CI Project

GROUP MEMBERS –

ADITYA AMBURE – 202301070154

SNEHA KABRA – 202301070162

KRISHNA KEDAR – 202301070177

DEVANG DESHMUKH – 202301070170

DATA PREPROCESSING AND EDA REPORT

Data Preprocessing Steps:

1. Loaded the dataset and verified column data types.

2. Filled missing numeric values using median.

3. Filled missing categorical values with the label 'Missing'.

4. Converted categorical columns to the correct data type.

5. Dropped irrelevant or constant columns such as EmployeeCount, StandardHours, Over18, and EmployeeNumber.

6. Performed one-hot encoding on all categorical features using pandas.get_dummies().

7. Engineered new features:

   - YearsAtCompany_by_Age = YearsAtCompany / (Age + 1)

   - YearsSinceLastPromotion_flag = 1 if YearsSinceLastPromotion > 0 else 0

8. Created the target variable Attrition_flag by mapping Yes → 1 and No → 0.

9. Balanced the dataset using SMOTE to prevent bias toward majority classes.

10. Scaled all numeric features using StandardScaler.

EDA Insights:

- Attrition distribution showed class imbalance; SMOTE corrected it.
- Younger employees (18–30) showed higher attrition compared to older employees.
- Employees working overtime had significantly higher attrition.
- MonthlyIncome and JobLevel showed a positive correlation.
- Satisfaction scores (JobSatisfaction, EnvironmentSatisfaction) directly influenced attrition.
- Heatmap revealed clusters of correlated variables such as:
  - TotalWorkingYears vs MonthlyIncome
  - JobLevel vs MonthlyIncome
  - Age vs YearsAtCompany

Visualizations Created:

- Attrition countplot
- Department vs Attrition comparison
- Age group attrition bar chart
- Heatmap of numeric correlations