Modern Academy
الأكاديمية الحديثة للهندسة والتكنولوجيا

AMIT

DATA SCEINCE

Employee HR Dataset

PRESENTED TO

AMIT Learning

PRESENTED BY

AI Team

JULY2025

Table of Contents

Team Members

1. Razan Ihab Abdel-Latif

   CODE:4230554

2. Abdel-Rahman Ahmed Hussein

   CODE:4230786

3. Mennah Abdel-Maroof

   CODE:4230811

4. Mohamed Khaled Mondy

   CODE:4230799

Dataset Overview

The Employee_HR.csv dataset contains information about 14,999 employees across various departments. Each record includes attributes related to employee satisfaction, performance evaluations, work history, compensation, and whether they left the company (churn).

Key Features:

- **EmpId**: Unique employee identifier
- **Satisfaction**: Employee satisfaction score (0.9-9.2)
- **Evaluation**: Performance evaluation score (4.5-10.0)
- **number_of_projects**: Number of projects assigned (2-7)
- **average_montly_hours**: Average monthly hours worked (126-310)
- **time_spent_company**: Years at company (2-6)
- **work_accident**: Whether employee had a work accident (0/1)
- **Promotion**: Whether employee was promoted (0/1)
- **Department**: Employee department (sales, accounting, hr, etc.)
- **Salary_INR**: Employee salary in Indian Rupees
- **Churn**: Whether employee left (1) or stayed (0)

Data Collection

THE DATASET APPEARS TO BE COLLECTED FROM HR SYSTEMS TRACKING:

- Employee performance metrics
- Work history and engagement
- Compensation data
- Turnover information

POTENTIAL COLLECTION METHODS:

- HRIS (Human Resource Information System) exports
- Performance management system data
- Employee satisfaction surveys
- Payroll system records

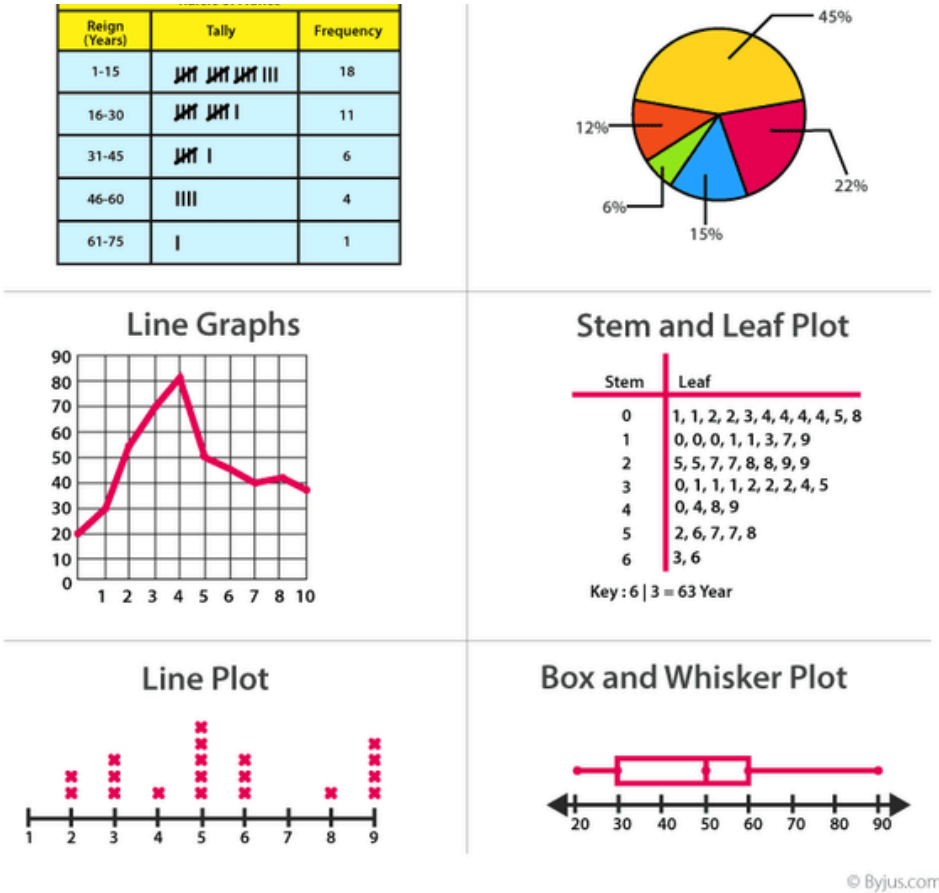
Data Representation

THE DATA IS STRUCTURED IN A TABULAR FORMAT WITH:

- **11 columns (features)**
- **14,999 rows**

MIXED DATA TYPES:

- Numerical: Satisfaction, Evaluation, Salary_INR
- Categorical: Department, work_accident, Promotion
- Binary: Churn



Data Wrangling

Issues Identified:

-Some numerical values show floating-point precision artifacts (e.g., "5.300000000000002")

-Potential outliers in Salary_INR (values ranging from 10,026 to 386,458 INR)

-Some departments have very few representatives

Cleaning Required:

- Normalize floating-point representations
- Handle potential outliers in salary data
- Consider department consolidation for rare categories

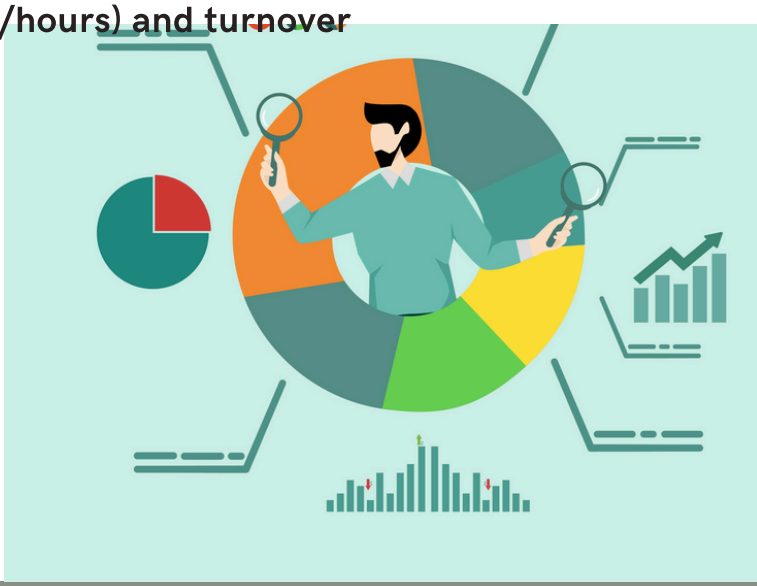
Data Analysis

PRELIMINARY INSIGHTS:

- The dataset only contains churned employees (Churn=1 for all records)
- Wide range in satisfaction (0.9-9.2) and evaluation scores (4.5-10.0)
- Significant variance in monthly hours (126-310)
- Salary distribution appears right-skewed

POTENTIAL ANALYSIS DIRECTIONS:

- Correlation between satisfaction and churn
- Impact of promotions on retention
- Department-wise churn patterns
- Relationship between workload (projects/hours) and turnover



Data Preprocessing

Feature Engineering:

- Create workload ratio (hours/projects)
- Calculate tenure-based features
- Normalize salary data

Handling Missing Values:

- Dataset appears complete but should verify

Scaling:

- Standardize numerical features (Satisfaction, Evaluation, etc.)

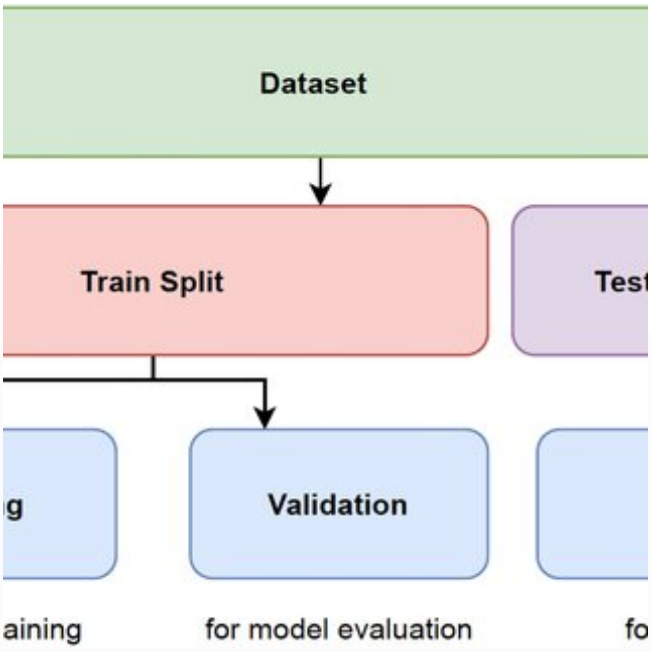
Categorical Encoding:

- One-hot encoding for Department
- Binary encoding for work_accident, Promotion

Data Encoding and Data Splitting

Data Encoding

- One-Hot Encoding: For Department (10+ categories)
- Binary Encoding: For work_accident and Promotion
- Numerical Scaling: StandardScaler for continuous features



Data Splitting

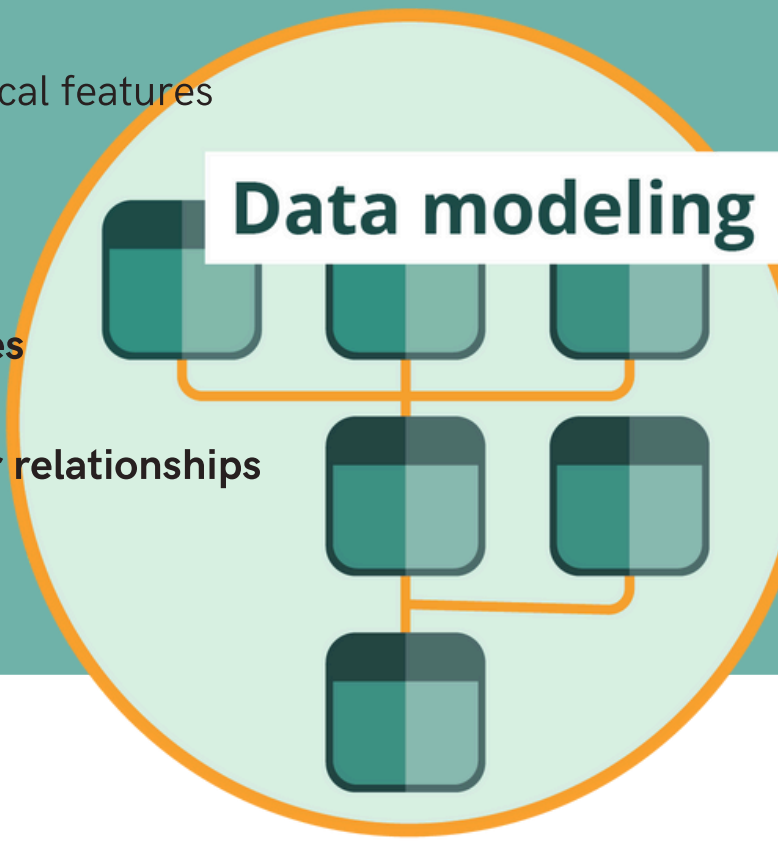
- Additional data with Churn=0 for proper binary classification
- If obtained, standard 70-30 or 80-20 train-test split
- Stratified sampling to maintain class balance

Data Modeling

Predictive Modeling (Classification)

Logistic Regression:

- Baseline model for binary classification (churn prediction)
- Interpretable coefficients showing feature importance
- Works well with standardized numerical features



Support Vector Machines (SVM):

- **Effective for high-dimensional spaces**
- **Kernel tricks can capture non-linear relationships**
- **Sensitive to feature scaling**

K-Nearest Neighbors (KNN):

- **Distance-based model using Euclidean/Manhattan distance**
- **Useful for capturing local patterns in employee behavior**
- **Requires careful selection of k and distance metric**

Data Modeling

Linear Regression:

- **Predict continuous outcomes like satisfaction or salary**
- **Could model relationship between hours/projects and satisfaction**
- **Provides interpretable coefficients**

Ensemble Models

Bagging [BaggingClassifier - RandomForestClassifier - ExtraTreesClassifier]

Boosting [XGBoostClassifier - CatBoostClassifier - LGBMClassifier]

Voting [VotingClassifier]

Stacking [StckingClassifier]

Implementation Considerations

Feature Importance:

Key features likely to impact models:

- **Satisfaction and Evaluation scores**
- **Workload metrics (projects/hours)**
- **Tenure and promotion history**
- **Department affiliation**

Model Evaluation Metrics:

- **For classification: Accuracy, Precision, Recall, F1, ROC-AUC**
- **For regression: RMSE, R-squared**

Model Performance Benchmarks

Model Performance Benchmarks

This table summarizes the performance of various machine learning models, from foundational algorithms to advanced ensemble and boosting methods, in predicting employee churn.

98.87%	98.73%	98.53%	98.53%
StckingClassifier	ExtraTreesClassifier	RandomForestClassifier	LGBMClassifier
98.50%	98.23%	98.17%	97.50%
XGBoostClassifier	CatBoostClassifier	BaggingClassifier	Decision Tree
95.83%	94.63%	77.13%	76.40%
KNeighborsClassifier	SVC	Logistic Regression	LinearSVC

VotingClassifier Accuracy

VotingClassifier Accuracy

98.90%

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	0.99	1.00	0.99	2294
1 (Churn)	0.99	0.96	0.98	706
Total	0.99	0.98	0.98	3000

Recommended Approach

Given the current dataset (all churned employees), the most valuable initial analyses would be:

Descriptive Analytics:

- **Department-wise churn patterns**
- **Satisfaction/evaluation distributions**
- **Workload analysis**

Regression:

- **Model satisfaction based on workload features**

Conclusion

This HR dataset provides valuable insights into employee churn patterns. While currently limited to churned employees, it offers opportunities for segmentation and workload analysis. Implementing a combination of clustering and (with additional data) classification models can help organizations understand and predict turnover risks, enabling proactive retention strategies.