EMPLOYEE TURNOVER PREDICTION WITH MACHINE LEARNING

A Data-Driven Approach to Understand and Reduce Workforce Attrition

Presented by : Zwe Thiha Naung, Saint Saint San, Soe Thandar Lwin, Wut Yee Phyo

Mentor : Ma Khin Su Myat Moe

TEAM CONTRIBUTIONS

Name	GitHub Username	Primary Roles and Contributions
Zwe Thiha Naung	@Jack243979	Data Preprocessing, Data Modeling (SVM)
Saint Saint San	@saintsaintsan	Document Preparation, Data Modeling (KNN)
Soe Thandar Lwin	@SoeThandarLwin	Data Modeling (Naïve Bayes Classifier)
Wut Yee Phyo	@ariesphyo	EDA, Data Modeling (Logistic Regression) Final Presentation

INTRODUCTION

Problem Statement : High employee turnover can significantly impact organizational performance, leading to increased hiring costs, loss of skilled talent, and productivity decline. Identifying potential attrition early is critical for effective retention strategies.

Business Context : Employee turnover is a major challenge in HR management. Predicting which employees are likely to leave enables HR departments to take proactive actions, reducing costs and improving workforce stability.

Proactive action:

- Offering career growth opportunities before employees feel unsatisfied
- Conducting employee satisfaction surveys regularly.
- Predicting turnover using data analysis and addressing causes early.

Objective : To develop a predictive model using machine learning that identifies employees at risk of leaving, enabling businesses to make data-driven retention decisions.

DATASET DESCRIPTION

Column Name	DataType	Description
satisfaction_level	float	Employee's job satisfaction score (0-1).
last_evaluation	float	Most recent performance evaluation score. (0-1)
number_project	integer.	Number of projects the employee is involved in.
average_montly_hours	integer	Average monthly working hours.
time_spend_company	integer	Number of years the employee has spent in the company.
work_accident	boolean	Whether the employee had a work accident (0 = No, 1 = Yes).
promotion_last_5years	boolean	Whether the employee got a promotion in the last 5 years (0 = No, 1 = Yes).
salary	string / object	Salary level of the employee (low, medium, high).
department	string / object	Department of the employee (e.g., sales, technical, support).
left	boolean	Whether the employee left the company.Target variable.

Size : **14,999** observations, 10 features (9 input dimensions + 1 target variable)

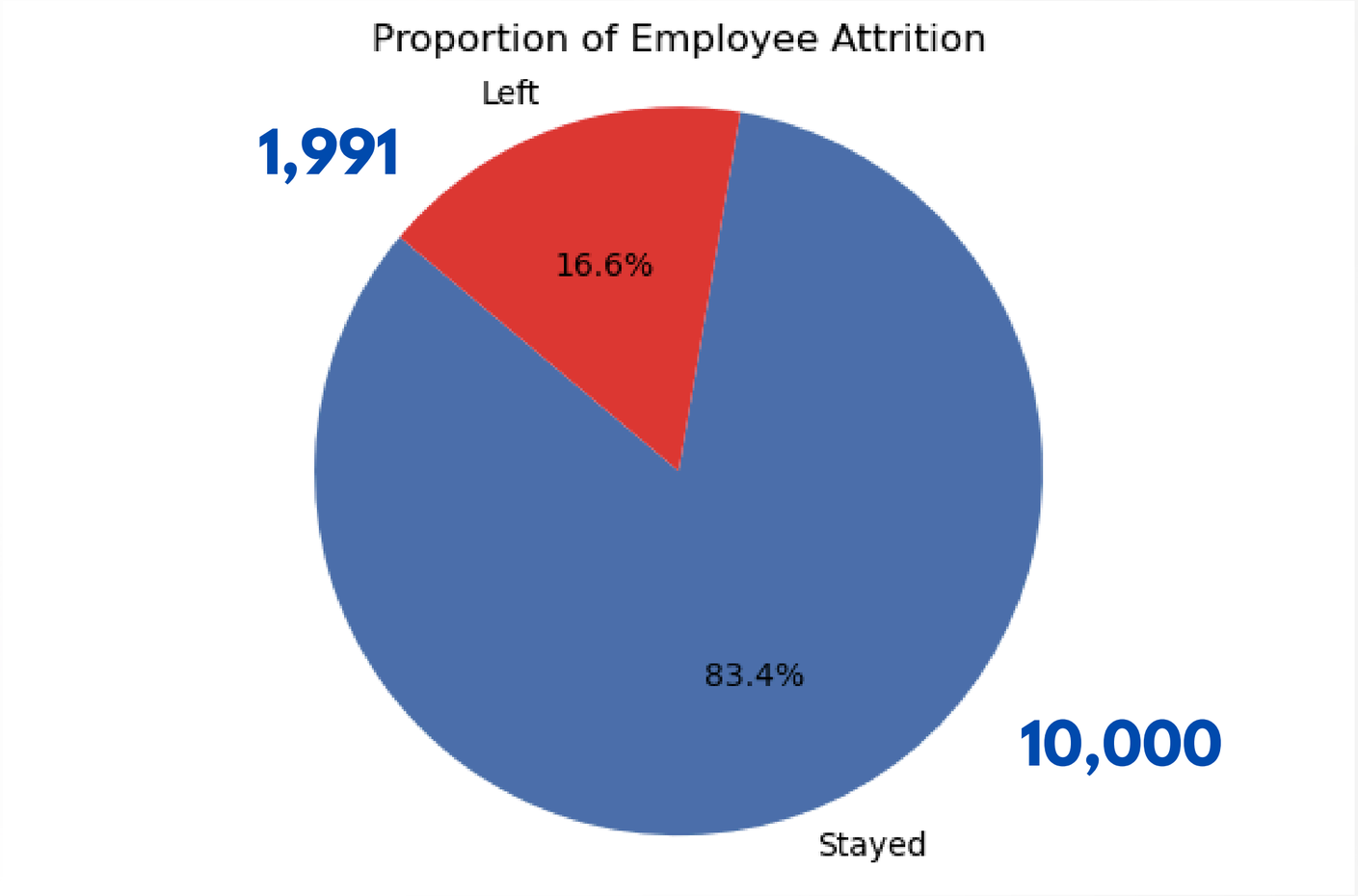
Data Source: Employee Satisfaction and Salary Dataset (Zenodo.org Record 13935313, 2024)

DATA PREPROCESSING

1. Checked and removed duplicate rows
2. Applied one-hot encoding for department (nominal : 0 and 1)
3. Applied label encoding for salary (ordinal: low < medium < high)
4. Scaled features using MinMaxScaler and StandardScaler
5. Split dataset into training (80%) and testing (20%) sets

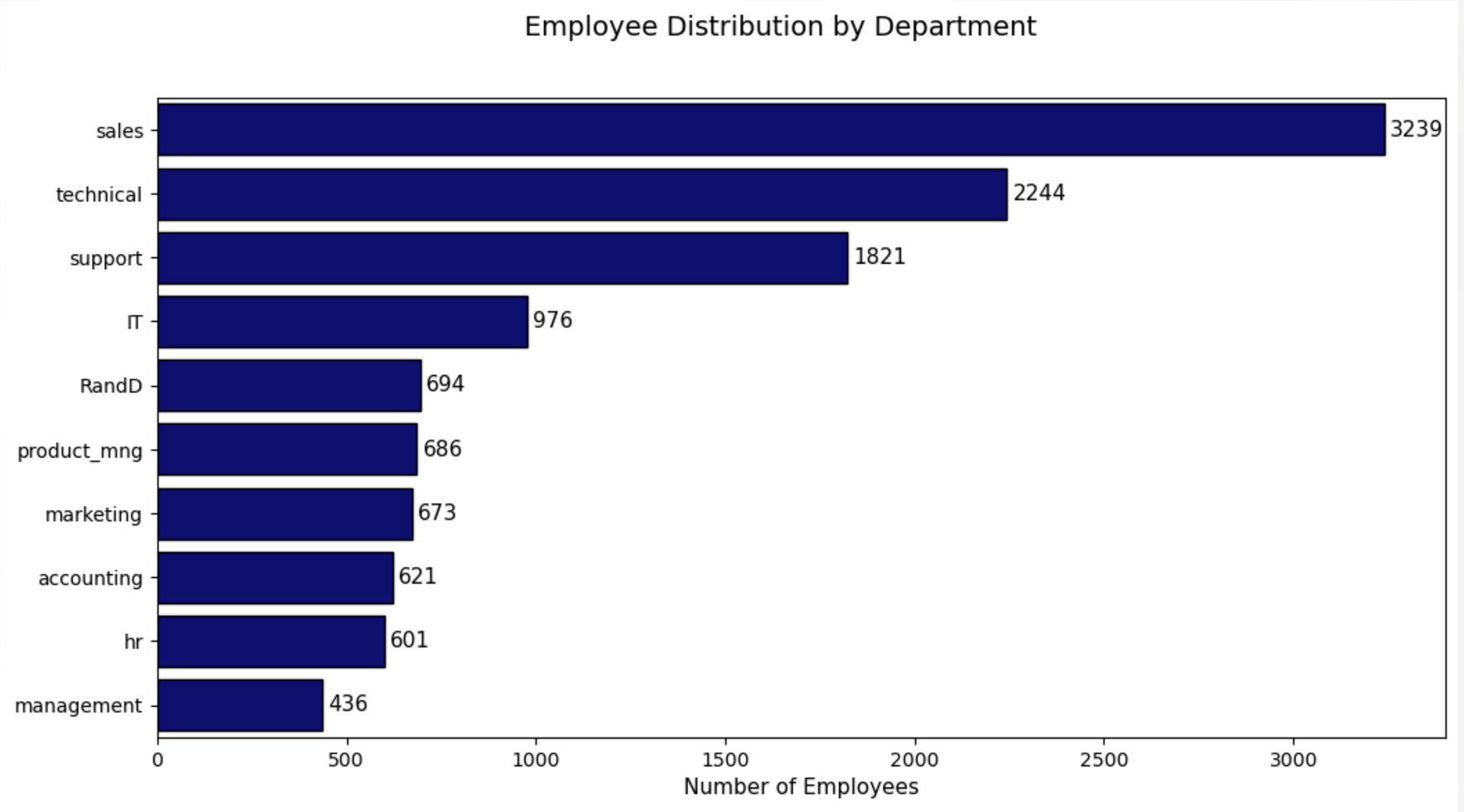
OVERALL DISTRIBUTION OF ATTRITION

What percentage of employees have left the company vs. stayed?



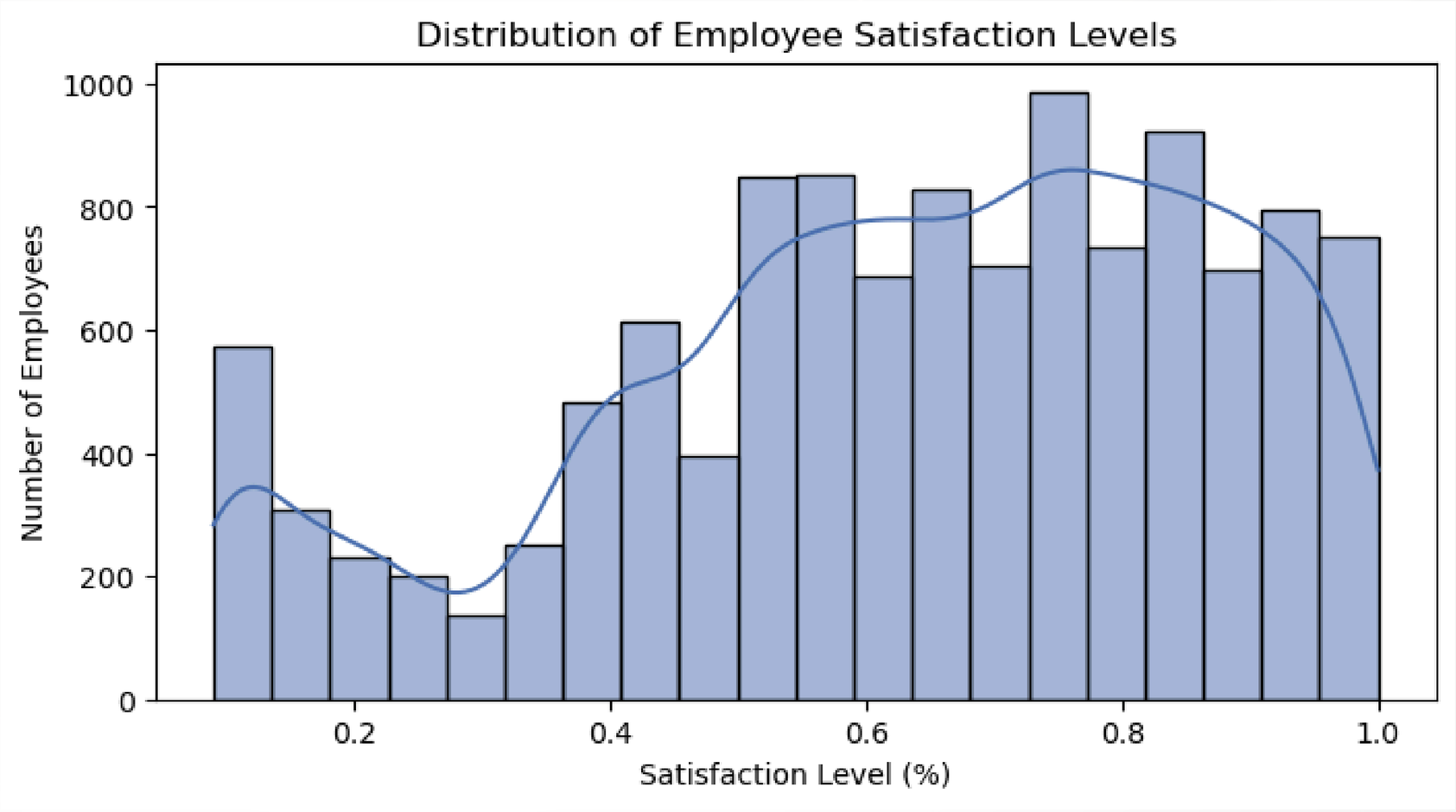
EMPLOYEE DISTRIBUTION BY DEPARTMENT

How are employees distributed across different departments?



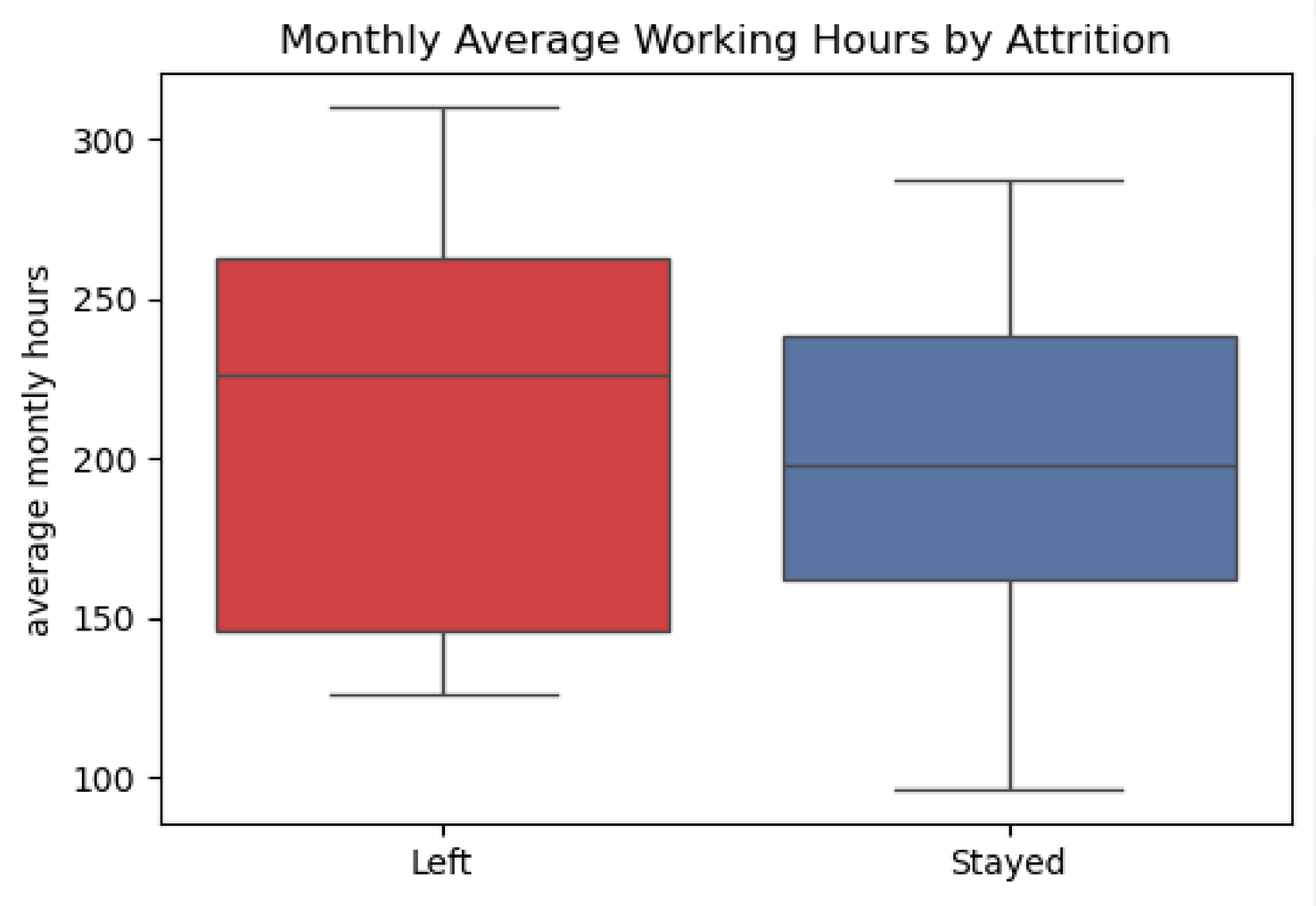
EMPLOYEE DISTRIBUTION BY SATISFACTION LEVEL

How does the distribution of satisfaction levels vary among employees?



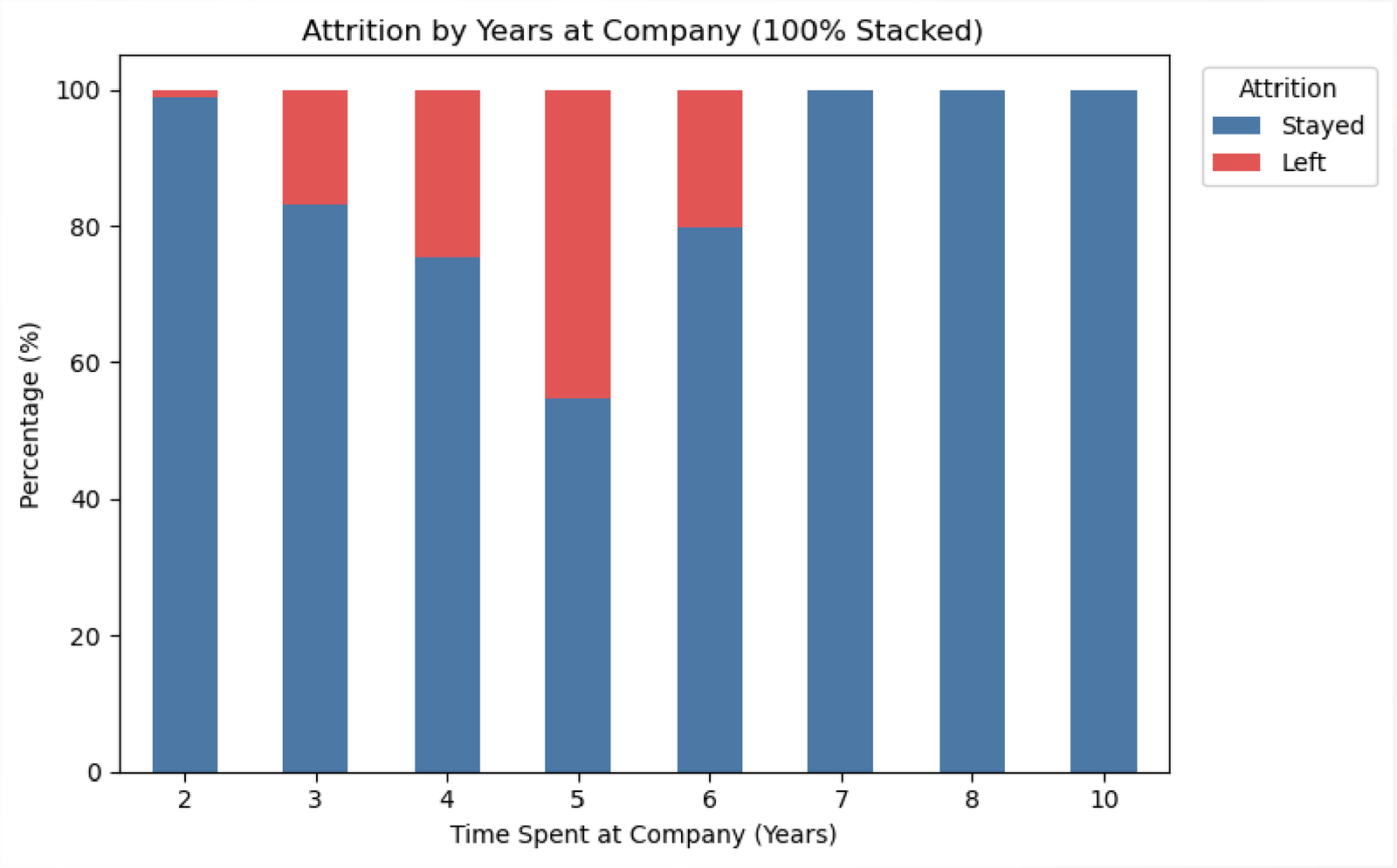
AVERAGE MONTHLY HOURS AND ATTRITION

Do employees who left the company have more working hours?



TIME SPENT IN COMPANY AND ATTRITION

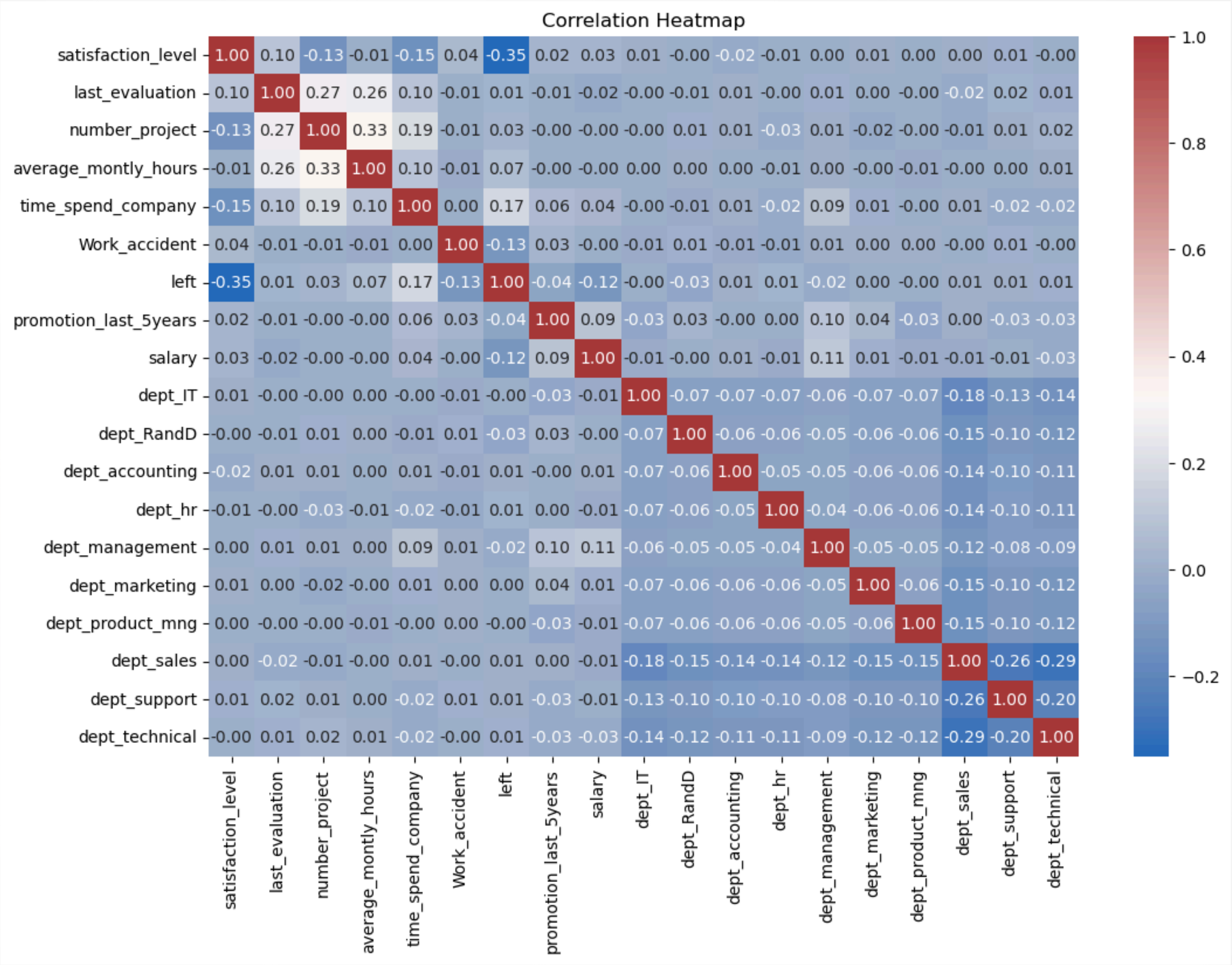
Does time spent at the company affect attrition?



SALARY AND ATTRITION

Are employees with lower salary more likely to leave?

Which features are strongly correlated with attrition?



EDA SUMMARY

- Employees with **lower satisfaction** and **more working hours** have a higher chance of leaving.
- Employees with **low salaries** are much more likely to leave compared to medium or high salaries.
- Employees with **≤3 years or ≥7 years** at the company exhibit relatively low attrition rates.
- Most features show weak to moderate correlations (±0.1 to ±0.4), indicating **non-linear patterns** rather than strictly linear relationships.

MODEL SELECTION

Logistic Regression

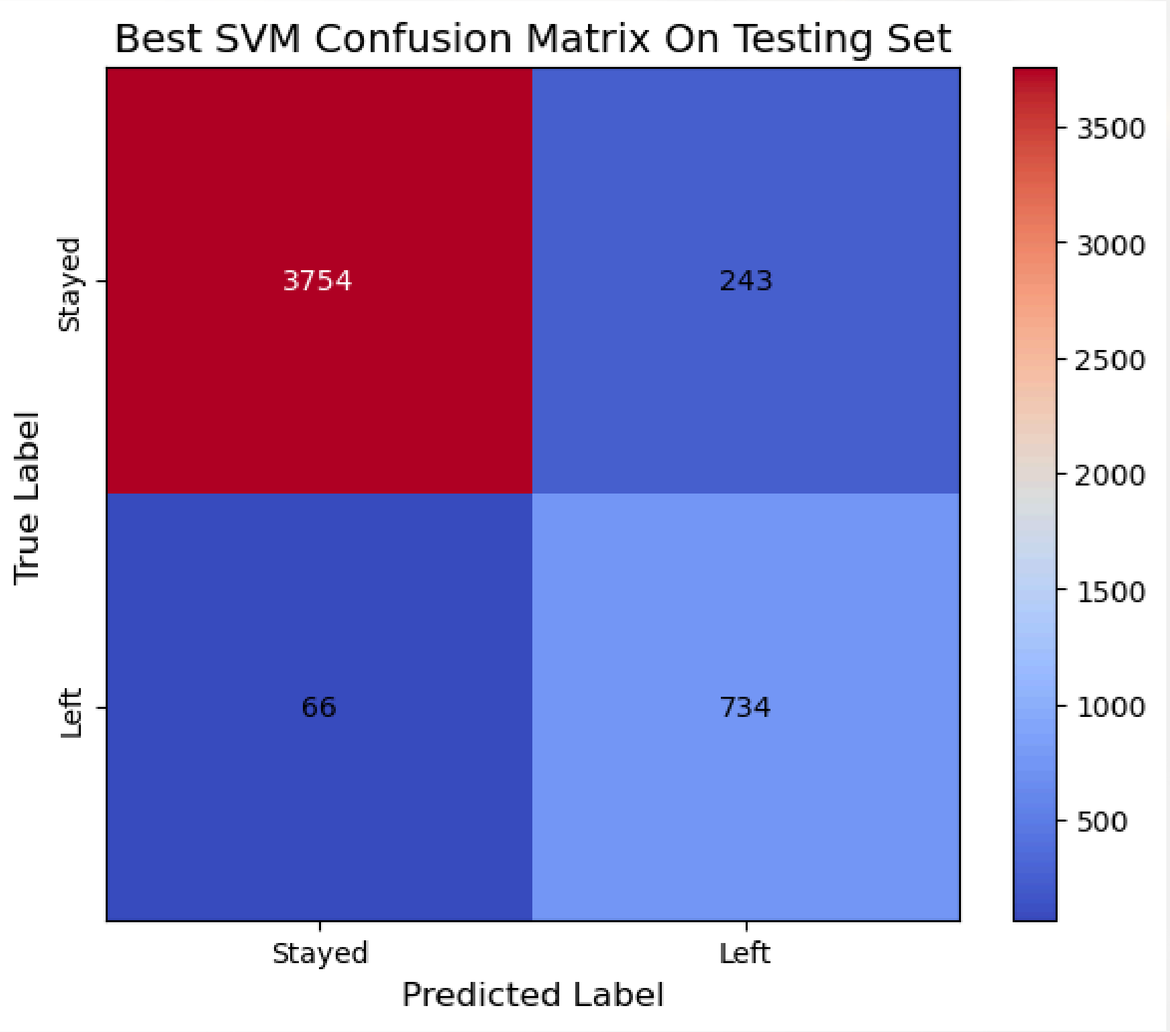
K-Nearest Neighbor

Support Vector Machine

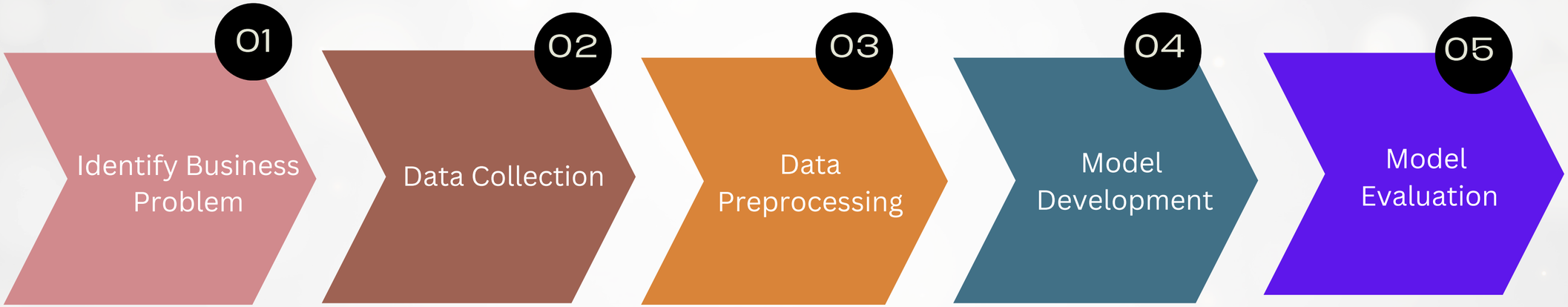
Naive Bayes Classifier

BEST MODEL : SVM CONFUSION MATRIX

class	precision	recall	f1-score	support
0	0.98	0.94	0.96	3997
1	0.75	0.92	0.83	800



IMPLEMENTATION PROCESS



- Identify employee turnover as a major issue in HR management

- Collect employee dataset from Zenodo's official source

Data Preparation

- Clean data, encode categorical variables, normalize features.

Perform EDA

- Analyze patterns, class imbalance, and correlations.

Train/Test Split

- Split the dataset into training and testing sets

Build a Pipeline

- Scaling +
- Sampling +
- Modeling

GridSearchCV

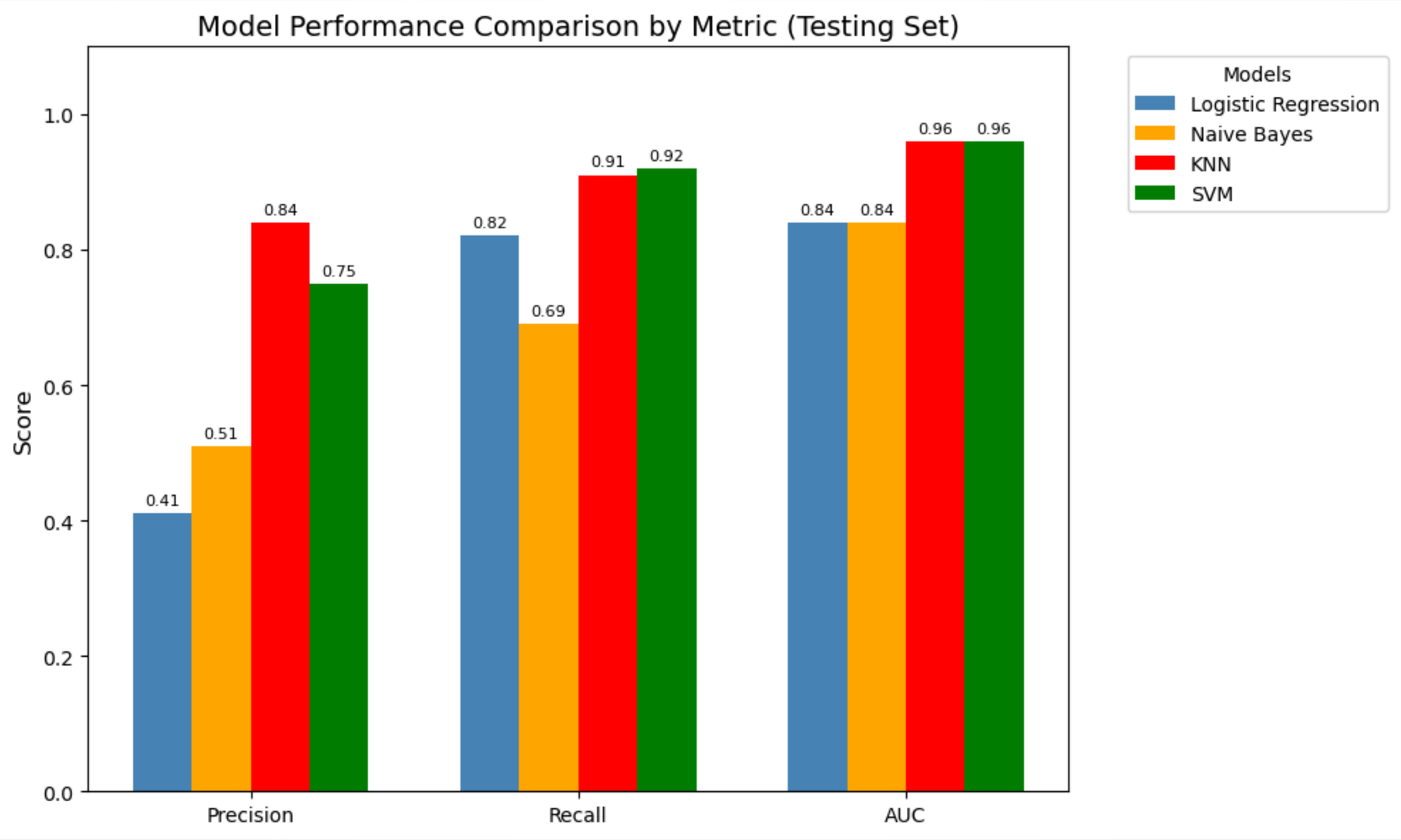
- Hyperparameter tuning

- Evaluate model performance with Precision, Recall, F1, AUC.

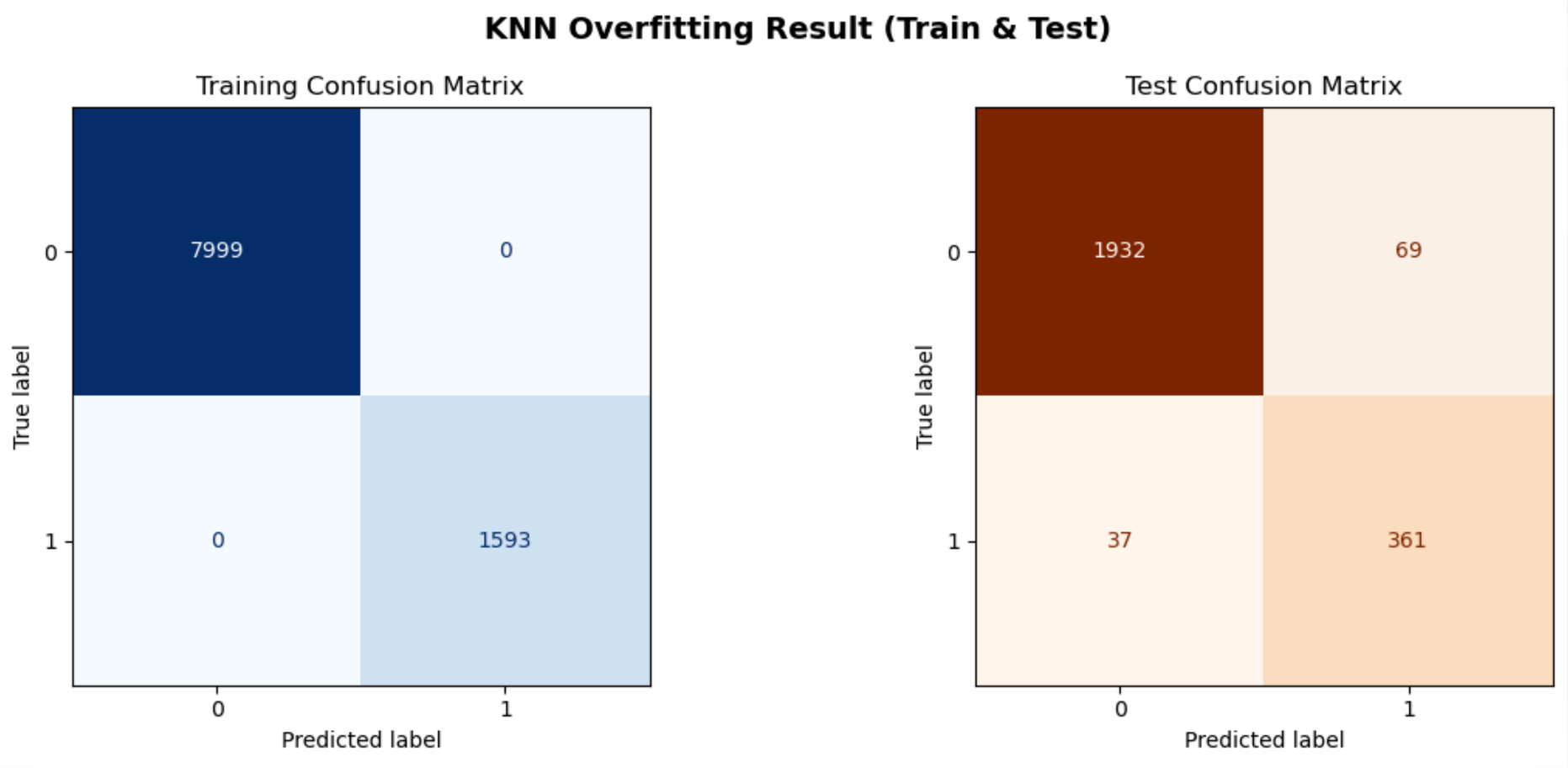
USED LIBRARIES

- **Pandas** (Data analysis and manipulation)
- **NumPy** (Mathematical computations)
- **Scikit-learn** (Machine learning models & evaluation)
- **Imbalanced-learn** (Handling class imbalance with SMOTE)
- **Matplotlib & Seaborn** (Data visualization)

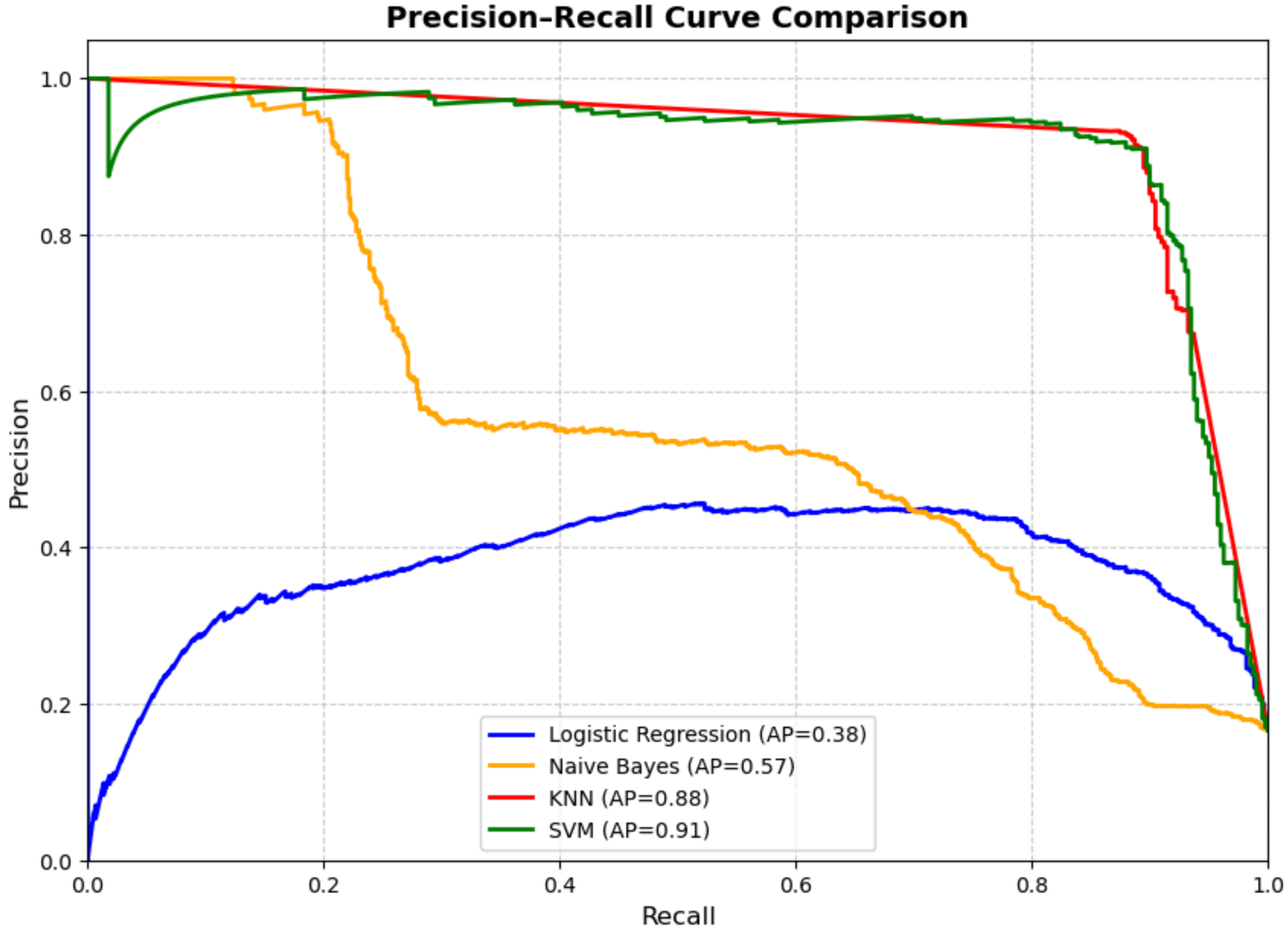
EVALUATION METRICS



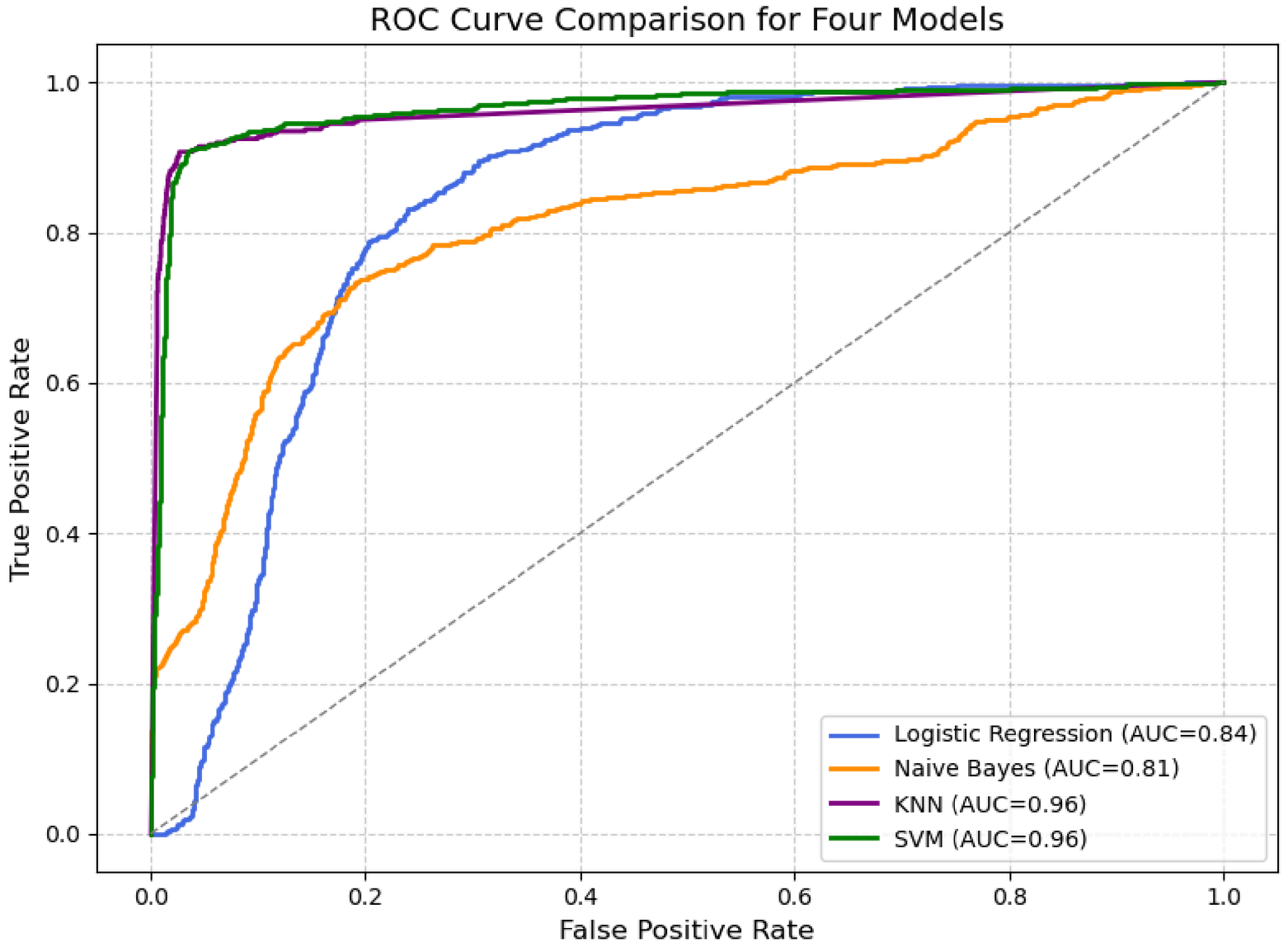
KNN : OVERFITTING RESULT



RESULTS & FINDINGS



RESULTS & FINDINGS



DISCUSSION OF RESULTS

Balanced Performance : SVM achieved a strong balance between precision, recall, and F1-score for both classes, especially for minority class (employees who left), avoiding bias toward majority class.

Handles Non-linearity : By using the RBF kernel, SVM effectively captured the non-linear relationships in the dataset, which other models like Logistic Regression and Naive Bayes struggled with.

Better Generalization : Unlike KNN, which showed signs of overfitting with 100% training accuracy, SVM maintained high accuracy on testing data while avoiding overfitting, ensuring reliable performance on unseen data.

Robustness to Class Imbalance : Even with imbalanced data, SVM delivered consistently high recall for the minority class, making it more effective in identifying employees who are likely to leave.

CONCLUSION

Project Insight & Business Relevance

Employee turnover is a major challenge for many businesses in Myanmar, and this project shows how Machine Learning can predict attrition using Myanmar employee data.

Business Impact

Early Risk Identification : Our SVM model predicts employees at risk of leaving by uncovering hidden data patterns (e.g., workload, satisfaction, tenure), enabling HR to take proactive retention actions.

Data-Driven Decisions : Instead of relying on guesswork, HR teams can use model predictions to prioritize high-risk employees and design targeted interventions (e.g., incentives, career development plans).

HR SUGGESTIONS

- Overworking is a strong factor driving attrition as HR should balance workloads and limit excessive overtime.
- The data indicates that attrition peaks between three and six years of tenure; therefore, HR should prioritize career development, promotions, and engagement initiatives during this period to mitigate turnover.
- Salary dissatisfaction or unfair compensation contributes to turnover, calling for a review of compensation to remain competitive.

FUTURE WORKS

- Plan to extend this project by collecting key features (e.g., satisfaction level, number of projects, working hours, tenure, work accidents, promotions, department, salary, left) from Myanmar employees.
- Collect data anonymously from employees via Google Forms, without including any personally identifiable information (PII).
- Experiment with other models like Random Forest and XGBoost.

THANK YOU

QUESTIONS ARE WELCOME!