EMPLOYEE ATTRITION PREDICTION PROJECT REPORT

1. Introduction

Employee attrition is a significant challenge for organizations as it can impact productivity, morale, and costs.
This project aims to predict whether an employee is likely to leave the company using historical HR data. Predictive insights allow HR teams to take proactive retention measures.

Dataset:

- WA_Fn-UseC_-HR-Employee-Attrition.csv
- 1,470 employees, 35 features including Age, Department, Job Role, Monthly Income, Overtime, Years at Company, and more.

Objective:

- Build a predictive model to classify employees as likely to **stay** or **leave**.

2. Data Exploration and Cleaning

- **Missing Values:** None — dataset is complete.
- **Numerical Analysis:** Histograms and summary statistics were used to understand the distribution of features like Age, Distance from Home, Monthly Income, and Total Working Years.
- **Categorical Analysis:** Columns such as Department, Job Role, Marital Status, and OverTime were reviewed for patterns in attrition.

3. Feature Engineering

- **Dropped Irrelevant Columns:**
    - HourlyRate (covered under Monthly Income)
- **Added New Features:**
    - **PromotionGap** = Years at Company − Years Since Last Promotion
        - Represents the duration since an employee's last promotion.
    - **WorkloadPressure** = Job Level × Overtime indicator (1 if Yes, 0 if No)

        - Captures employee stress/workload.

- **Encoding Categorical Variables:**
    - All categorical variables were transformed using one-hot encoding for machine learning.

4. Model Building

- **Target Variable:**
    - Attrition (Yes → 1, No → 0)
- **Features:** All columns except Attrition after encoding.
- **Train-Test Split:**
    - 80% training, 20% testing, random state = 42.
- **Model Used:** Logistic Regression
    - Max iterations = 1000 to ensure convergence.

5. Model Evaluation

Metric	Result
Accuracy	87%
Precision (Left)	56%
Recall (Left)	26%
F1 Score (Left)	35%

- **Confusion Matrix:**

	Predicted Stayed	Predicted Left
Actual Stayed	247	8
Actual Left	29	10

- **Interpretation:**

- Model correctly identifies employees who stay most of the time.
- Recall for employees who leave is low (26%), indicating the model misses some leaving employees.
- Overall accuracy is high (87%), but improvements are needed for better attrition detection.

6. Visualizations

1. **Numerical Feature Distributions:**
    - Histograms of Age, Monthly Income, Total Working Years, etc.
2. **Confusion Matrix:**
    - Heatmap showing true positives, true negatives, false positives, and false negatives.

7. Tools and Libraries

- Python 3
- Pandas & NumPy — Data manipulation
- Seaborn & Matplotlib — Visualization
- Scikit-learn — Machine learning

8. Conclusions

- Logistic Regression achieved **87% overall accuracy**, effectively identifying employees who stay.
- Model struggles with predicting employees who leave (low recall), suggesting further feature engineering or alternative models could improve performance.
- Key takeaways:
    - Employee workload and promotion gaps may be important predictors of attrition.
    - Overtime combined with job level can highlight stressed employees.

9. Recommendations / Future Work

1. **Scale numerical features** to improve logistic regression convergence.
2. **Test alternative models**: Random Forest or Gradient Boosting for better predictive power.
3. **Feature Importance Analysis** to determine which factors drive attrition most.
4. **HR Actionable Insights**:
    - Monitor employees with high WorkloadPressure and long PromotionGap.
    - Implement retention strategies targeting high-risk employees.