**Executive Summary**

Employee turnover is a critical issue for organizations due to its impact on productivity, costs, and workplace morale. This project aims to leverage data-driven approaches using Python and machine learning to understand the key factors that drive employee attrition. By analyzing employee satisfaction, workload, performance, and organizational factors, we develop predictive models that help HR departments identify at-risk employees and take proactive retention measures. The insights from this analysis will serve as a foundation for formulating better HR strategies and improving employee engagement and retention.

**Objective and Problem Statement**

**Objective:**

The primary objective of this project is to **predict whether an employee is likely to leave the company** and to **identify the most influential factors contributing to attrition** using machine learning techniques.

**Problem Statement:**

High employee turnover rates can significantly affect a company's performance and financial health. Without understanding why employees leave, companies cannot build effective retention strategies. This project addresses the following questions:

* What employee-related factors are most associated with attrition?
* Can we build a model that accurately predicts whether an employee will leave?
* What actionable insights can be derived to reduce employee turnover?

**Data Description and Assumptions**

**Dataset Overview:**

The dataset contains records of 14,999 employees with the following features:

| **Feature** | **Description** |
| --- | --- |
| satisfaction\_level | Employee job satisfaction (0 to 1) |
| last\_evaluation | Last performance evaluation score (0 to 1) |
| number\_project | Number of projects the employee has worked on |
| average\_montly\_hours | Average monthly hours worked |
| time\_spend\_company | Number of years with the company |
| Work\_accident | Whether the employee had a work accident (0 = No, 1 = Yes) |
| left | Target: Whether the employee left (1 = Yes, 0 = No) |
| promotion\_last\_5years | Was promoted in the last 5 years (0 = No, 1 = Yes) |
| Department | Department name (e.g., sales, HR, IT) |
| salary | Salary category (Low, Medium, High) |

**Assumptions:**

* There are no duplicate employee records.
* Missing values, if any, are minimal and handled appropriately.
* salary is an ordinal variable (Low < Medium < High).
* The dataset represents a complete and balanced picture of the company's employee history at a point in time.

**🔍 Methodology**

**1. Data Preprocessing**

* **Missing Values**: Detect and impute/drop missing data.
* **Categorical Encoding**:
  + Department: One-hot encoding.
  + salary: Ordinal encoding (Low = 0, Medium = 1, High = 2).
* **Feature Scaling**:
  + Normalize numerical features (satisfaction\_level, last\_evaluation, etc.) for models sensitive to scale.
* **Train-Test Split**: Split dataset into training and testing subsets (e.g., 80/20).

**2. Exploratory Data Analysis (EDA)**

* **Univariate Analysis**: Histograms and density plots to examine distributions.
* **Bivariate Analysis**: Correlation heatmap, boxplots, and bar charts to explore relationships with left.
* **Trend Analysis**:
  + Attrition vs Satisfaction Level
  + Attrition by Department and Salary
  + Project workload and promotions vs Attrition

**3. Modeling**

* Train classification models:
  + Logistic Regression (baseline)
  + Random Forest
  + Gradient Boosting (e.g., XGBoost)
  + Support Vector Machine (SVM)
  + Multi-layer Perceptron (Neural Network)
* Use cross-validation to tune hyperparameters.

**4. Evaluation**

* Metrics:
  + **Accuracy**
  + **Precision & Recall**
  + **F1-Score**
  + **ROC-AUC Score**
  + **Confusion Matrix**
* Compare model performance and select the best one.

**5. Interpretation & Insights**

* Use:
  + **Feature importance** from tree-based models.
  + **SHAP** values for model interpretability.
* Identify key drivers (e.g., low satisfaction, long work hours, lack of promotion).
* Recommend HR strategies based on insights.