tỷ lệ nghỉ việc thấp => Ôn định, kinh nghiệm

Insight:

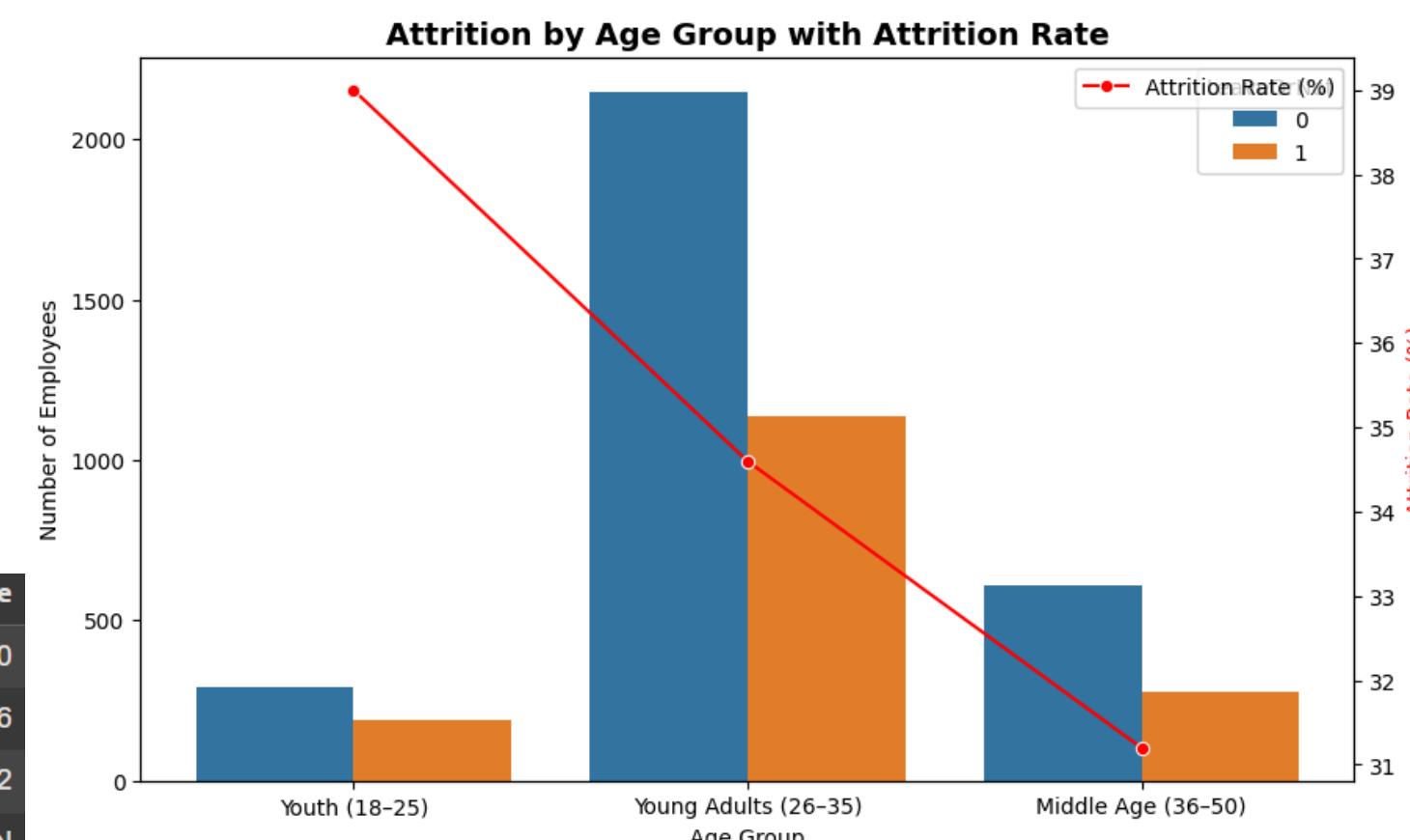
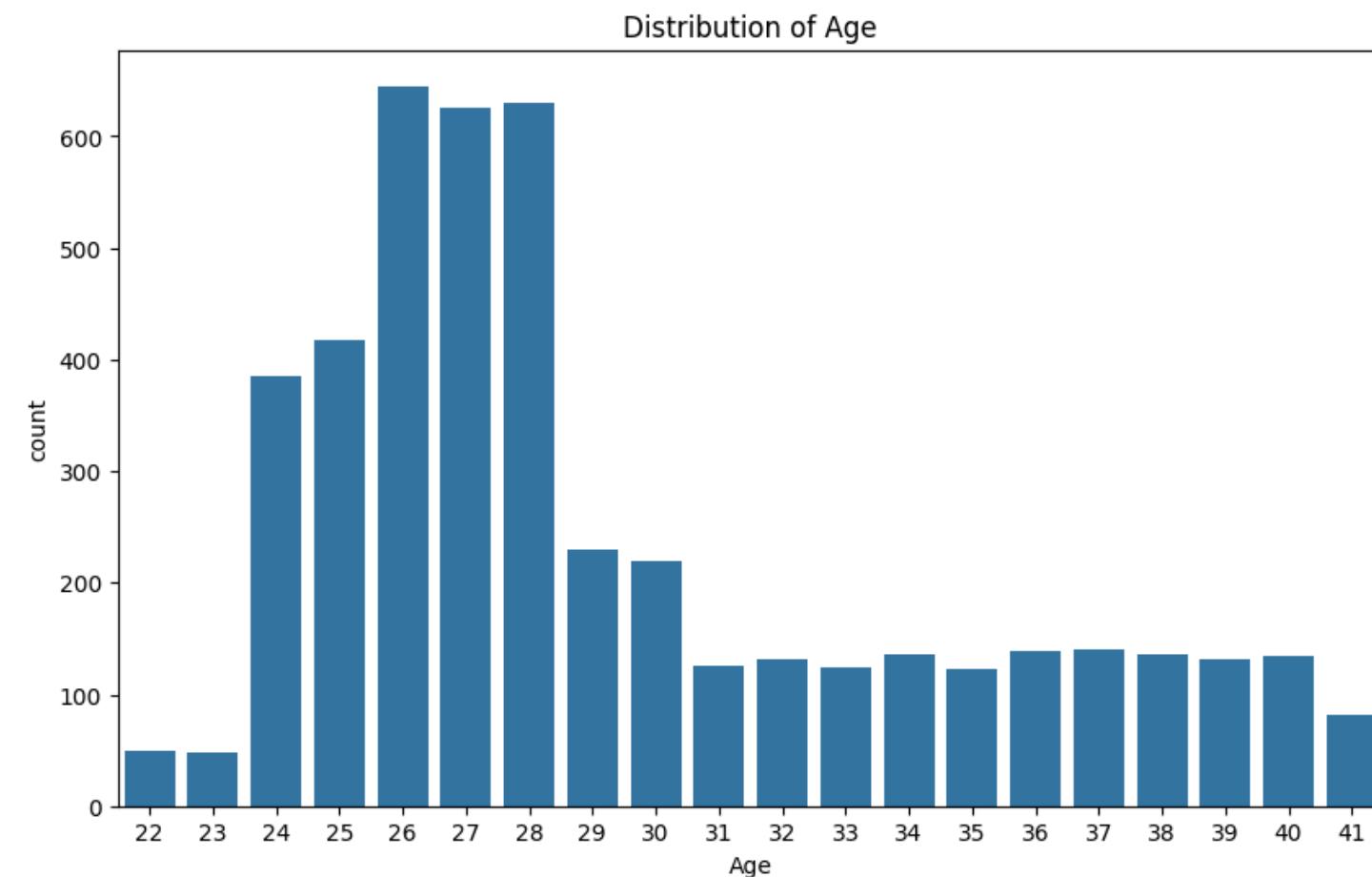
- AgeGroup liên quan chặt chẽ đến khả năng nghỉ việc.
- Nghỉ việc tập trung ở nhóm tuổi trẻ, đang phát triển sự nghiệp.

Kết luận:

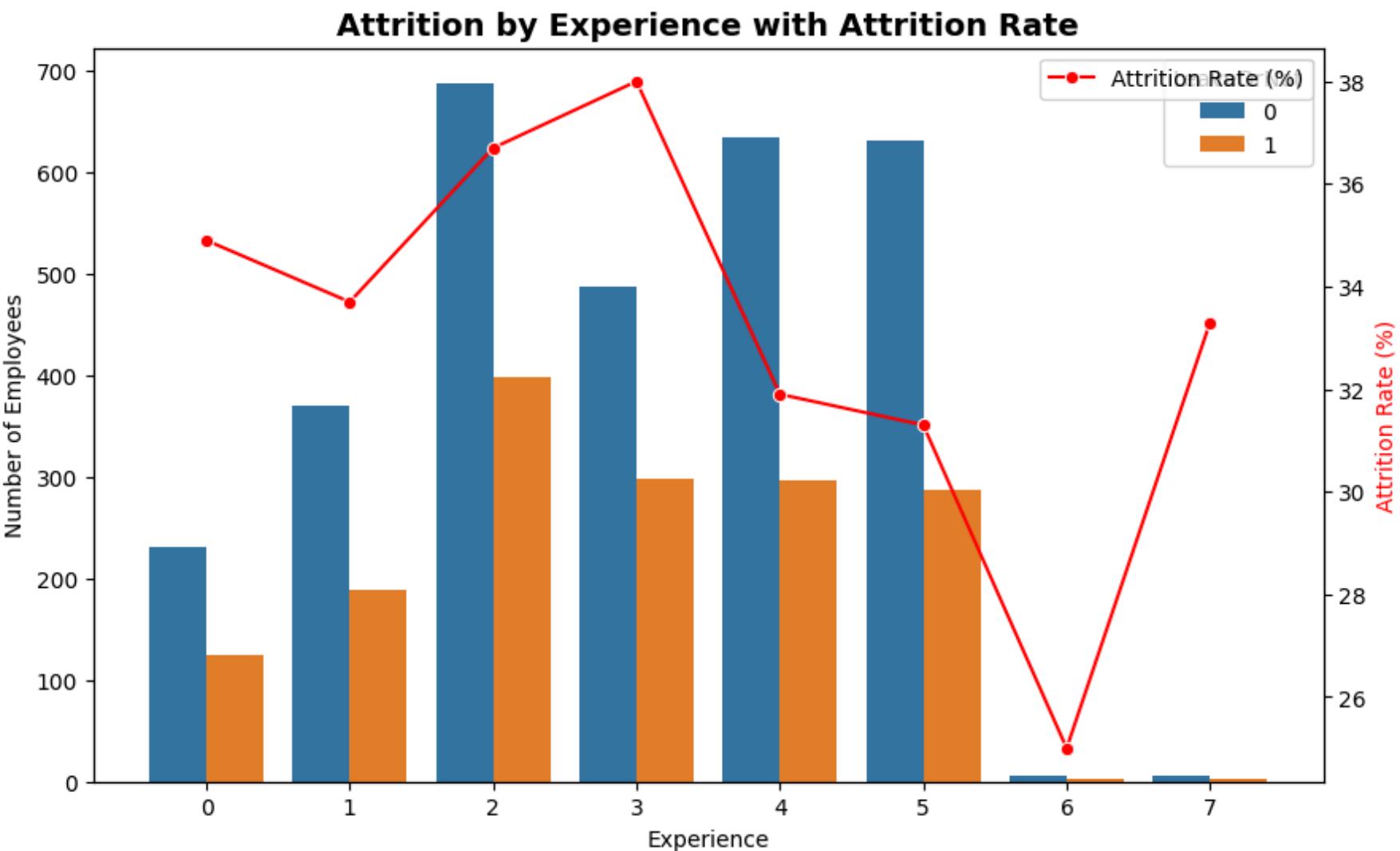
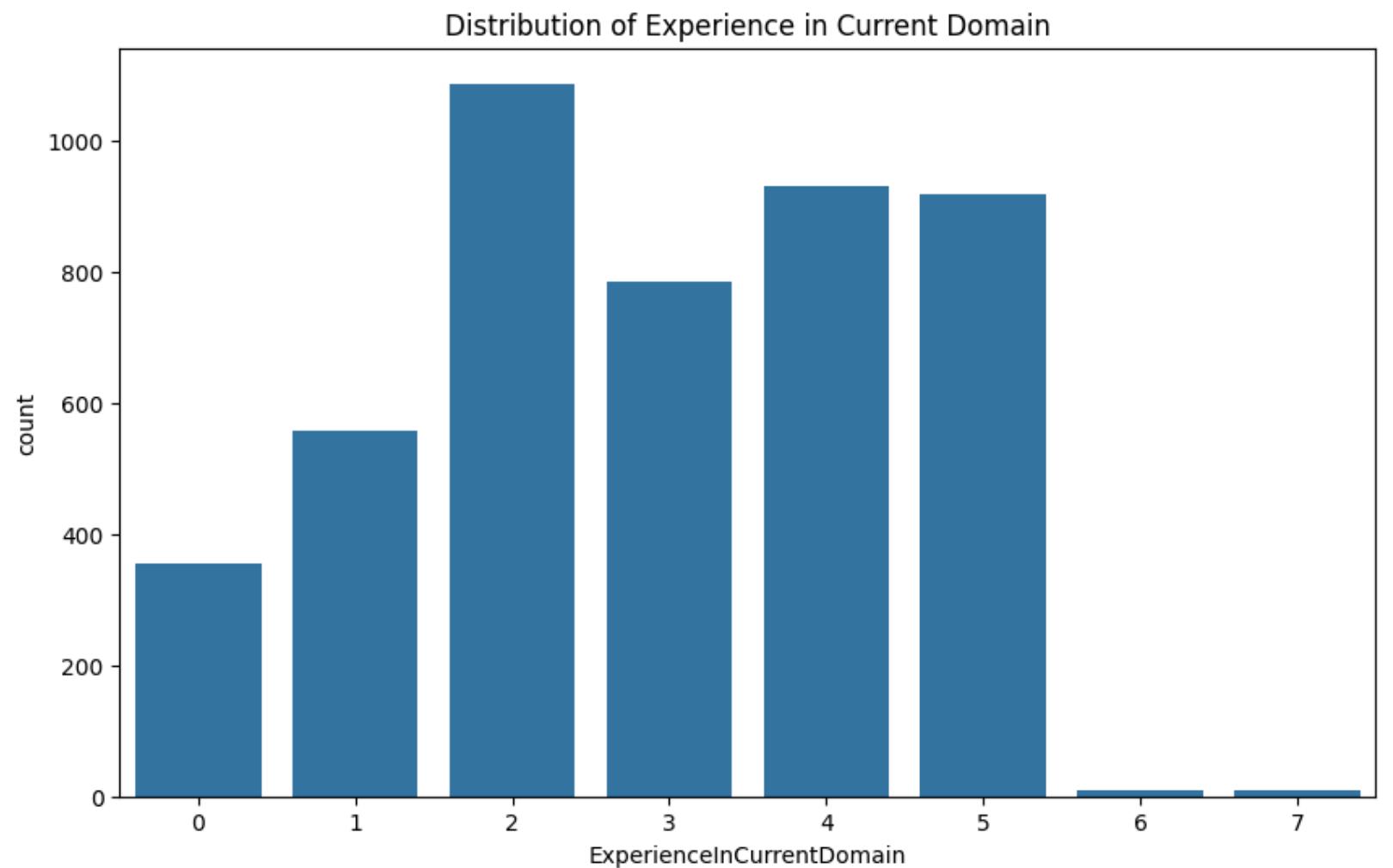
→ AgeGroup là biến quan trọng cần đưa vào mô hình dự đoán

Age Range	Group Name	Meaning
18–25	Youth	Sinh viên mới ra trường, bắt đầu sự nghiệp
26–35	Young Adults	Trưởng thành, phát triển nghề nghiệp
36–50	Middle Age	Ôn định, nhiều kinh nghiệm, tiềm năng lãnh đạo
51–65	Older Adults	Dày dặn kinh nghiệm, có thể chuẩn bị nghỉ hưu

AgeGroup	AttritionRate
0	Youth (18–25)
1	Young Adults (26–35)
2	Middle Age (36–50)
3	Older Adults (51–65)



PHÂN TÍCH THEO NĂM KINH NGHIỆM



Nhận xét:

- Nhân viên tập trung ở nhóm 2-5 năm kinh nghiệm.
- Tỷ lệ nghỉ việc cao nhất ở nhóm 3 năm (~38%).

Insight:

- **Kinh nghiệm càng thấp → Nguy cơ nghỉ việc càng cao.**
- **Nghỉ việc tập trung ở nhóm < 3 năm kinh nghiệm.**

→ ExperiencelnCurrentDomain là biến nên đưa vào mô hình vì liên quan rõ rệt đến khả năng nghỉ việc.

ExperienceInCurrentDomain	AttritionRate
0	34.9
1	33.7
2	36.7
3	38.0
4	31.9
5	31.3
6	25.0
7	33.3

PHÂN TÍCH THEO GIỚI TÍNH

Nhận xét:

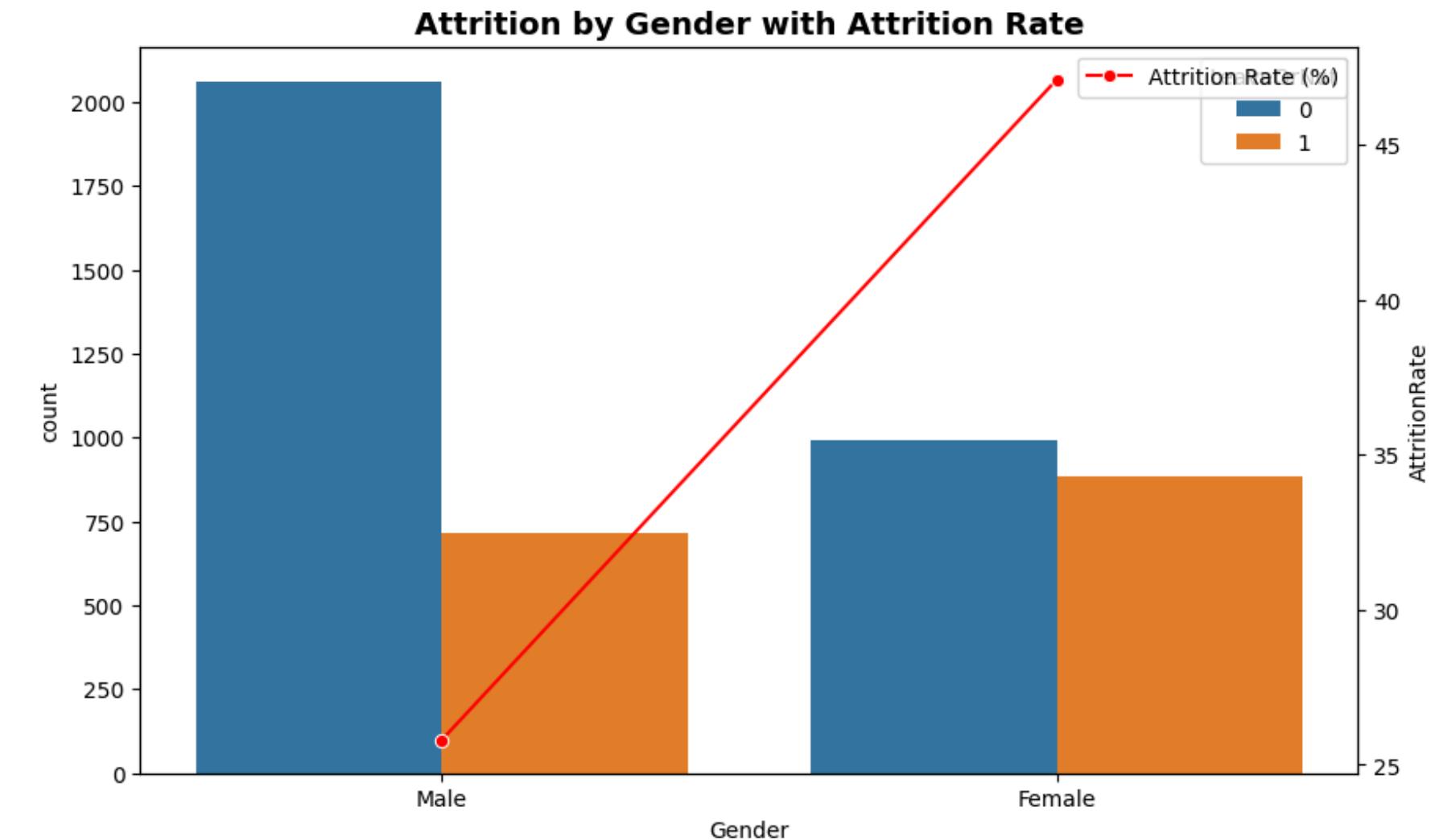
Nữ giới có tỷ lệ nghỉ việc cao hơn nam (47.1% vs 25.8%).

📌 **Insight:**

Nhân viên nữ là nhóm có nguy cơ nghỉ việc cao, cần chú trọng trong chính sách giữ chân.

✓ **Kết luận:**

Gender là biến quan trọng, nên đưa vào mô hình.



Gender	AttritionRate
0 Female	47.1
1 Male	25.8

KẾT LUẬN CHUNG



Kết luận từ 3 yếu tố chính:

Độ tuổi (Age Group):

- Nhóm trẻ (18–35) nghỉ việc cao, nhất là 26–35 → Cần môi trường tốt, cơ hội thăng tiến.

Kinh nghiệm (Experience):

- <3 năm kinh nghiệm nghỉ việc cao nhất (~38%) → Giai đoạn nhạy cảm, cần giữ chân.
- Sau 4 năm: nghỉ việc giảm → Tăng tính ổn định.

Giới tính (Gender):

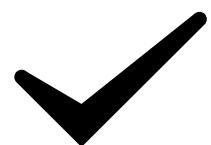
- Nữ giới có tỷ lệ nghỉ việc cao gấp đôi nam → Cần chính sách hỗ trợ & phát triển công bằng.

Tóm lại:

→ **Ba yếu tố này cần đưa vào mô hình dự đoán và là trọng tâm của chiến lược giữ chân nhân sự.**

PHÂN TÍCH DỮ LIỆU

BIẾN SỐ CÁ NHÂN
(Tuổi, năm kinh nghiệm,
giới tính)

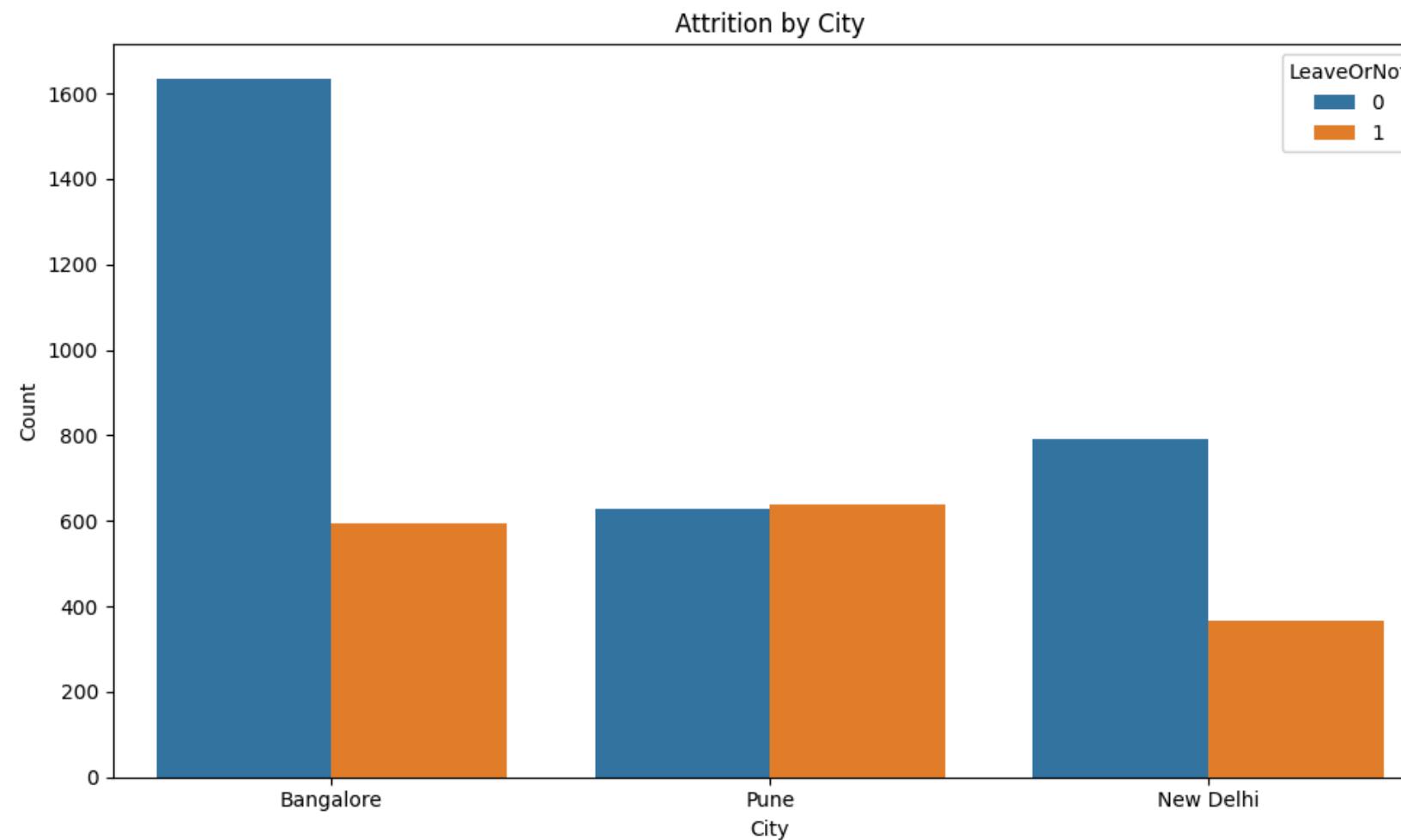


BIẾN SỐ TỔ CHỨC
(Thành phố, bậc lương,
Everbenched)

HỌC VẤN VÀ NĂM GIA NHẬP
(Học vấn, năm gia nhập cty)



PHÂN TÍCH THEO THÀNH PHỐ



City	AttritionRate
Pune	0.503943
New Delhi	0.316335
Bangalore	0.267056

Nhận xét:

- Pune: Tỷ lệ nghỉ việc cao nhất (~50%) → Đáng báo động.
- New Delhi: Tỷ lệ nghỉ việc trung bình (~31,6%) → Cần chú ý.
- Bangalore: Số lượng nhân viên lớn nhất, tỷ lệ nghỉ việc thấp nhất.

✓ Kết luận:

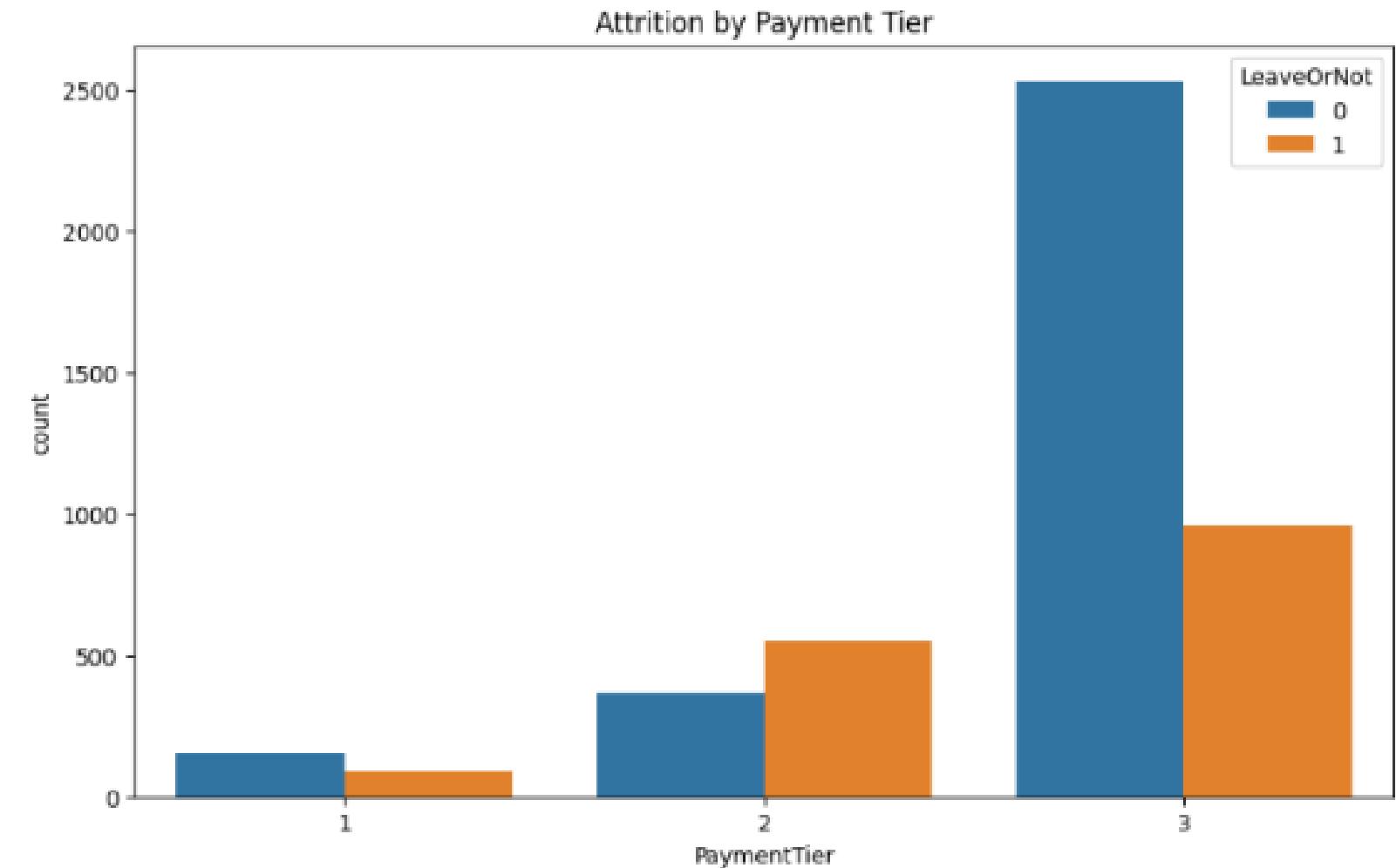
- **Cần điều tra nguyên nhân & xây dựng chính sách giữ chân theo từng thành phố.**

BẬC LƯƠNG (PAYMENT TIER)

Nhận xét:

- Tier 3: Số lượng lớn, tỷ lệ nghỉ việc thấp → Giữ chân tốt.
- Tier 2: Tỷ lệ nghỉ việc cao hơn → Nhóm rủi ro cần chú ý.
- Tier 1: Số lượng ít, tỷ lệ nghỉ việc cao → Lương thấp khó giữ chân.

→ **Mức lương liên quan chặt chẽ đến quyết định nghỉ việc → Nên đưa vào mô hình & xem xét chính sách lương hợp lý.**



PaymentTier	AttritionRate
0	1
1	2
2	3

EVERBENCHED ẢNH HƯỞNG TỚI LEAVEORNOT?

Benched & Nghỉ việc:

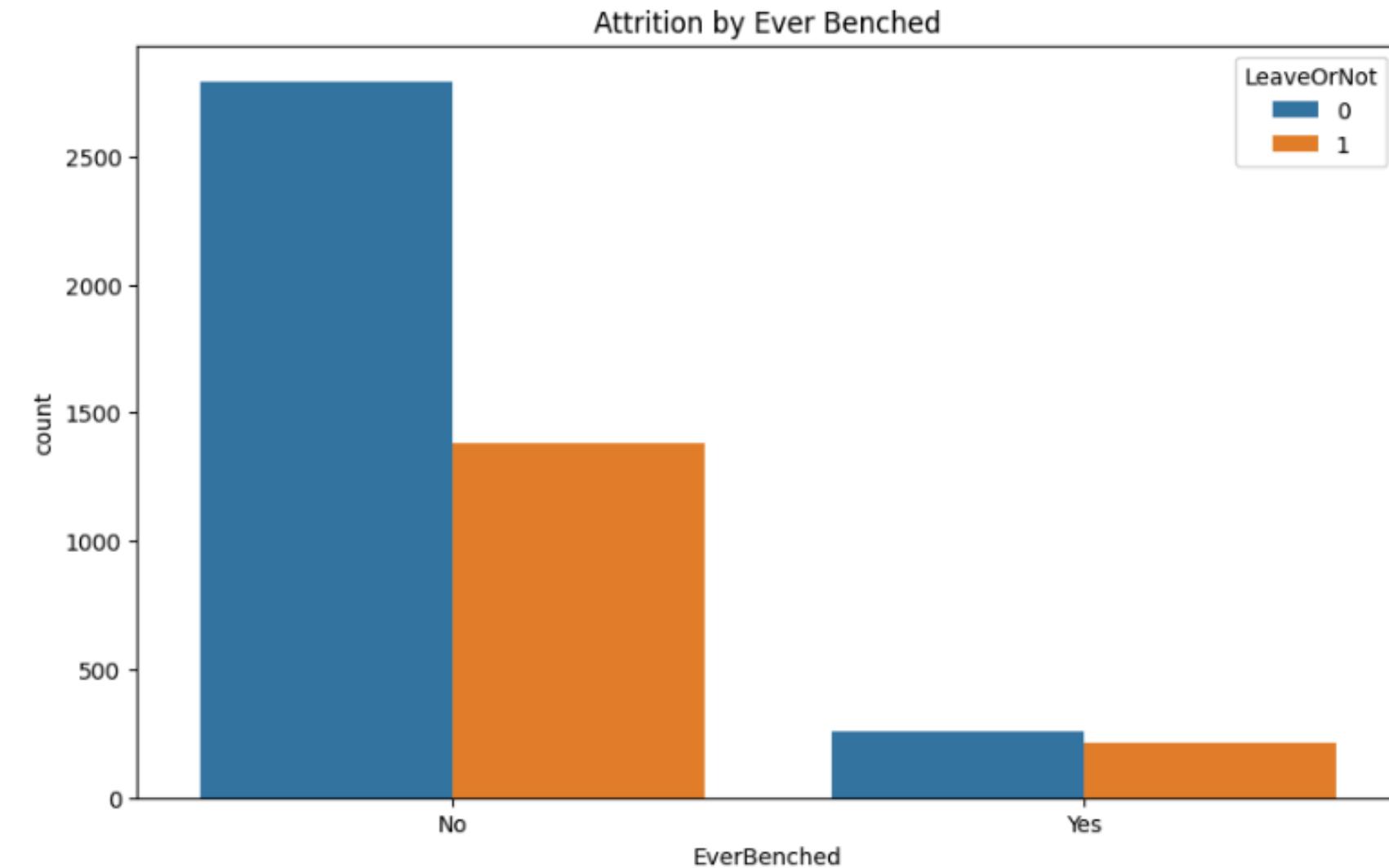
- Nhóm Benched (Yes): Tỷ lệ nghỉ việc ~45% (cao hơn đáng kể)
- Nhóm Không Benched (No): Tỷ lệ nghỉ việc ~33%

📌 **Insight:**

Bị “benched” liên quan mạnh đến nguy cơ nghỉ việc

✓ Kết luận:

→ Nên đưa vào mô hình và xem xét chính sách hỗ trợ.



EverBenched	AttritionRate
Yes	0.453975
No	0.331257

KẾT LUẬN CHUNG



Vị trí làm việc (City), Bậc lương (Payment Tier), Trạng thái Benched đều có mối liên hệ chặt chẽ với quyết định nghỉ việc của nhân viên.

Các nhóm có tỷ lệ nghỉ việc cao:

→ Pune, Payment Tier 2 & 1, Benched (Yes) → Nhóm rủi ro cao, cần ưu tiên giữ chân.

Các nhóm có tỷ lệ nghỉ việc thấp:

→ Bangalore, Payment Tier 3, Không Benched (No) → Nhóm ổn định hơn.

Khuyến nghị:

Nên đưa các yếu tố này vào mô hình dự đoán.

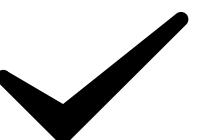
Điều chỉnh chính sách phù hợp từng nhóm để cải thiện tỷ lệ nghỉ việc.

PHÂN TÍCH DỮ LIỆU

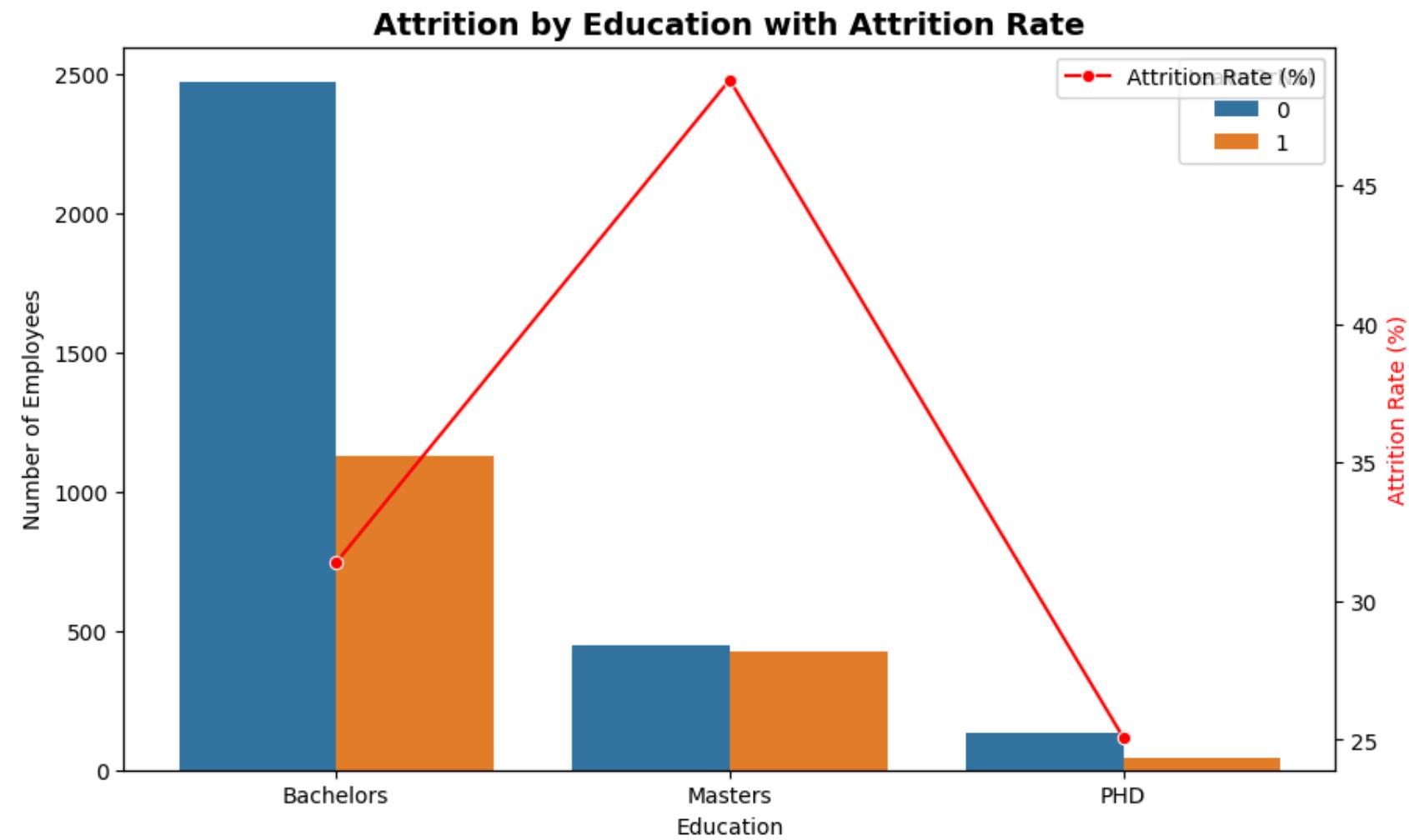
BIỂN SỐ CÁ NHÂN
(Tuổi, năm kinh
nghiệm, giới tính)

BIỂN SỐ TỔ CHỨC
(Thành phố, bậc
lương, Everbenched)

HỌC VĂN VÀ NĂM GIA NHẬP
(Học văn, năm gia nhập cty)



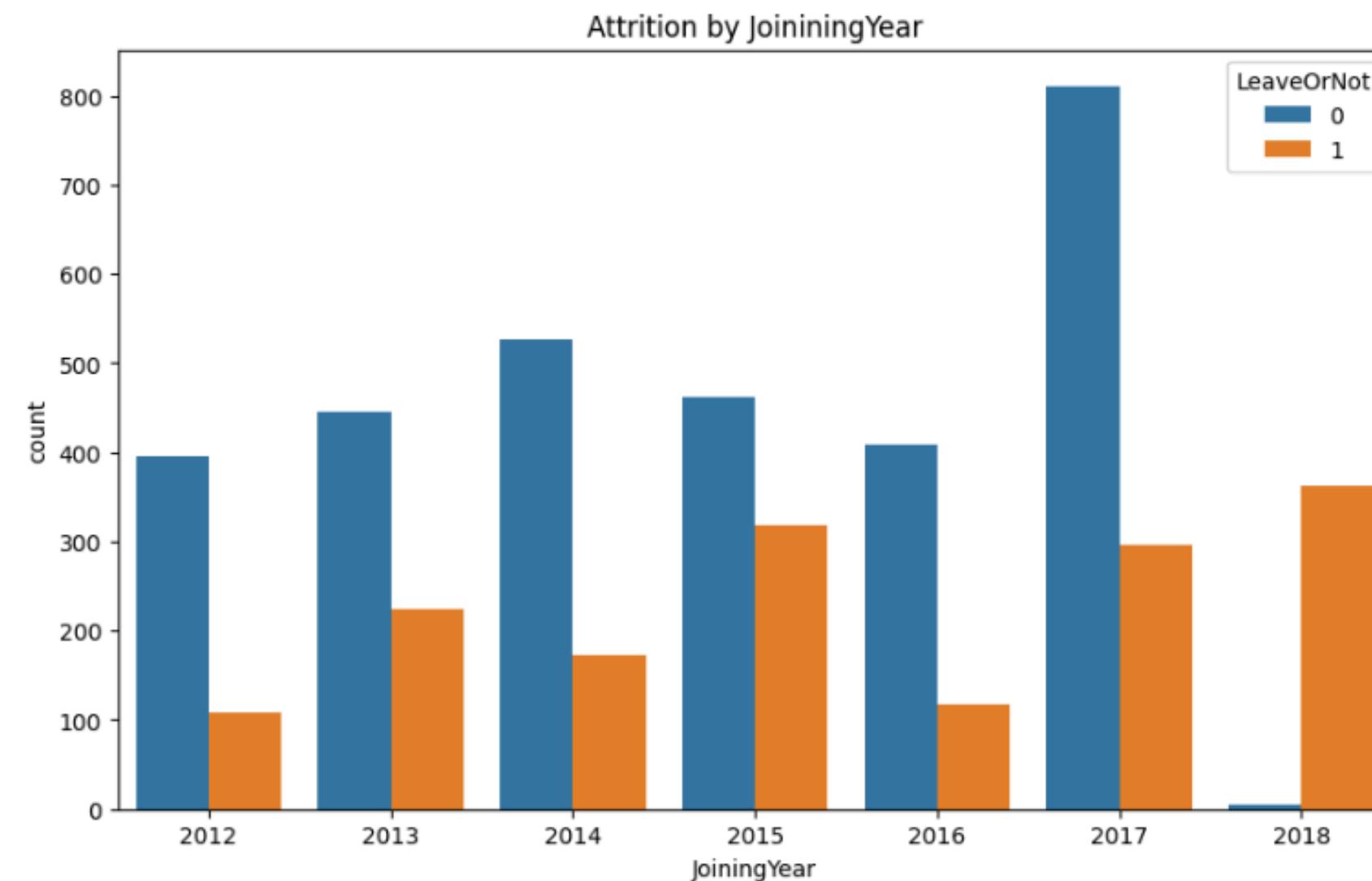
TRÌNH ĐỘ HỌC VẤN ĐƯỢC PHÂN BỐ NHƯ THẾ NÀO ?



- Masters: Tỷ lệ nghỉ việc cao nhất → Nhóm nguy cơ cao.
 - PHD: Tỷ lệ nghỉ việc thấp → Nhóm ổn định, cần đầu tư phát triển.
 - Bachelors: Số lượng đông, nghỉ việc nhiều → Cần chiến lược giữ chân hiệu quả.
- **Kết luận:** Trình độ học vấn ảnh hưởng đến quyết định nghỉ việc → Nên đưa vào mô hình & xây dựng chính sách phù

Education	AttritionRate
0 Bachelors	31.4
1 Masters	48.8
2 PHD	25.1

JOININGYEAR CÓ ẢNH HƯỞNG THẾ NÀO ĐẾN LEAVEORNOT?



JoiningYear	AttritionRate
2018	0.986376
2015	0.407170
2013	0.334828
2017	0.268051
2014	0.247496
2016	0.222857
2012	0.216270

Năm gia nhập & Nghỉ việc:

- 2013–2015: Tỷ lệ nghỉ việc cao, đặc biệt 2015 ~40.7%.
- 2016–2017: Tỷ lệ nghỉ việc thấp, ổn định (~22–26%).
- 2018: Tỷ lệ nghỉ việc rất cao (~98.6%) → Có thể do Layoff hoặc biến động lớn.

→ **Kết luận:** Thời gian gia nhập ảnh hưởng rõ rệt đến nguy cơ nghỉ việc → Nên đưa vào mô hình & theo dõi biến động theo năm

KẾT LUẬN CHUNG



📌 Nhận định & Insight:

- Masters: Tỷ lệ nghỉ việc cao → Cần khảo sát kỳ vọng & chính sách đãi ngộ.
- PhD: Ổn định → Nên đầu tư & giữ chân lâu dài.
- Bachelors: Đóng nhất, nghỉ việc cao → Nhóm cần chiến lược giữ chân trọng tâm.
- 2013–2015: Tỷ lệ nghỉ việc cao, đặc biệt 2015 (~40.7%).
- 2016–2017: Ổn định, nghỉ việc thấp.
- 2018: Nghỉ việc gần như toàn bộ → Có thể do layoff hoặc bất thường dữ liệu.

✓ Kết luận:

- **Education & Joining Year là biến quan trọng cần đưa vào mô hình.**
- **Nên tập trung giữ chân nhóm Masters & Bachelors, đồng thời điều chỉnh chính sách theo năm gia nhập.**

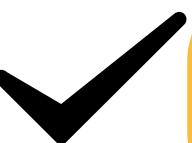


MỤC LỤC



I. Bối cảnh và tổng quan về dữ liệu

II. Phân tích EDA + Biểu đồ



III.Tiền xử lí dữ liệu

IV. Mô hình hóa và đánh giá mô hình

V. Ý nghĩa và kết luận, đề xuất giải pháp

TIỀN XỬ LÍ DỮ LIỆU

Nhóm độ tuổi thành AgeGroup

18–25: Youth, 26–35: Young Adults. 36–50 :Middle Age, 51–65: Older Adults

Mã hóa biến phân loại: One-hot Encoding

```
# One-hot encode all categorical columns  
data = pd.get_dummies(data, columns=['Education', 'City', 'AgeGroup', 'JoiningYear', 'PaymentTier'], drop_first=True)
```

Map giá trị Yes/No, Male/Female thành số

```
data['Gender'] = data['Gender'].map({'Male': 1, 'Female': 0})  
data['EverBenched'] = data['EverBenched'].map({'Yes': 1, 'No': 0})
```

Áp dụng SMOTE để xử lý mất cân bằng nhãn.

Áp dụng GridSearchCV để tìm tham số hợp lí





MỤC LỤC



I. Bối cảnh và tổng quan về dữ liệu

II. Phân tích EDA + Biểu đồ

III.Tiền xử lí dữ liệu



IV. Mô hình hóa và đánh giá mô hình

V. Ý nghĩa và kết luận, đề xuất giải pháp

XÂY DỰNG MÔ HÌNH

```
[40] x = data.drop(columns=['LeaveOrNot'])
y = data['LeaveOrNot']

[42] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42, stratify=y)

[43] scaler = StandardScaler()

[44] x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)
```

```
from imblearn.over_sampling import SMOTE

sm = SMOTE(random_state=42)
x_resampled, y_resampled = sm.fit_resample(x_train_scaled, y_train)
```

Chia dữ liệu:

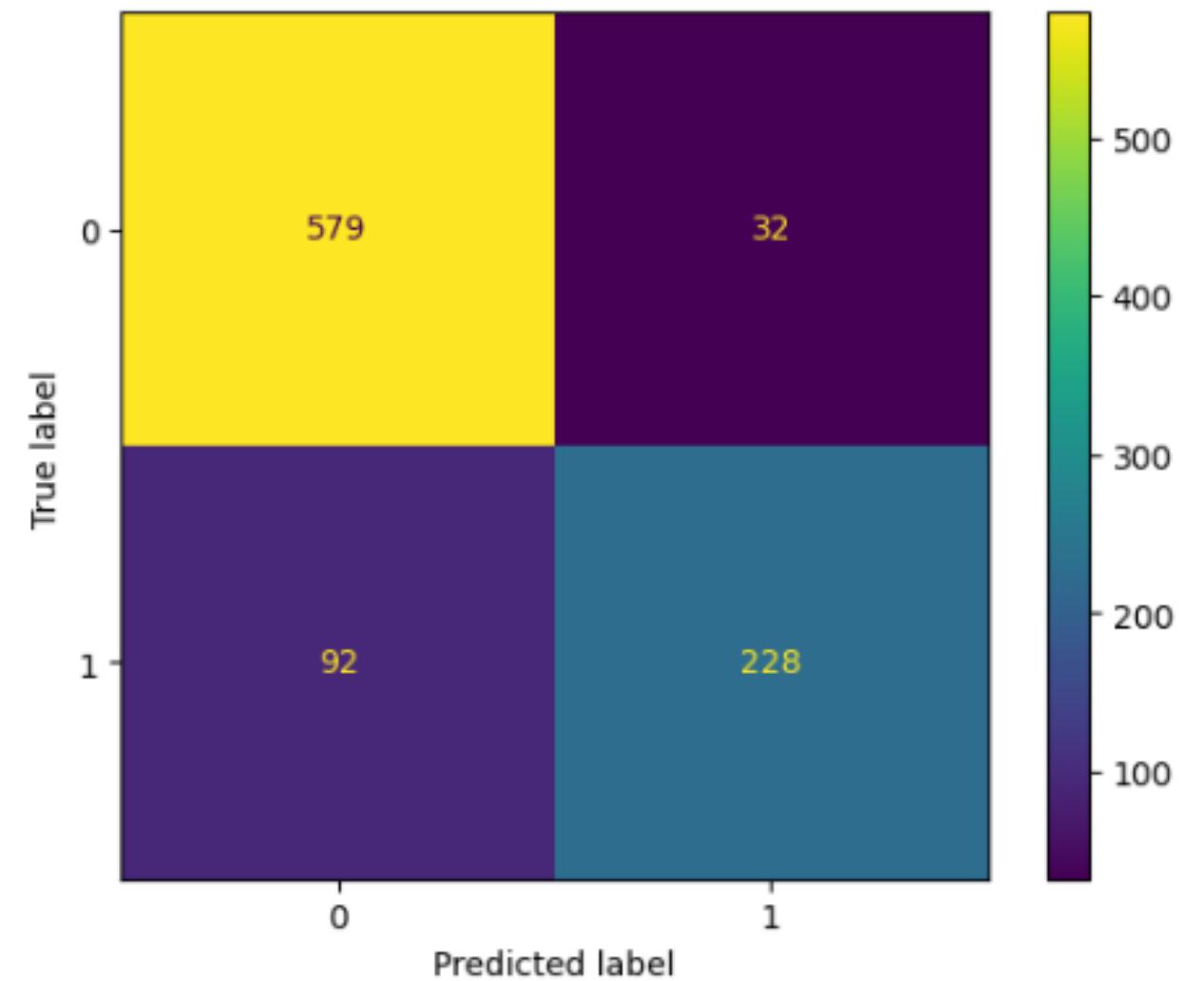
- Train/Test Split (80% / 20%).
- Áp dụng SMOTE để xử lý mất cân bằng nhãn.

Sử dụng 6 mô hình của classification để so sánh và tìm ra mô hình nào dự đoán tốt nhất:

- RandomForest, Logistic Regression, XGBoosting, LightBGM, Native Bayes, Support Vector Machine (SVM)



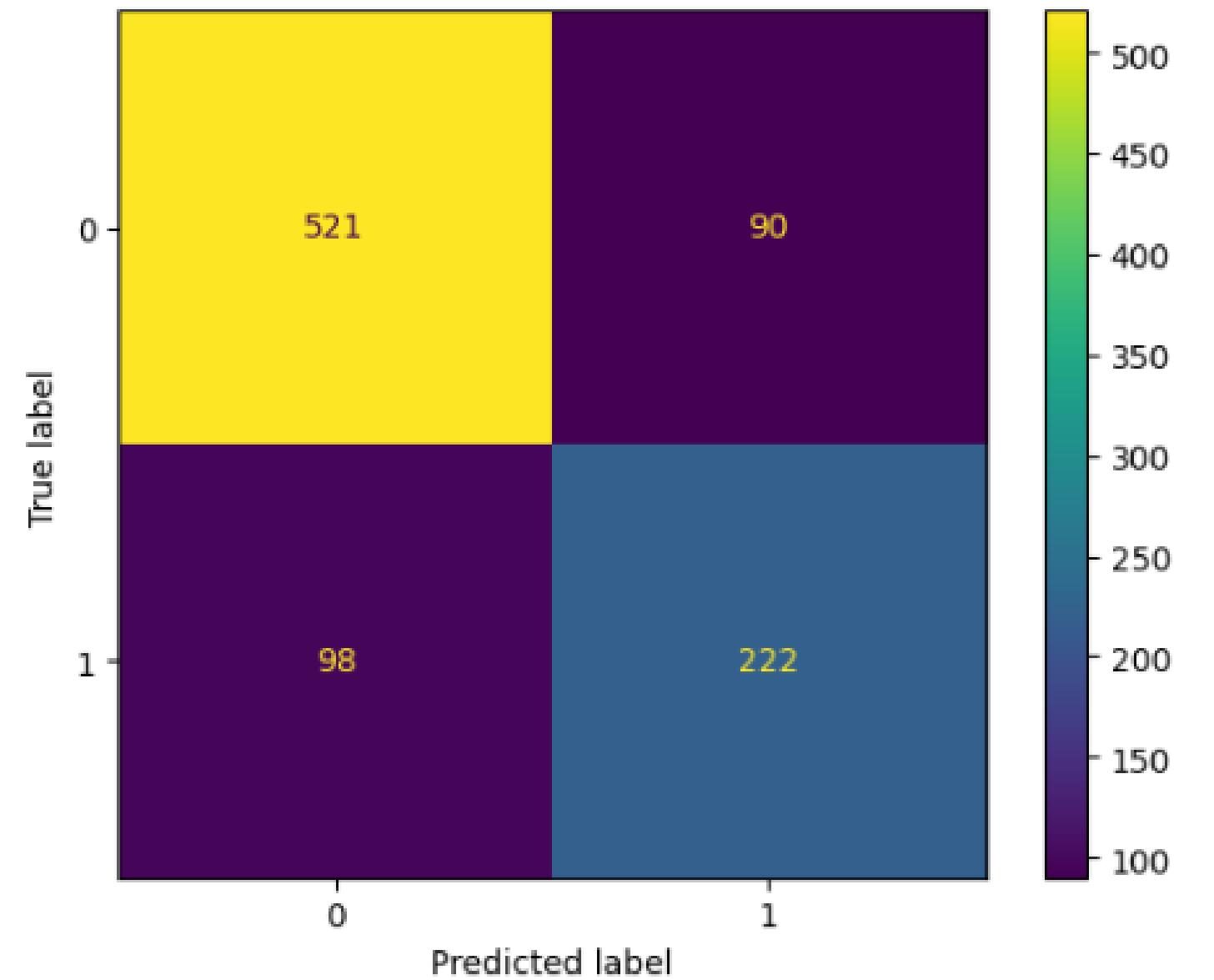
RANDOM FOREST



	precision	recall	f1-score	support
0	0.86	0.95	0.90	611
1	0.88	0.71	0.79	329



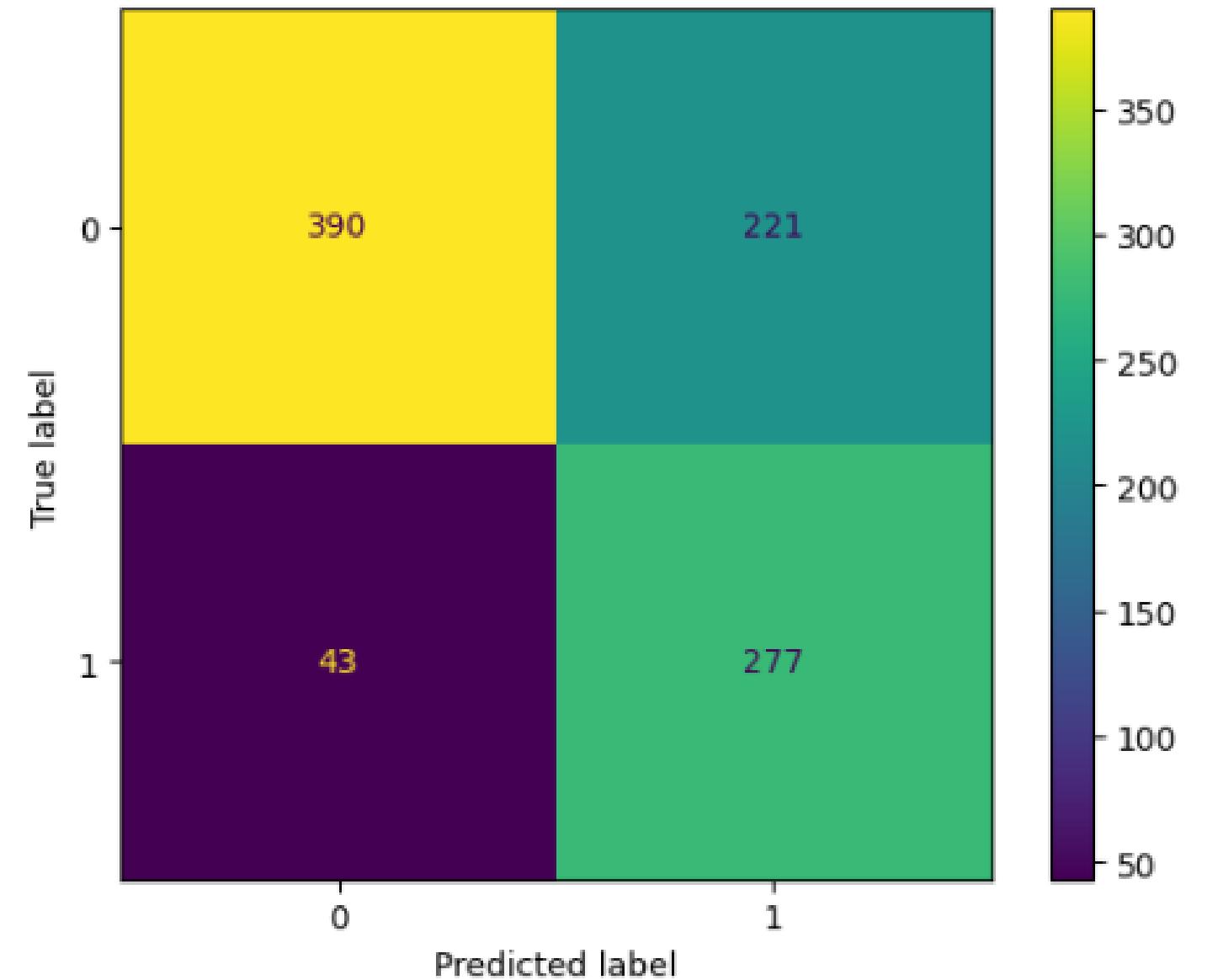
LOGISTIC REGRESSION



	precision	recall	f1-score	support
0	0.84	0.85	0.85	611
1	0.71	0.69	0.70	320



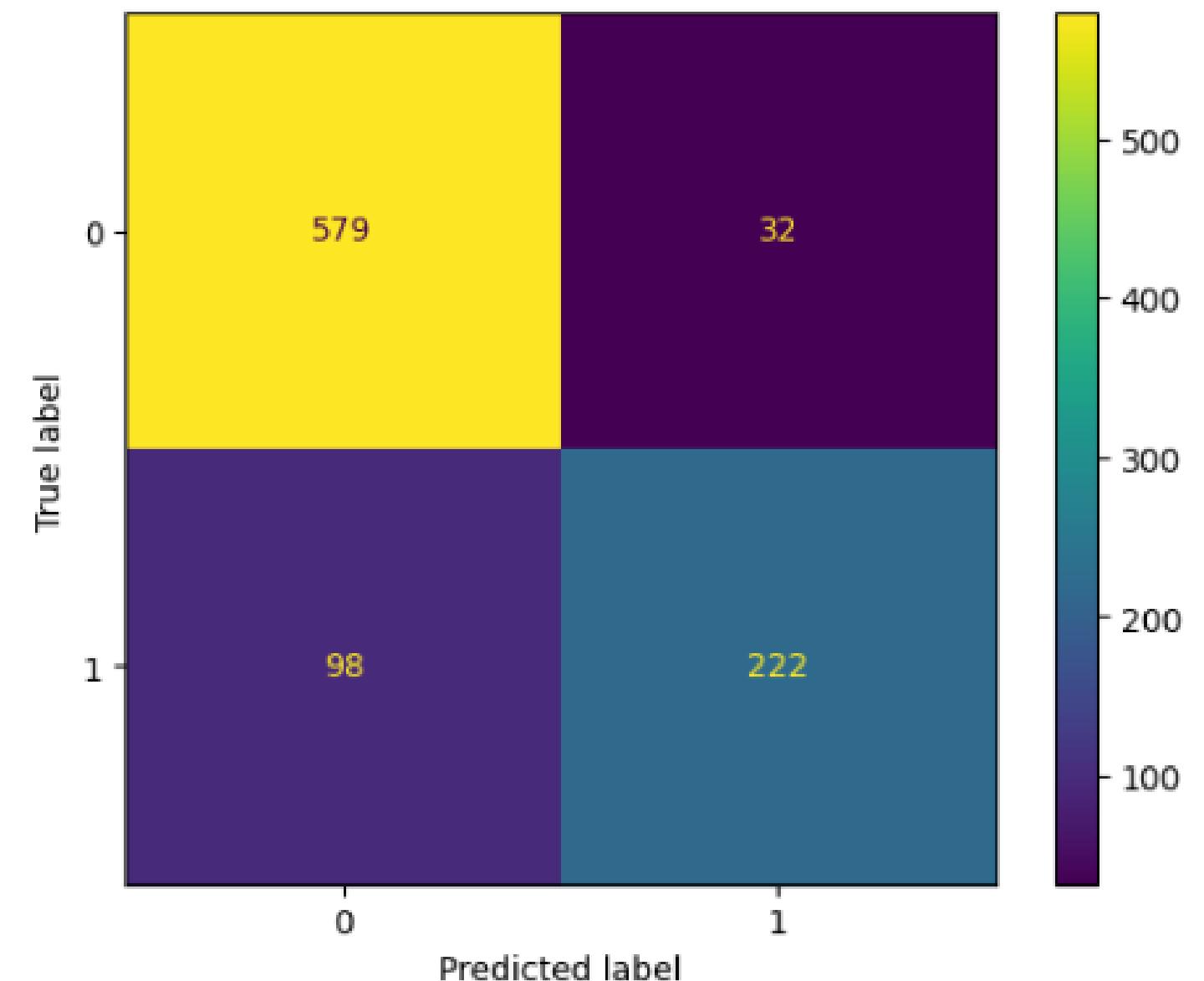
XGBOOSTING



	precision	recall	f1-score	support
0	0.90	0.64	0.75	611
1	0.56	0.87	0.68	320



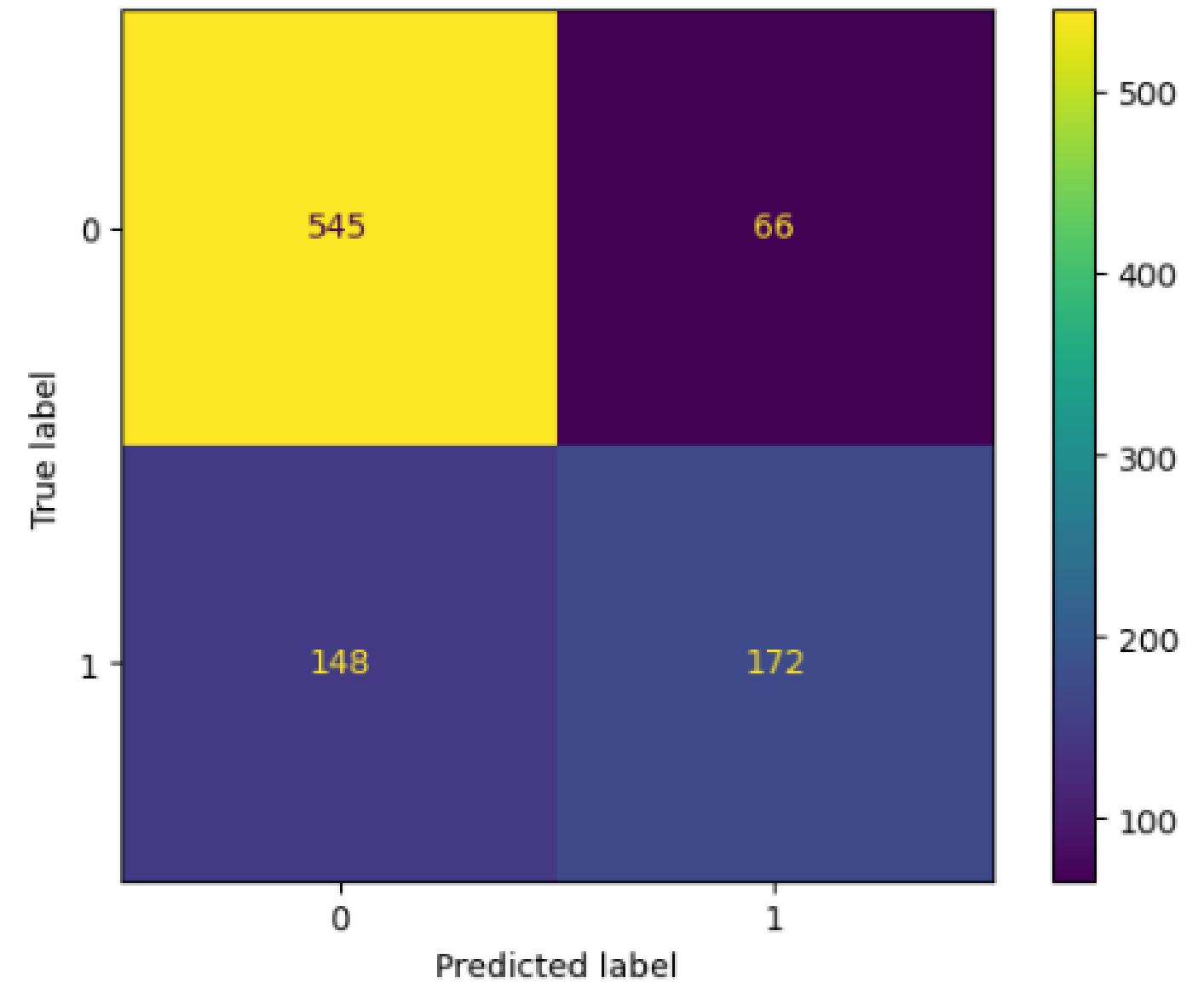
LIGHTBGM



	precision	recall	f1-score	support
0	0.86	0.95	0.90	611
1	0.87	0.69	0.77	320



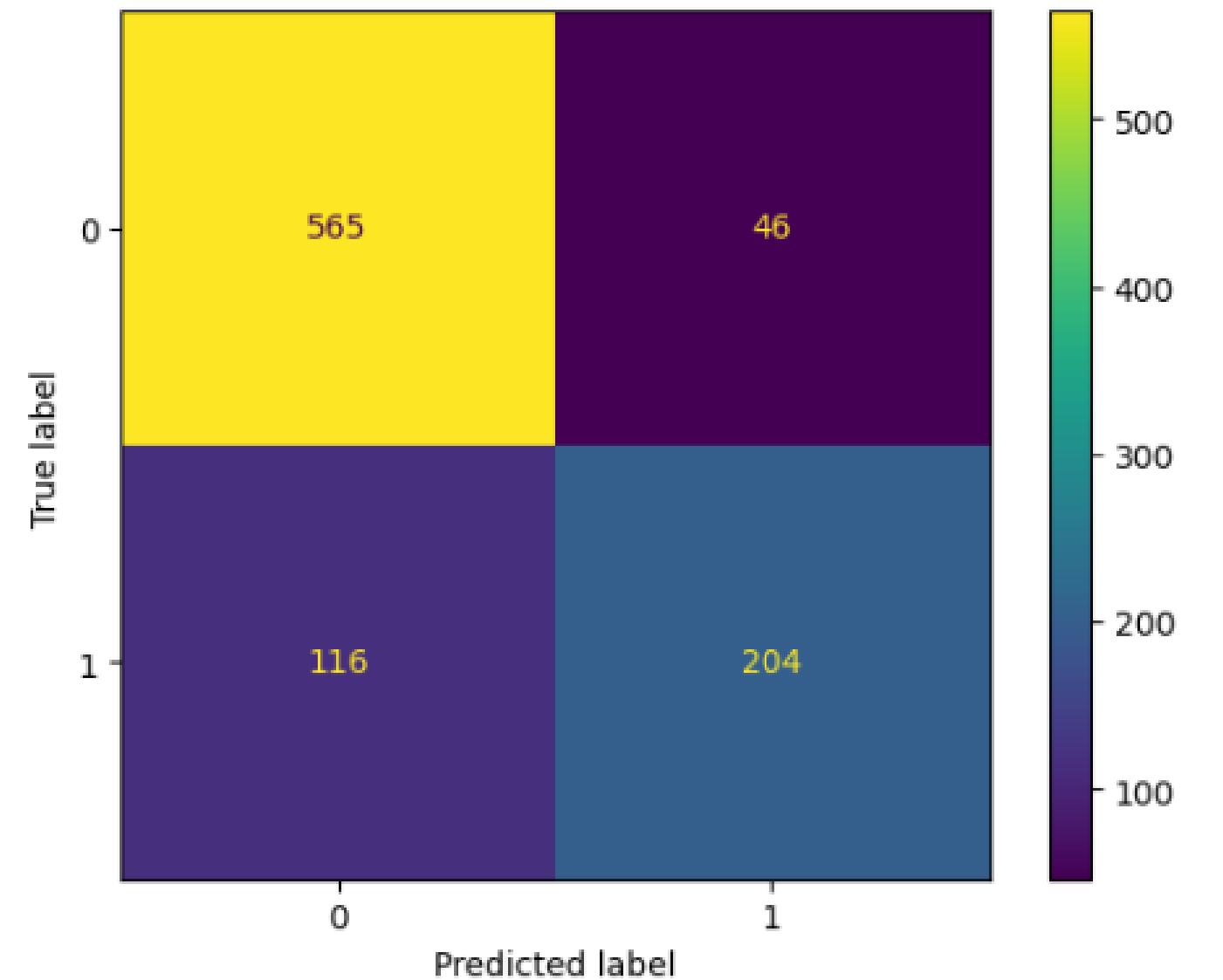
NATIVE BAYES



	precision	recall	f1-score	support
0	0.79	0.89	0.84	611
1	0.72	0.54	0.62	320



SUPPORT VECTOR MACHINE



	precision	recall	f1-score	support
0	0.83	0.92	0.87	611
1	0.82	0.64	0.72	320



TỐI ƯU VÀ ĐÁNH GIÁ

Mô hình	Precision	Recall	F1-Score
Random Forest	0.88	0.71	0.79
Logistic Regression	0.71	0.69	0.70
XGBoosting	0.56	0.87	0.68
LightGBM	0.87	0.69	0.77
Naive Bayes	0.72	0.54	0.62
SVM	0.82	0.64	0.72

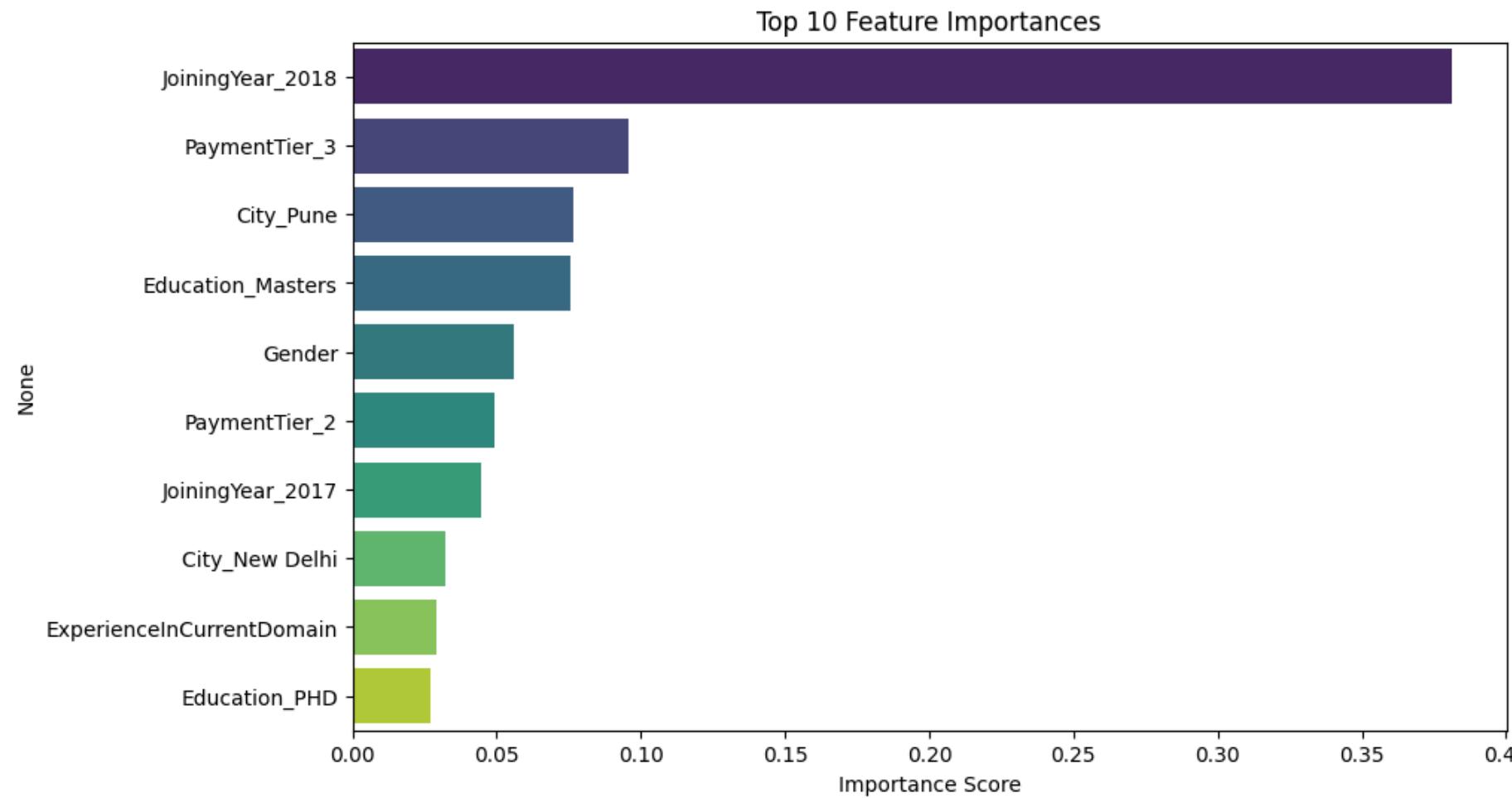
Các bước đã làm để tăng hiệu suất:

- GridSearchCV để tối ưu siêu tham số.
- Chọn thước đo Recall để tập trung vào nhóm nguy cơ nghỉ việc.

✓ Kết quả cho thấy:

- XGBoost Recall cao nhất (87%) → Khuyến nghị mô hình ưu tiên.

TOP 10 FEATURE IMPORTANCES TỪ MÔ HÌNH XGBOOSTING



Feature Importances

- JoiningYear_2018: Ảnh hưởng mạnh nhất → Tỷ lệ nghỉ việc cực cao.
- PaymentTier_3: Lương cao → Giữ chân tốt hơn.
- City_Pune: Nghỉ việc cao.
- Masters, Gender, PaymentTier_2, JoiningYear_2017: Ảnh hưởng trung bình → Nguy cơ nghỉ việc cao.
- New Delhi, Kinh nghiệm, Tuổi, PHD: Ảnh hưởng thấp.

✓ Kết luận:

Năm gia nhập, lương, địa điểm, học vấn, giới tính là yếu tố chính
→ Nên ưu tiên trong mô hình.



MỤC LỤC



I. Bối cảnh và tổng quan về dữ liệu

II. Phân tích EDA + Biểu đồ

III.Tiền xử lí dữ liệu

IV. Mô hình hóa và đánh giá mô hình



V. Ý nghĩa và kết luận, đề xuất giải pháp

Ý NGHĨA CỦA MÔ HÌNH

>>

HỖ TRỢ NHẬN DIỆN SỚM NHÂN VIÊN CÓ NGUY CƠ NGHỈ VIỆC.

>>

CUNG CẤP CƠ SỞ RA QUYẾT ĐỊNH NHÂN SỰ Dựa TRÊN DỮ LIỆU.

>>

GIẢM CHI PHÍ TUYỂN DỤNG, NÂNG CAO GIỮ CHÂN NHÂN VIÊN.



HIỆU SUẤT MÔ HÌNH ỔN ĐỊNH → CÓ THỂ ÁP DỤNG THỰC TIỄN.



KẾT LUẬN

🔍 Kết luận & Đề xuất

Các mô hình đạt Accuracy 79%–84%.

XGBoost có Recall cao nhất (0.87) → Tốt để phát hiện nguy cơ nghỉ việc.

Random Forest cân bằng giữa Recall và Precision → Phù hợp nếu cần ổn định và dễ giải thích.

✓ Khuyến nghị:

Ưu tiên XGBoost nếu cần phát hiện sớm nguy cơ nghỉ việc.

Xem xét Random Forest nếu cần giảm cảnh báo sai.

⚠ Hạn chế:

Thiếu nhiều yếu tố quan trọng (phòng ban, vị trí, chức vụ) → Mô hình chưa phản ánh đầy đủ thực tế.

Chỉ nên dùng mô hình như công cụ hỗ trợ, cần kết hợp thêm phân tích chuyên sâu khi ra quyết định.



● ĐỀ XUẤT GIẢI PHÁP ●

1. Tăng cường phúc lợi và đãi ngộ

- Cải thiện chế độ lương thưởng, đặc biệt với nhóm nhân viên có xác suất nghỉ việc cao.
- Cung cấp bảo hiểm, chăm sóc sức khỏe.

2. Lắng nghe và cải thiện môi trường làm việc.

3. Phát triển lộ trình nghề nghiệp rõ ràng

- Thiết kế lộ trình thăng tiến cụ thể cho từng nhóm nhân viên.
- Cho phép luân chuyển vị trí để tránh nhảm chán, mệt mỏi.

4. Sử dụng kết quả dự đoán từ mô hình

- Tập trung vào những nhân viên có xác suất nghỉ việc cao (theo mô hình).
- Gặp gỡ, trò chuyện cá nhân để tìm hiểu nguyên nhân.

5. Quan tâm nhóm nhân viên mới hoặc dễ nghỉ

- Dữ liệu thường cho thấy nhân viên mới, hoặc có ít năm kinh nghiệm dễ nghỉ việc hơn



THANK YOU

