

# Employee Attrition Prediction Model

This documentation outlines the process of developing a predictive model for employee attrition. It encompasses comprehensive dataset analysis, preprocessing steps, model development, and evaluation results. Below is a summary of the key components:

## Dataset Analysis:

- **Summary Statistics:**
  - Provides an overview of statistical measures like mean, standard deviation, min, max, etc., for numerical features.
- **Missing Values:**
  - Shows the count of missing values in each column.
- **Dataset Shape:**
  - Indicates the number of rows and columns in the dataset.
- **Columns:**
  - Lists all the column names in the dataset.
- **Data Types:**
  - Shows the data type of each column.

## Preprocessing Steps:

1. **Dropping Unnecessary Columns:**
  - Removed columns: 'EmployeeCount', 'EmployeeNumber', 'StandardHours', 'Over18'.
  - These columns might not contribute to the prediction task or have constant values.
2. **Encoding Categorical Variables:**
  - Used one-hot encoding for categorical features: 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'.

- Used ordinal encoding for 'BusinessTravel' with categories: 'Non-Travel', 'Travel\_Rarely', 'Travel\_Frequently'.

### 3. Encoding Target Variable:

- Encoded 'Attrition' as binary: 'Yes' as 1 and 'No' as 0.

### 4. Splitting Data into Train and Test Sets:

- Split the dataset into training and testing sets with a ratio of 80:20.
- Used a random state of 42 for reproducibility.

### 5. Feature Scaling:

- Standardized numerical features using **StandardScaler**.
- Standardization transforms data to have a mean of 0 and a standard deviation of 1.

## Model Development:

- **Model:**
  - Used Logistic Regression for binary classification.
  - Logistic Regression is a simple and effective algorithm for binary classification tasks.
- **Model Training:**
  - Trained the Logistic Regression model using the training data.
- **Model Evaluation:**
  - Evaluated the model using the test data.
  - Metrics used: Accuracy, Precision, Recall, F1-score.

## Evaluation Results:

- **Accuracy:**
  - Accuracy measures the ratio of correctly predicted instances to the total instances.
  - Accuracy: 0.891 (or 89.1%).
- **Precision:**

- Precision measures the ratio of correctly predicted positive observations to the total predicted positive observations.
- Precision: 0.621 (or 62.1%).
- **Recall:**
  - Recall (or sensitivity) measures the ratio of correctly predicted positive observations to the all observations in actual class.
  - Recall: 0.462 (or 46.2%).
- **F1-score:**
  - F1-score is the weighted average of Precision and Recall. It considers both false positives and false negatives.
  - F1-score: 0.529 (or 52.9%).

These evaluation metrics give an indication of how well the model is performing in predicting employee attrition based on the given features.