|  |  |
| --- | --- |
| **Course- B. Tech** | **Type-** Specialization Elective |
| **Course Code-** CSET371 | **Course Name-** Big Data Analytics and Business Intelligence |
| **Year-** 2024 | **Semester**- Odd |
|  | **Session-** 2022-2026 |

**Student’s Name: Ayush Maurya Enrollment No.: E22CSEU1586 Batch: 49**

**Problem Statement**

**“Predicting Employee Attrition in an Organization”**

# INTRODUCTION

Employee attrition, therefore, is an ongoing issue, cut across industries, because attrition directly impacts productivity since it breaks the chain of work and involves significant costs for replacements-recruitment and training, above all. Experienced employees have essential organizational knowledge, increase team cohesiveness, and facilitate effective and consistent supply of products and services. High levels of attrition can signal underlining organizational problems and even job dissatisfaction. Perhaps the compensation is not sufficient, and sometimes, the work-life demands are not met. Dealing with issues described above demands better recognition of factors that will influence an employee's staying or leaving the organization.

This project will focus on using the power of data analytics and machine learning to predict attrition in employees as well as generate actionable insights. With a comprehensive analysis of demographic, job-related, compensation, and engagement features, we will identify patterns and relationships that will contribute to attrition. The approach allows HR departments to identify at-risk employees early on, thus enabling them to take targeted retention strategies. Predicting attrition reduces turnover costs, but also increases employee satisfaction and the capacity of an organization to hold onto the best talent in a competitive market. These tools take what is usually a reactionary challenge and make it into a proactive opportunity for business improvement.

# OBJECTIVES

**Identify the major determinants of attrition.**

Analyze the traits of demographics, job, compensation, and engagement as an underpinning for developing drivers of employee turnover.

**Build a reliable predictive model.**

Train and test classification models on whether the employee will continue or not to work with the company.

**Understanding Organizations Composition**

Missing values, class imbalance, and feature scaling need to be addressed in the optimization process of the dataset toward good machine learning.

**Understanding Organizations Composition**

Which people are working, what age group, gender they belong to and what plans company offers to which employees.

**Provide actionable insights**

Feature importance analysis and outputs from these models can help HR departments understand the specific areas for intervention.

**Support HR strategies.**

It helps the human resource teams implement evidence-based retention programs, work-life balance optimization, and compensation policy changes for at-risk employees.

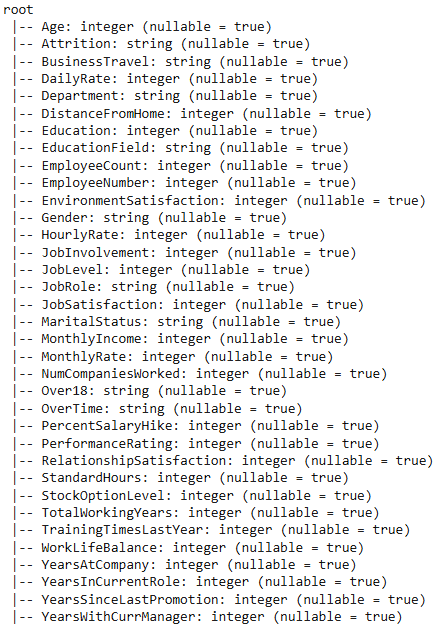
# DATASET

The provided employee attrition dataset contains information about the employees, their work environment, and personal details; hence, the main objective of this dataset is to know factors that contribute to attrition, whether an employee leaves or stays with the organization. Here is a brief overview of the structure of the dataset:

1. Demographics and Personal Details such as Age, Gender and MaritalStatus
2. Work Details such as BusinessTravel (Frequency of business travel), Department, JobRole, JobLevel, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
3. Compensation and Benefits: Salary details ( DailyRate, HourlyRate, MonthlyIncome, MonthlyRate), PercentSalaryHike, StockOptionLevel
4. Performance and Satisfaction: JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, PerformanceRating, JobInvolvement
5. Work-Life Balance and Training: WorkLifeBalance, TrainingTimesLastYear
6. Attrition: Whether the employee left the company (Yes or No).
7. Other Factors: DistanceFromHome, Education, OverTime

The dataset has 1470 rows and 35 columns.

Following is the schema of the dataset:



# DATA PREPROCESSING

1. **Remove redundant columns**Dropped all unnecessary columns "EmployeeCount," "EmployeeNumber," "StandardHours," and "Over18.". These columns were redundant for analysis or contained only a single constant value, e.g. "Over18" contained only one value that is ‘Y’.
2. **Check Missing Value.**Checked for missing values in all columns by using isNull() and confirmed that there are no missing values.
3. **Column Classification:**Identified numerical columns (int, double, float) and categorical columns (string) based on data types for separate handling.
4. **One-Hot Encoding (OHE):**Applied StringIndexer to convert categorical columns into numerical indices. The columns that were converted are: Gender, JobRole, Department, BusinessTravel, OverTime, MaritalStatus.

OneHotEncoder was used to create one-hot encoded features from the indexed columns.

1. **Extracted Features**  
   Introduced new features using column transformation  
   "Tenure" = YearsAtCompany - YearsInCurrentRole.  
   "IncomeToJobLevelRatio" = MonthlyIncome / JobLevel.
2. **Target Variable Transformation**  
   This column, "Attrition," is converted to numerical "Attrition\_index" using StringIndexer on it,
3. **Correlation Analysis and Feature Selection:**  
   Computed the correlation of combined numerical and indexed categorical columns within a feature vector using VectorAssembler.

Features dropped with low correlation with target “Attrition\_index” (threshold = 0.03).

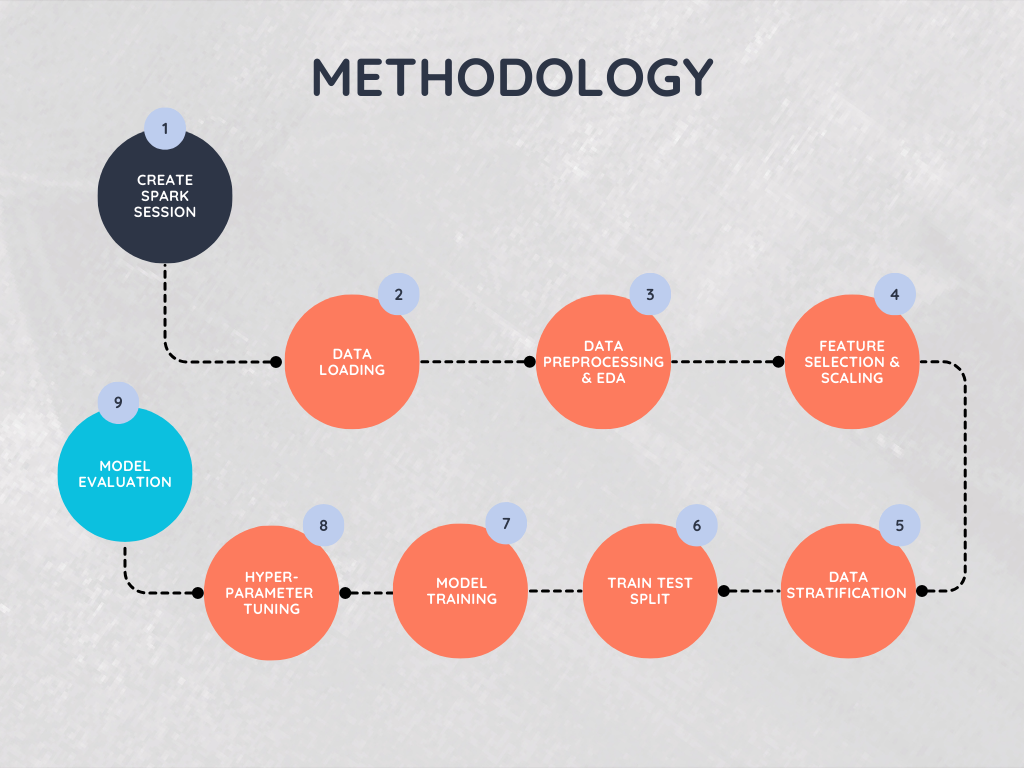
1. **Feature Scaling:**  
   Scaled the selected features into a [0, 1] range using MinMaxScaler in preparation for training the model.
2. **Stratified Split:**  
   Split the data for training and validation, maintaining the class balance of the target variable.

**Final features used:** 'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'NumCompaniesWorked', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Department\_index', 'OverTime\_index'

# Tables Hardware/Software/ Technique Used

|  |  |
| --- | --- |
| **Criteria** | **Details** |
| Hardware Configuration | Windows 11 Home Single Language 64-bit with 8 GB AMD Ryzen 5 with Radeon Graphics and GeForce GTX 1650 |
| Software Configuration | Google Colab with runtime TPU v2-8 |
| Big Data Tools Used with Version | PySpark Version: 3.5.3 |
| Python Library Used | NumPy, Pandas, sklearn, pyspark.ml, pyspark.sql |
| Visualization Tool | Matplotlib, Seaborn |
| Any other tools, libraries used | None |

# IMPLEMENTATION & METHODOLOGY



First the spark session was created and then the implementation started.

1. **Setting Up a Spark Meeting**  
   To enable interaction with the PySpark API, SparkSession.builder is used to start a PySpark session.  
   The master node setting ("local[\*]") and application name ("Employee Attrition Prediction") are set up for the session to make use of all CPU cores.

The session is the starting point for loading and processing the dataset.

1. **Data loading and Preparation**  
   Load dataset - Use spark.read.csv() with header=True and appropriate types to load the HR data set.  
   First Exploration: Print the schema using .printSchema() and view the first few rows to understand the dataset.  
   Descriptive analysis was done to know about the features.  
   Dropped unnecessary columns : Drop columns like EmployeeCount, EmployeeNumber, StandardHours, and Over18 since they predict nothing.

Checked for missing values. There were no missing values in the dataset.

1. **Exploratory Data Analysis (EDA)**

EDA includes statistical summaries and visualizations to understand the nature of data.

Used .describe() to calculate mean, std, min and max for numerical features.  
Feature correlations analyzed using .corr() to find out relations.

Histograms: Displayed distributions for features like Age and MonthlyIncome.

Box Plots: Highlighted outliers in DistanceFromHome and YearsAtCompany.

Correlation Heatmap: Visualized relationships between features.

Count Plots: Explored categorical features like JobRole and Department.

1. **Feature Engineering**  
   Categorical Encoding: Apply StringIndexer to convert the string-based categorical features into numerical indices, for example, Gender, JobRole.  
     
   Used OneHotEncoder to transform indexed categories to sparse vectors for more meaningful representation in machine learning algorithms.  
     
   Create meaningful features such as Tenure-the difference between YearsAtCompany and YearsInCurrentRole-and IncomeToJobLevelRatio, which give normalized insights.

Correlation Analysis:  
Use VectorAssembler to combine all features and calculate correlation with target (Attrition\_index).  
Have dropped features with low correlation(<0.03)  resulting in dimensionality reduction and improvement on model performance.

Feature Scaling:

Normalize numerical features to a [0, 1] range using MinMaxScaler to ensure uniform scaling for model training.

1. **Data Splitting**

Split the dataset into **training** (80%) and **testing** (20%) setsto ensure reliable model evaluation. Stratified the split to maintain the distribution of the target variable (Attrition\_index).

1. **Model Building**

Trained models like logistic regression, decision tree, random forest and gbt classifier using PySpark's MLlib library to predict employee attrition.

Also checked which featuresImportance of the models.

1. **Model Evaluation:**  
   The performance of models is checked by metrics such as accuracy, weighted precision, weighted recall, f1 score, and ROC-AUC score as there was class imbalance in the data.
2. **Hyperparameter tuning:**

Hyperparameter tuning in the code is done using CrossValidator in PySpark. It evaluates multiple hyperparameter combinations through k-fold cross-validation, selecting the best model based on performance metrics like accuracy. The CrossValidator uses an estimator (model), an evaluator (e.g., BinaryClassificationEvaluator), and a ParameterGridBuilder to define the search space for hyperparameters. It helps optimize the model's hyperparameters for better performance.

1. **Visualization of Results:**

Plotted confusion metrics and ROC-AUC curves and comparison between the different models.

# Comparative Analysis of PYSPARK ML Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Name** | **Accuracy** | **Weighted Precision** | **Weighted Recall** | **F1-Score** | **AUC** |
| Logistic Regression | 89.79 | 88.89 | 89.79 | 88.35 | 0.79 |
| Decision Tree | 83.26 | 80.04 | 83.26 | 81.27 | 0.35 |
| Random Forest | 86.12 | 82.85 | 86.12 | 81.38 | 0.77 |
| Gradient Boosting Classifier | 84.89 | 81.48 | 84.89 | 82.41 | 0.74 |

# Comparative Analysis of PYSPARK ML Models After Hyperparameter Tuning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Name** | **Accuracy** | **Weighted Precision** | **Weighted Recall** | **F1-Score** | **AUC** |
| Logistic Regression | 88.16 | 86.86 | 88.16 | 85.57 | 0.79 |
| Decision Tree | 83.67 | 81.68 | 83.67 | 82.50 | 0.35 |
| Random Forest | 87.75 | 86.21 | 87.75 | 84.89 | 0.80 |

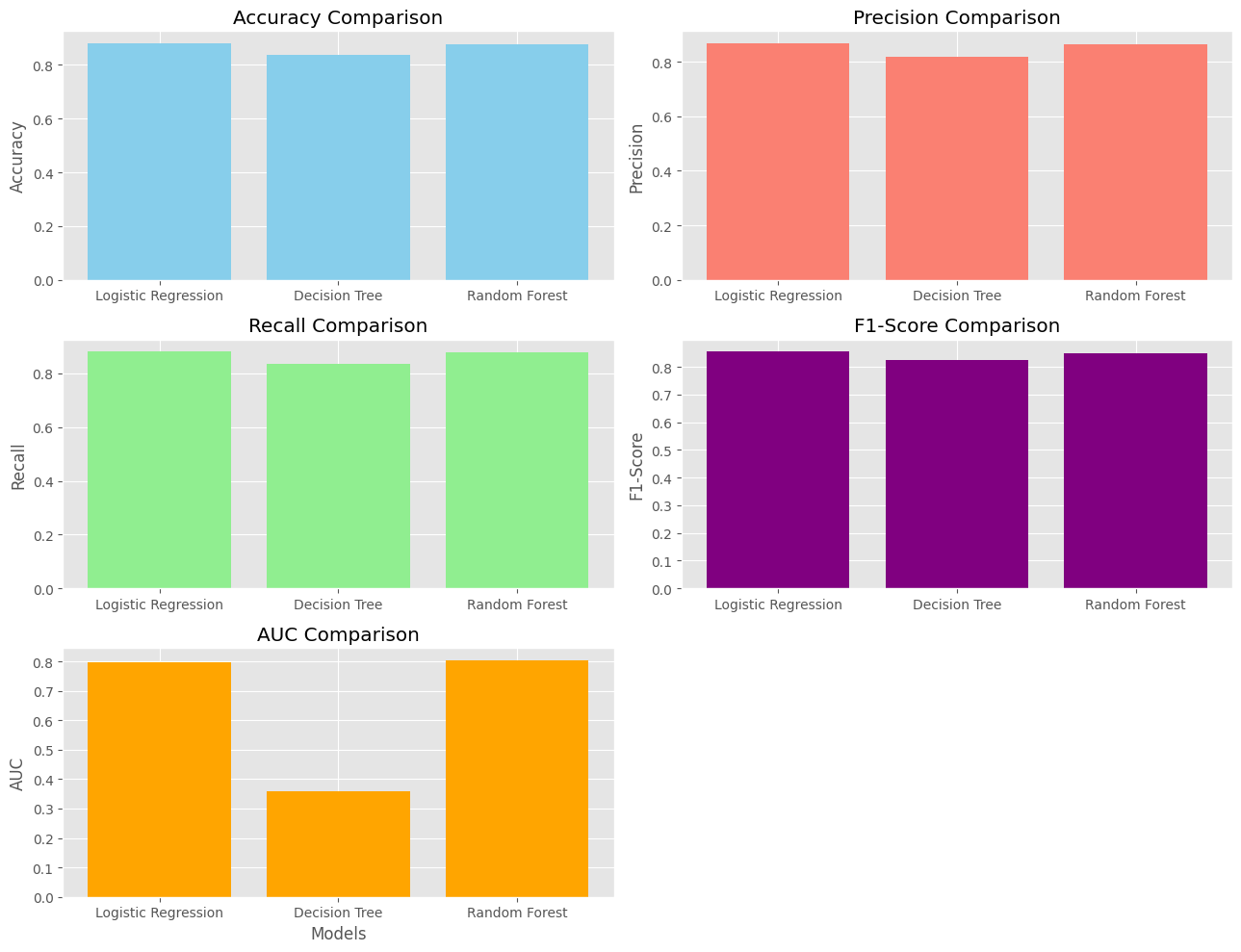
# ANALYTICS FINDINGS

Before Hyperparamter Tuning:

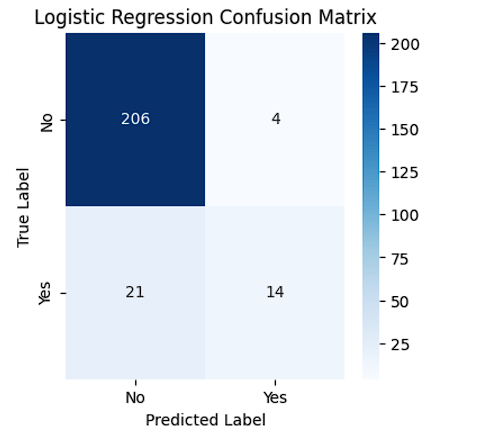
A group of colorful bars

Description automatically generated with medium confidence

After Hyperparameric Tuning:



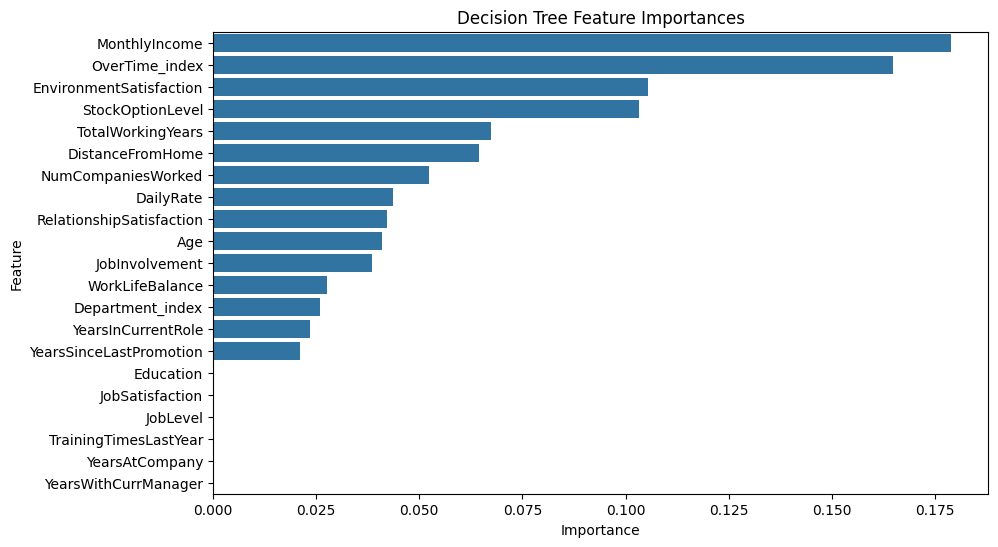
Interesting patterns of performance of various machine learning models inpredicting employee attrition before and after hyperparameter tuning are observed. Logistic Regression is a baseline model that maintains high accuracy and weighted recall in terms of identifying "Stay" and "Leave" classes. Even after hyperparameter tuning, its accuracy is stable with a mild drop from 89.79% to 88.16% but yields better F1-Score and weighted precision, that is, better overall balance regarding class predictions. Its AUC score of 0.79 also indicates great discrimination between the two classes.



The Decision Tree and the Random Forest models demonstrate the differences that hyperparameter tuning provides in performance. The metrics for the Decision Tree, F1-Score and weighted precision, are marginally improved after tuning, while its AUC score remains low at 0.35, indicating that it cannot cope with complex relationships. The Random Forest model resulted in great gain after the tuning with an AUC value of 0.80 and better weighted precision and recall. This proves that a properly tuned Random Forest is best for handling complex feature interaction.

Lastly, the Gradient Boosting Classifier, having a mediocre performance at the start, was good enough at improving its F1-Score post-tuning for imbalanced predictions. It still had a relatively poorer AUC of 0.74 compared to that of Random Forest. Conclusion: Logistic Regression and Random Forest are the strongest models of this experiment. The model of Logistic Regression has better simplicity and interpretability as compared to the other models of this experiment. The model Random Forest has robust performance, and it is flexible enough to capture fine patterns in the dataset.

The feature importance of the models to help us understand which factors most influence employee attrition in a model:



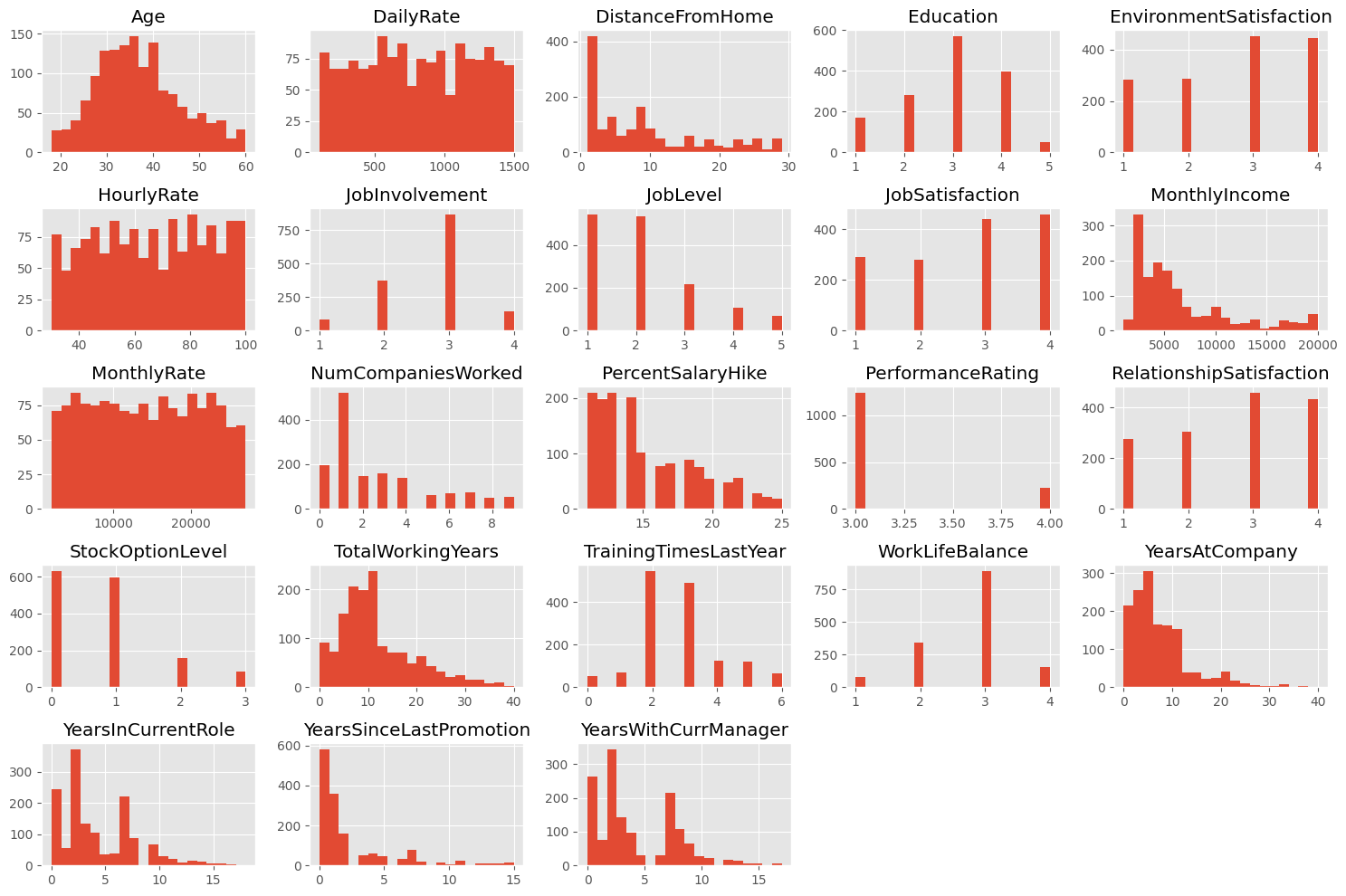
A graph of a forest

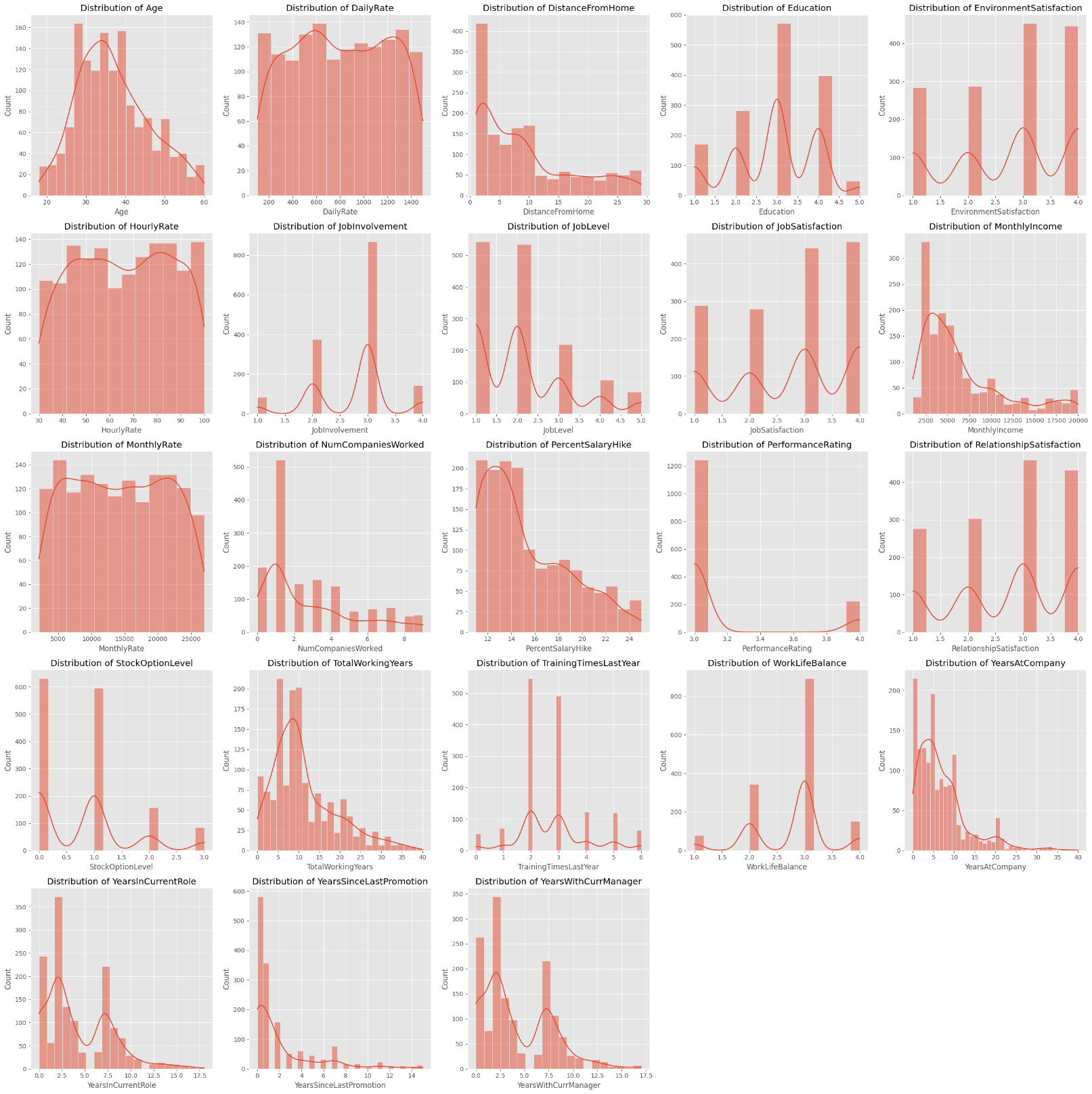
Description automatically generated with medium confidence

A graph with blue and white bars

Description automatically generated

# EDA Insights





The distribution of different numerical columns in the dataset is visually summarized by the histograms. Some important observations:

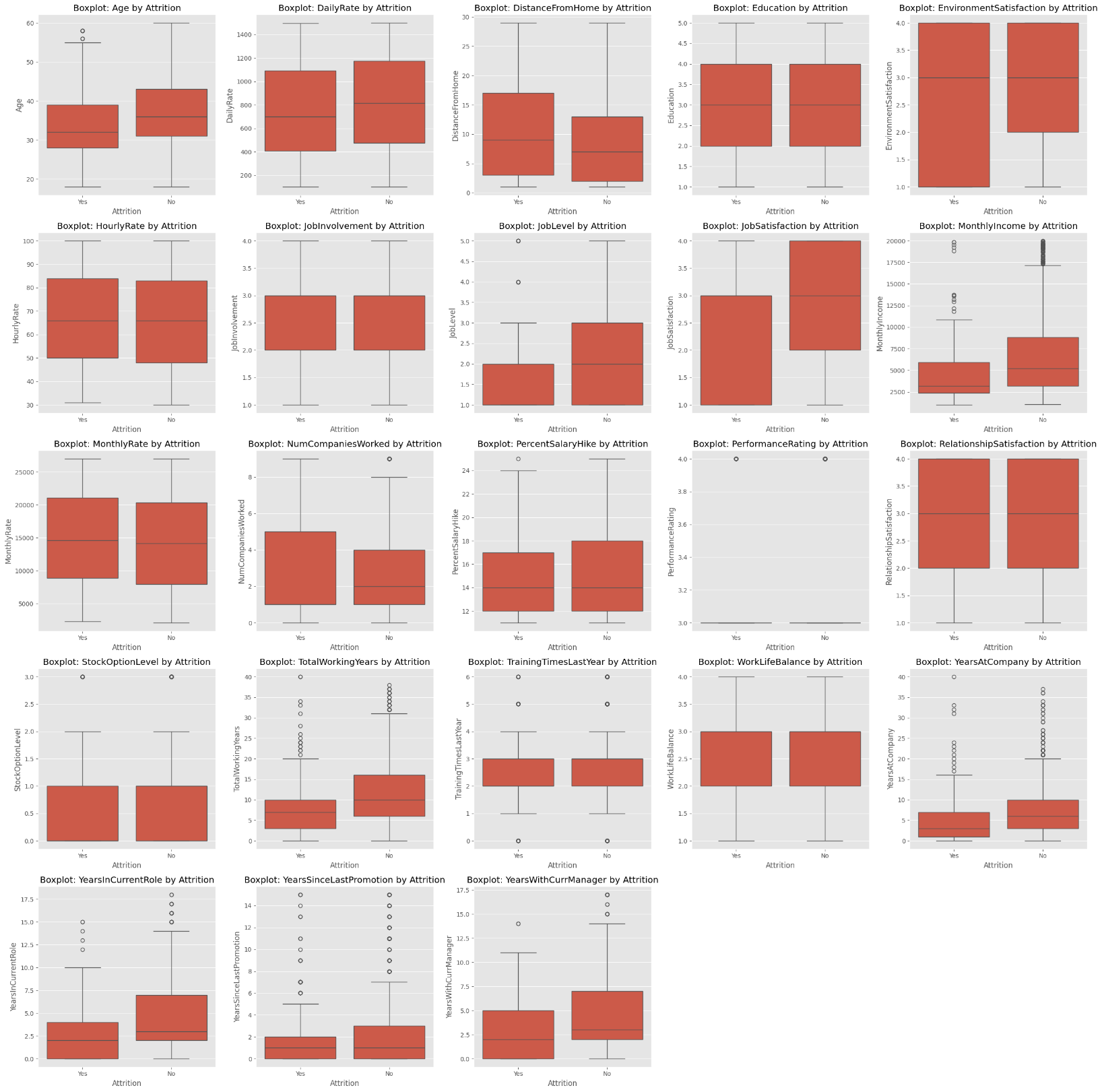
The majority of staff members are in their 30s to 40s.  
The rate is dispersed evenly each day.  
After ten units of distance, there is a noticeable decline in the number of employees who reside far from their place of employment.  
Most workers have a third or fourth-level education.  
The majority of workers give their surroundings a score of three or four.  
There is a uniform distribution of the hourly rate.  
The majority of workers have a job participation score of three or four.

The majority of workers give their jobs a score of three or four.

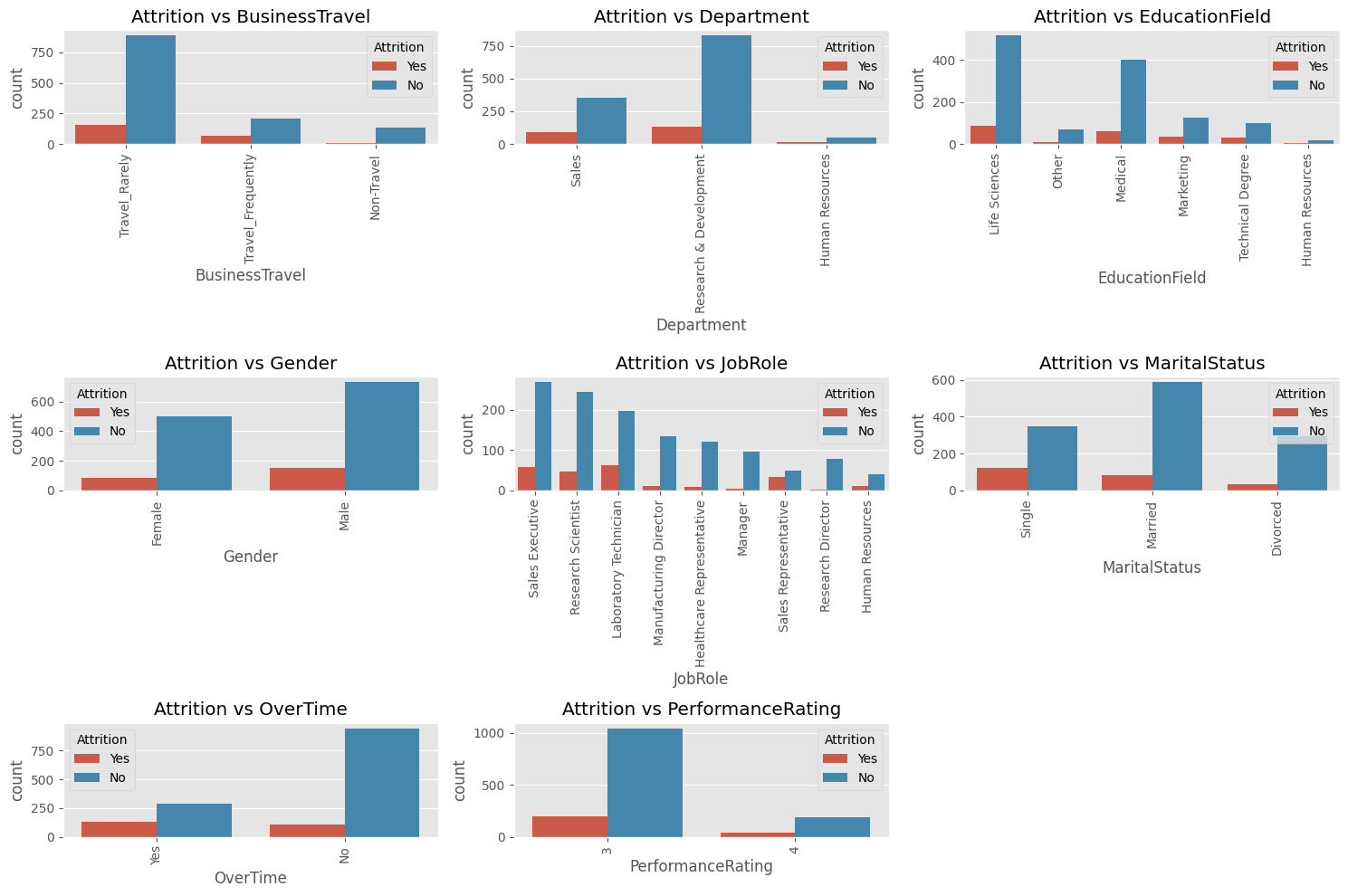
The majority of employees earn between 5,000 and 10,000 units per month, indicating a right-skewed distribution. There is a uniform distribution of the monthly rate.

There is a uniform distribution of the monthly rate.  
The majority of employees have worked for one or two companies.  
The majority of workers saw pay increases of 11% to 15%.  
The majority of workers receive ratings of three or four.  
The majority of workers give their relationships a score of three or four.

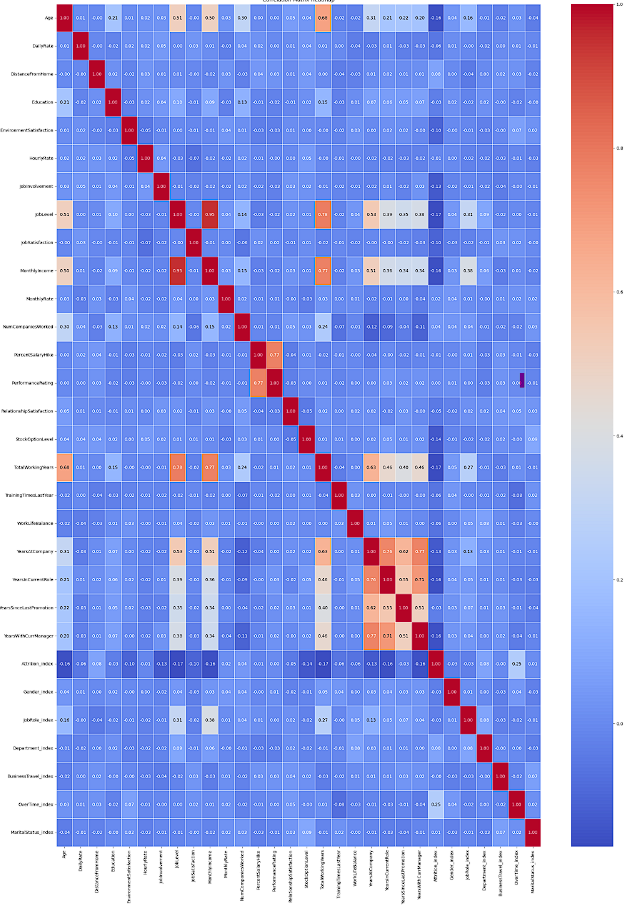
The majority of workers have either 0 or 1 stock options.  
The distribution of total working years is biased to the right, with the majority of employees having worked for five to fifteen years.  
Last year, the majority of staff attended two or three training sessions.  
The majority of workers give their work-life balance a score of three.  
The majority of staff members have worked there for less than ten years.  
The majority of workers have held their current position for fewer than five years.  
The majority of staff members received promotions in the last five years.  
The majority of workers have worked for their present boss for fewer than five years.  
  
Understanding staff demographics, job satisfaction, and work-related metrics can be aided by these insights, which offer a thorough summary of the distribution of different features in your information.



Several important insights regarding employee attrition are shown by the boxplots. Younger workers who live further away from their place of employment are more likely to have departed the company. Additionally, they express less contentment with things like their work and surroundings. Furthermore, departing employees typically hold lower-level positions, make less money, and have previously worked for more organizations. They have worked for the company for fewer years overall, spent less time in their current position, and had a promotion less recently. They have also had their current manager for fewer years. These observations point to a number of potential causes of staff attrition.



Several important facts on employee attrition are revealed by the countplots. Attrition rates are typically greater for workers with medical or technical degrees, those who work in sales, and those who travel regularly for work. Non-attrition is typically higher among male employees. In addition, attrition rates are higher for manufacturing directors and sales executives than for other positions. Employees that work overtime are more likely to quit, and single employees have a greater turnover rate than married or divorced employees. Furthermore, attrition is higher for workers with a performance rating of 4 than for those with a rating of 3. These elements offer insightful information on the causes of employee attrition.



According to the correlation matrix heatmap, employees who have worked for the company longer, are older, and make more money are less likely to quit. Attrition is also lower among workers who have worked for their current manager for a longer period of time. On the other hand, workers who put in extra hours are more likely to quit. Key elements that affect employee retention are highlighted by the correlation between lower attrition and higher job satisfaction, environment satisfaction, and work-life balance.

# Conclusion

The comparative study of machine learning models for predicting attrition of employees shows that Logistic Regression and Random Forest are robust solutions and have their own strengths, one is better than the others in certain aspects. As a matter of fact, this model always provides high accuracy and interpretability, which often makes it a good selection for applications requiring straightforward information. Random Forest, after hyperparameter tuning, shows the best performance with the best AUC and balanced metrics. Though Decision Tree and Gradient Boosting models show a bit of improvement, they are lagging in reliability and separability. Thus, it becomes a must to choose models based on the needs of the organization, and whereas Random Forest is suitable for this nuanced analysis, Logistic Regression would be suitable for ease of implementation.

Also, we can conclude that to reduce employee attrition, HR should focus on increasing monthly income, managing overtime effectively, and enhancing the overall work environment. By addressing these key areas, HR can improve employee satisfaction and retention.

[COLAB NOTEBOOK LINK](https://colab.research.google.com/drive/10r4utjADSwBsMQXS3a-q0wKyLl_mFmXH?usp=sharing)