Predicting Employee Retention - Project Report

1. Problem Statement

Employee retention is a critical factor for organizational success, particularly in the competitive technology sector where the demand for skilled professionals is high. High turnover leads to increased costs in recruitment, on boarding, and productivity loss. Therefore, a mid-sized technology company aims to proactively predict employee retention using historical data, enabling them to improve workforce stability and satisfaction.

2. Methodology

To address the employee retention prediction problem, a structured approach was adopted:

- **Data Collection**: HR data including demographics, job satisfaction, tenure, and performance.
- **Data Exploration**: Summary statistics, missing values analysis, and visualization.
- **Data Pre-processing**: Encoding categorical variables (e.g., department, job role), and standardizing numeric features (e.g., salary, tenure).
- **Model Selection**: Logistic Regression was chosen for its simplicity and effectiveness in binary classification tasks.
- **Model Training**: The dataset was split into 80% training and 20% testing subsets.
- **Model Evaluation**: Accuracy, confusion matrix, precision, recall, and F1-score were calculated.

3. Techniques Used

- **Exploratory Data Analysis (EDA)**: Used to uncover initial trends and patterns.
- **One-Hot Encoding**: To convert categorical data into numeric format suitable for machine learning models.
- **StandardScaler**: Used to normalize features and improve model performance.
- **Logistic Regression**: Selected for its interpretability and suitability for binary outcomes.
- **Model Validation**: Employed metrics such as accuracy and classification reports to ensure the model's effectiveness.

4. Visualisations

Several visual tools were used to gain insight into the dataset:

- **Correlation Heatmap**: Illustrated relationships between features, helping to identify redundancy.
- **Count Plots**: Compared retention across categories such as department or education level.
- **Boxplots**: Highlighted distributions and outliers in numeric variables like tenure or satisfaction score.

These visualizations provided both confirmation of assumptions and discovery of new patterns.

5. Insights

The model and visual analysis uncovered the following insights:

- High satisfaction and performance scores strongly correlate with retention.
- Newer employees (shorter tenure) are more likely to leave, indicating that early engagement is key.
- Some departments exhibit higher attrition rates, suggesting department-specific challenges.
- Features like job level, marital status, and education also influence retention decisions.

6. Actionable Outcomes

From the insights, the following actions are recommended:

- Use model predictions to identify and retain high-risk employees.
- Design personalized career and development plans to improve satisfaction.
- Conduct targeted surveys or interviews in high-turnover departments.
- Continue updating and retraining the model with new data to maintain relevance.