
📑 Preprocessing Guide for Employee Attrition Dataset

🎯 Target Column

- **Attrition (Yes/No)** → Label Encoding
    - Mapping: Yes → 1, No → 0

🔹 Feature Types & Encoding Strategy

1. Numerical Features

Columns:
Age, DistanceFromHome, MonthlyIncome, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, HourlyRate, MonthlyRate, PercentSalaryHike

- **Encoding:** None
- **Scaling:**
    - ✅ Apply **StandardScaler** for Logistic Regression, SVM, KNN, Neural Networks.
    - ❌ Not required for tree-based models (Random Forest, XGBoost, LightGBM).

---

2. Ordinal Features (natural order exists)

Columns:
Education, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, PerformanceRating, StockOptionLevel

- **Encoding:** Already integers (1–5 scale), **keep as is**.
- **Scaling:** None

---

3. Nominal Features (no natural order)

✅ **Low Cardinality → One-Hot Encoding**

- Gender (Male/Female)
- OverTime (Yes/No)
- BusinessTravel (3 categories)

- MaritalStatus (3 categories)
- Department (3 categories → One-Hot or Frequency Encoding okay)

✅ **High Cardinality → Target/Frequency Encoding**

- JobRole (9 categories)
- EducationField (6 categories)

📊 **Summary Table**

Feature Type	Columns	Encoding	Scaling
Numerical	Age, DistanceFromHome, MonthlyIncome, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, HourlyRate, MonthlyRate, PercentSalaryHike	None	StandardScaler (for linear/distance models only)
Ordinal	Education, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, PerformanceRating, StockOptionLevel	Keep as is (1–5 scale)	None
Nominal (Low-Cardinality)	Gender, OverTime, BusinessTravel, MaritalStatus, Department	One-Hot Encoding	None
Nominal (High-Cardinality)	JobRole, EducationField	Target/Frequency Encoding	None
Target	Attrition	Binary (0/1)	None

⚡ **Model-Specific Notes**

- **Tree Models (RandomForest, XGBoost, LightGBM):**
  - No scaling needed.
  - Use One-Hot + Target Encoding.
- **Linear / Distance Models (Logistic Regression, SVM, KNN, Neural Networks):**
  - Scaling is required for numeric features.
  - Same encoding strategy for categorical.

✅ This is your **final roadmap**:

- Apply encodings as per feature type.
- Scale numericals only if your model requires it.
- Target column stays binary (0/1).
