Employee Attrition Analysis

Machine Learning Graded Microproject

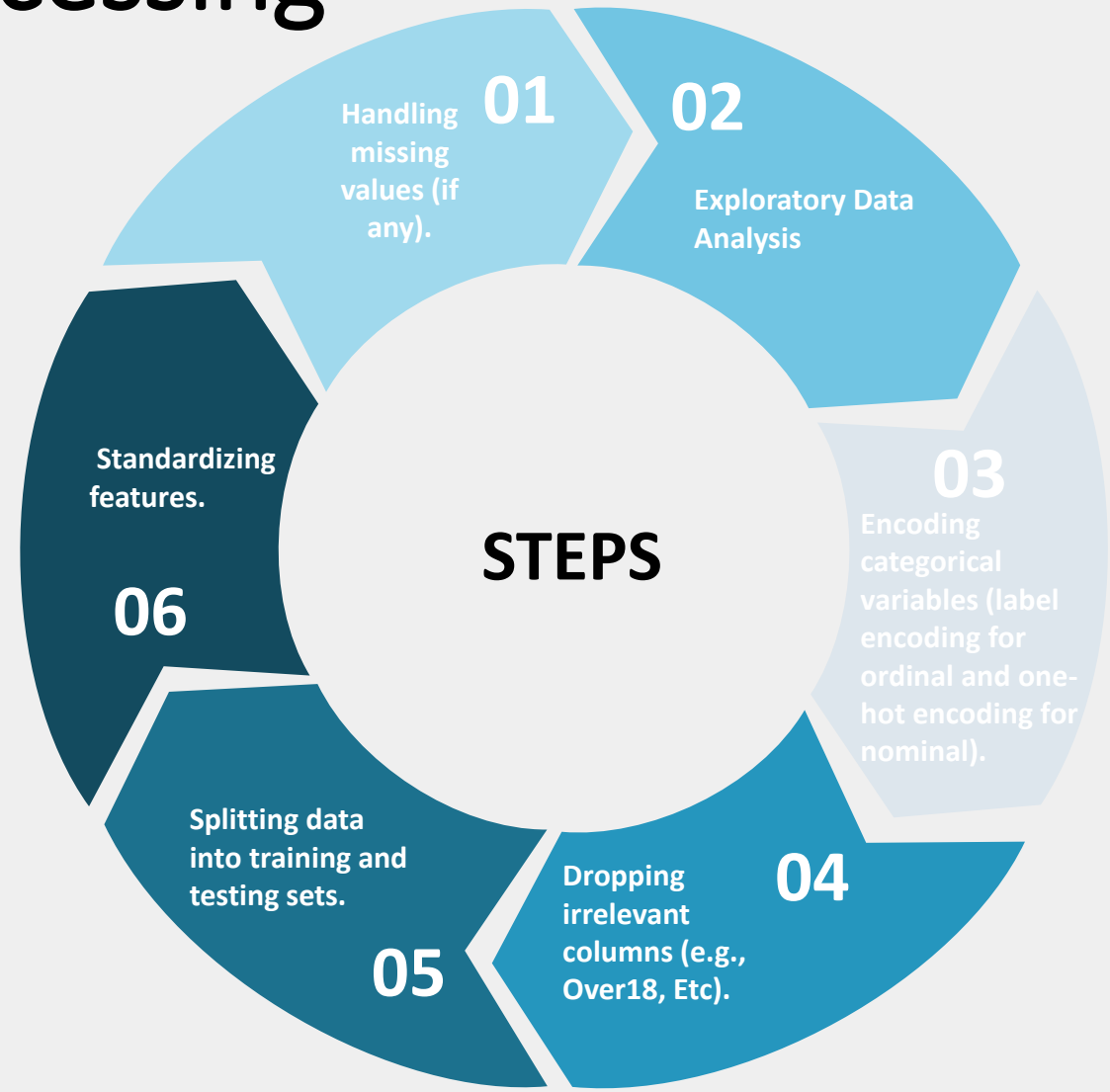
Uparna Das

25/07/2024

Project Overview

- The objective of this project is to uncover insights about possible employee attrition and build predictive models.
- The dataset consists of various attributes related to employees, including their satisfaction levels, involvement, and other factors that may contribute to attrition.

Data Preprocessing

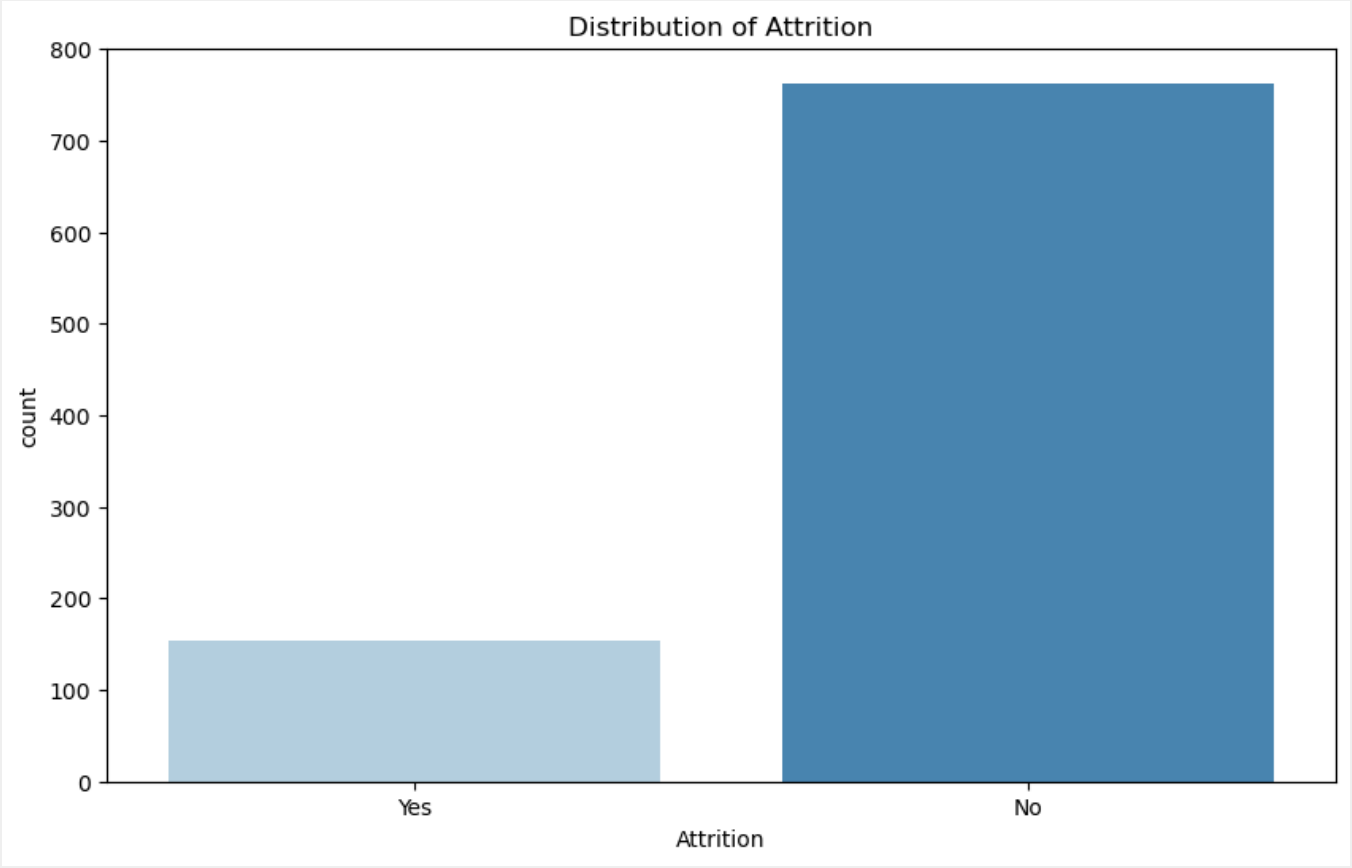


Before beginning to build the models, the following steps have been taken –

Exploratory Data Analysis (EDA)

Here are some descriptive statistics for the dataset:

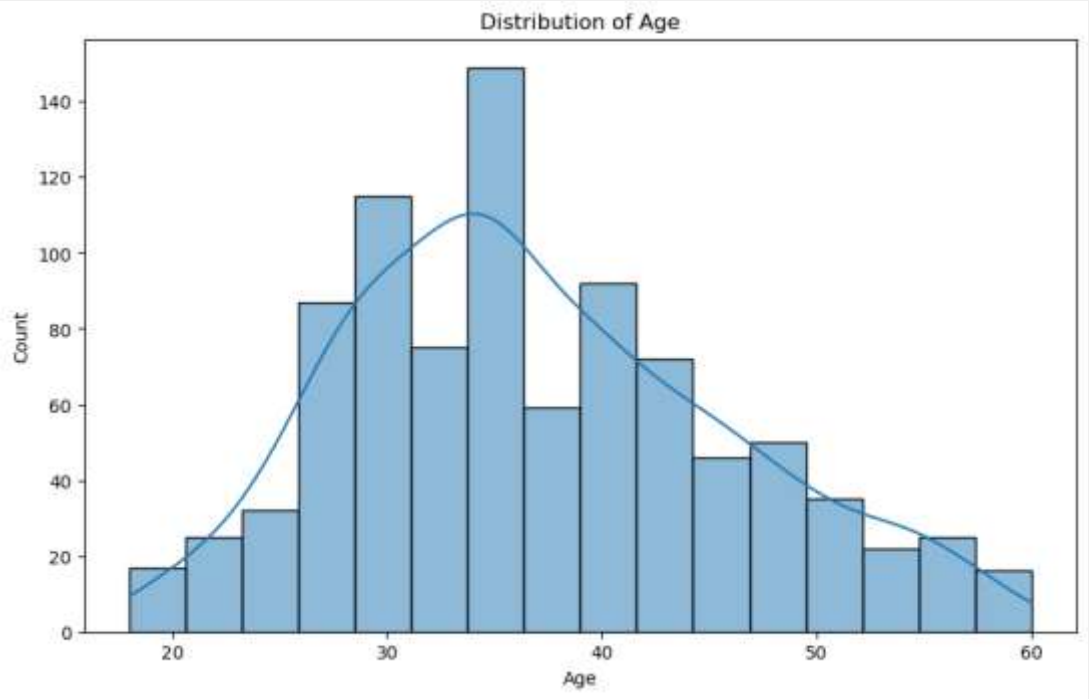
- **Age**: The average age is around 37 years.
- **Daily Rate**: The average daily rate is approximately 785.
- **Distance From Home**: On average, employees live around 9.3 units of distance from their workplace.
- **Education**: Most employees have an education level of 'College' or 'Bachelor'.
- **Monthly Income**: The distribution shows some outliers, indicating a few employees have significantly higher income.

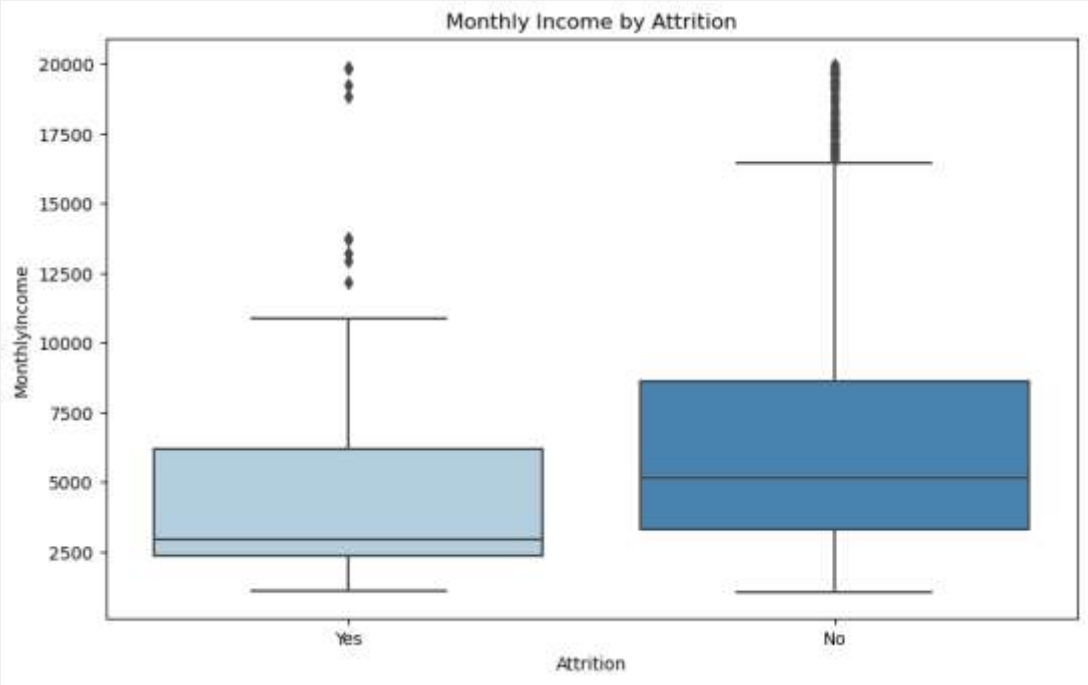


Distribution of Attrition

The target variable Attrition is imbalanced:

- **Attrition = No**: 80%
- **Attrition = Yes**: 20%



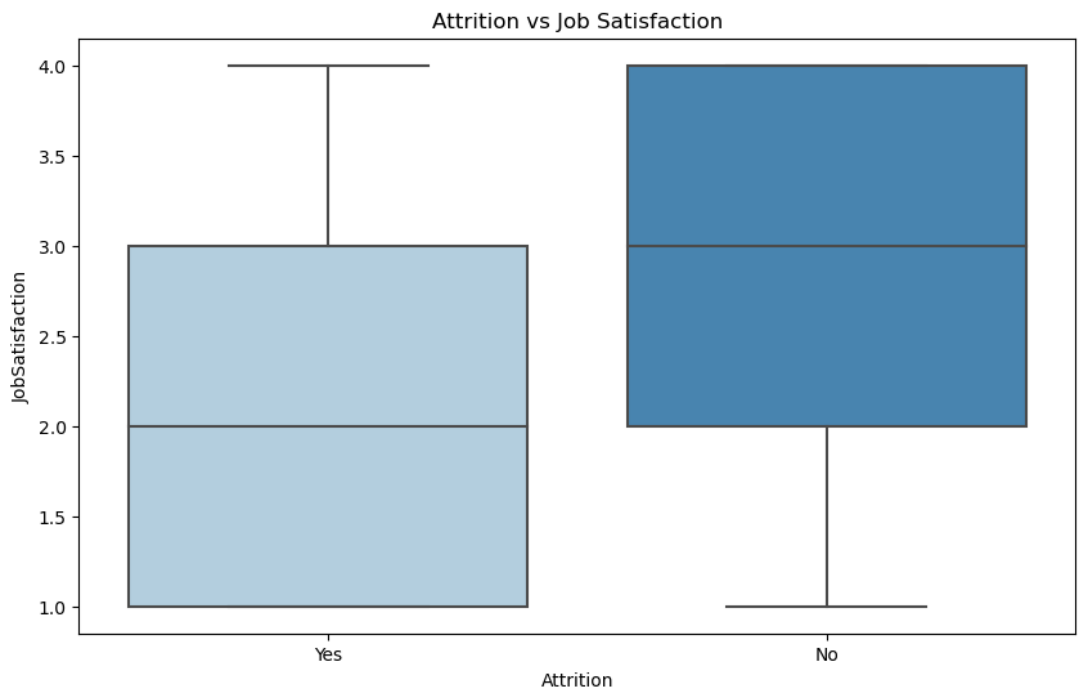


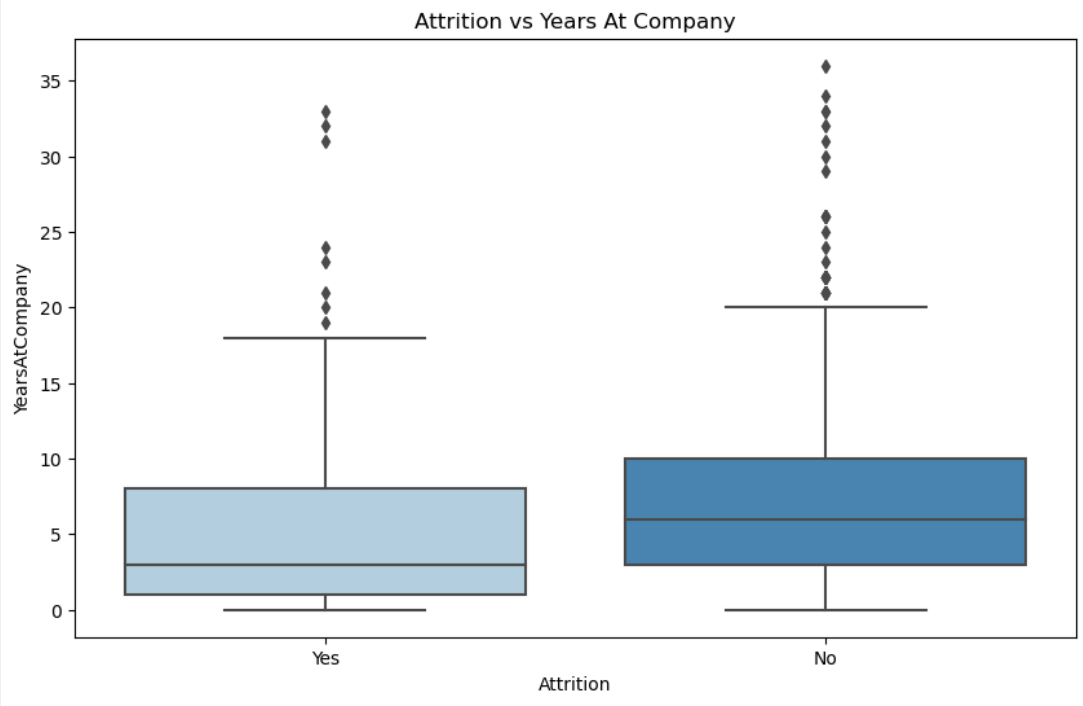
Distribution of Age

The age distribution is approximately normal, with a peak around 35-40 years old.

Monthly Income by Attrition

Employees who did not leave the company tend to have higher median monthly incomes compared to those who left.



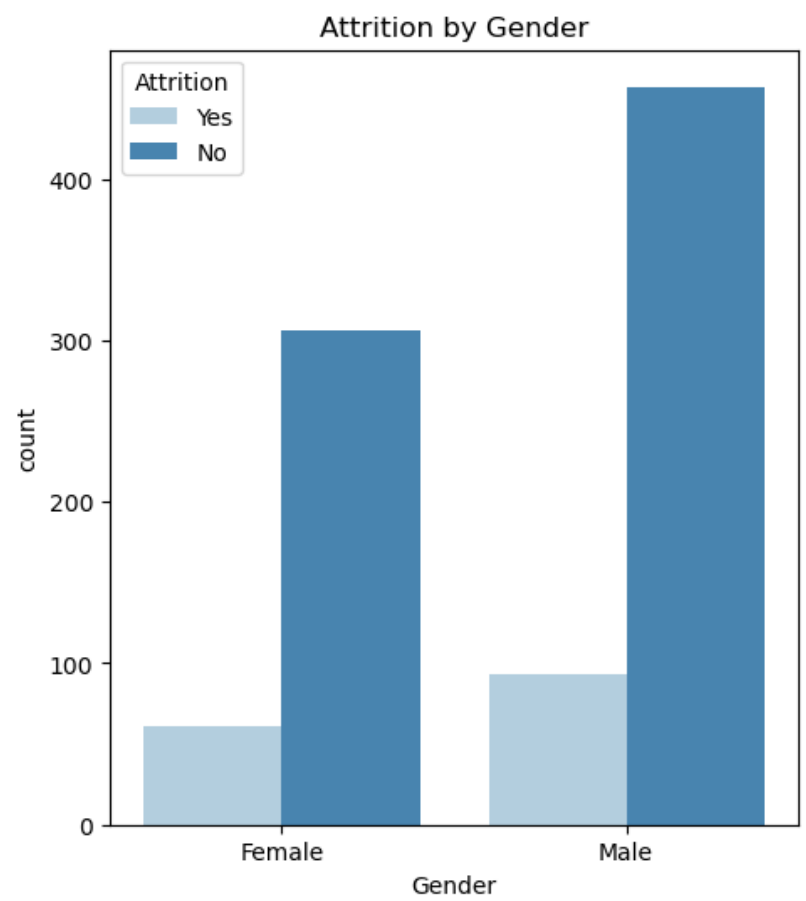


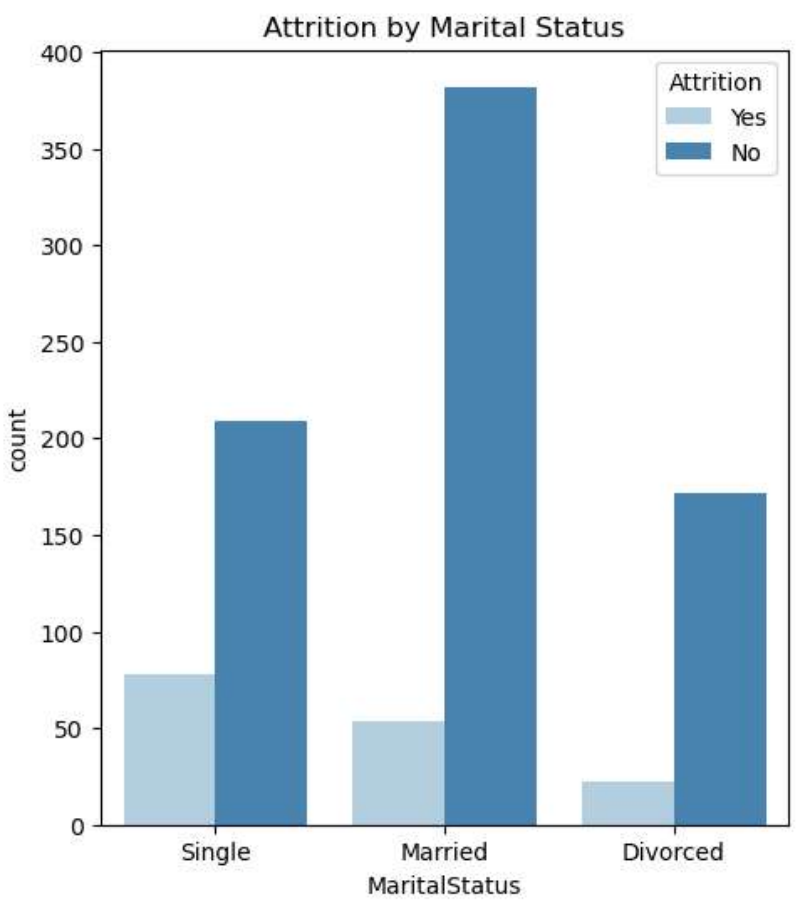
Attrition vs Job Satisfaction

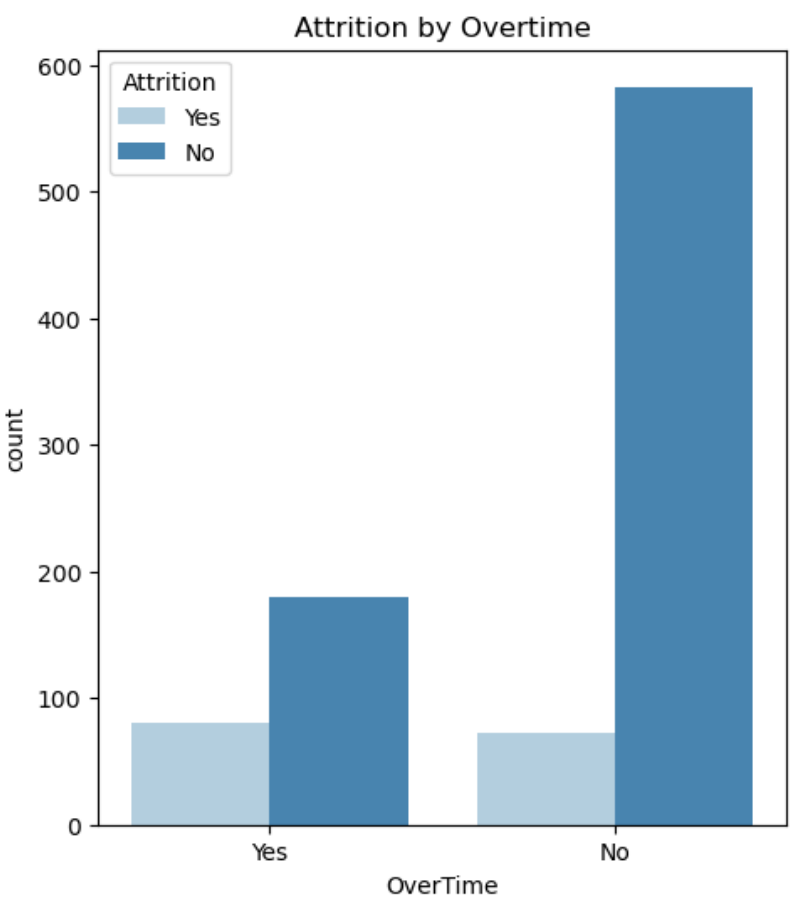
Higher job satisfaction correlates with lower attrition rates.

Attrition vs Years at Company

Employees with longer tenure are less likely to leave.



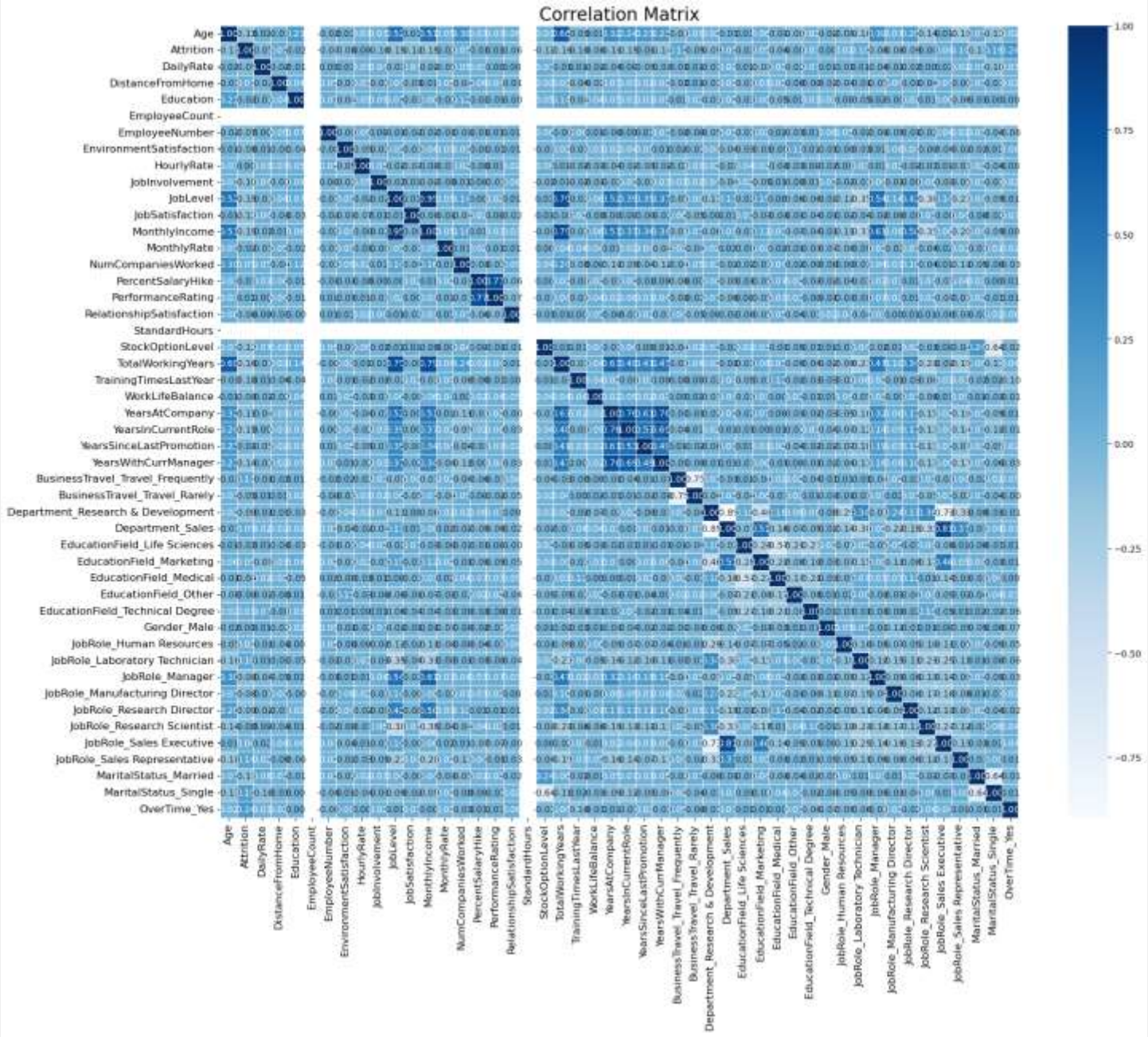




Attrition by Gender
Both males and females have similar attrition patterns.

Attrition by Marital Status
Single employees have higher attrition rates compared to married employees.

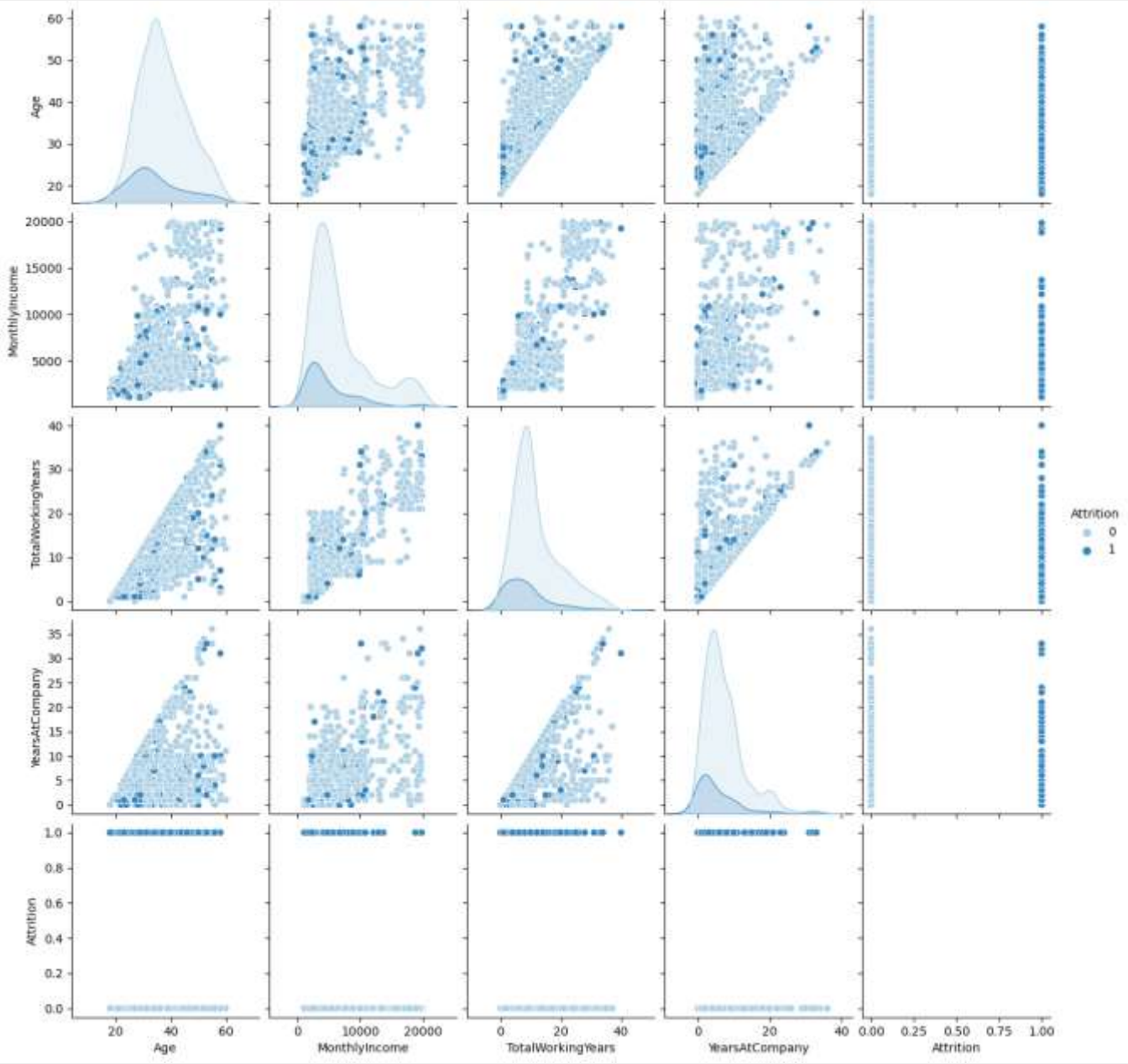
Attrition by Overtime
Employees who work overtime are more likely to leave.



Correlation Matrix

The correlation matrix shows the relationships between different numerical features. Some key observations:

- Monthly Income is moderately negatively correlated with Attrition (-0.17), suggesting that employees with higher incomes are less likely to leave
- Age is moderately negatively correlated with Attrition (-0.16), indicating that younger employees are more likely to leave.



Pair Plot

1) Age : Younger employees (below 30) and older employees (above 45) seem to have higher attrition rates, indicated by the presence of more 'Attrition=1' points.
2) Monthly Income : There is a clear distinction in attrition rates based on income, with lower-income employees showing higher attrition.
3) Total Working Years : As total working years increase, attrition rates decrease, indicated by fewer 'Attrition=1' points.
4) Years At Company : There is a significant drop in attrition for employees who have been with the company for longer periods (10+ years).

Key Insights from EDA

- **Important Correlations**:

- **Total Working Years** is moderately negatively correlated with Attrition (-0.17), indicating that employees with more working years are less likely to leave.
- **Over Time** is positively correlated with Attrition (0.14), suggesting that employees who work overtime are more likely to leave.
- The age distribution is approximately normal, with a peak around 35-40 years old.
- There is a noticeable gap in the income levels, indicating a possible link between salary and employee retention.
- Employees with higher **Monthly Income**, **Total Working Years**, and **Years At Company** tend to have lower attrition rates.

Model Building

- For building the models, I have used the following methods –

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbours
- Gradient Boosting Machine

- Each model is trained on the training data using the fit method. The model learns the relationship between the input features and the target variable (Attrition) by adjusting its parameters.

Model Performance Summary

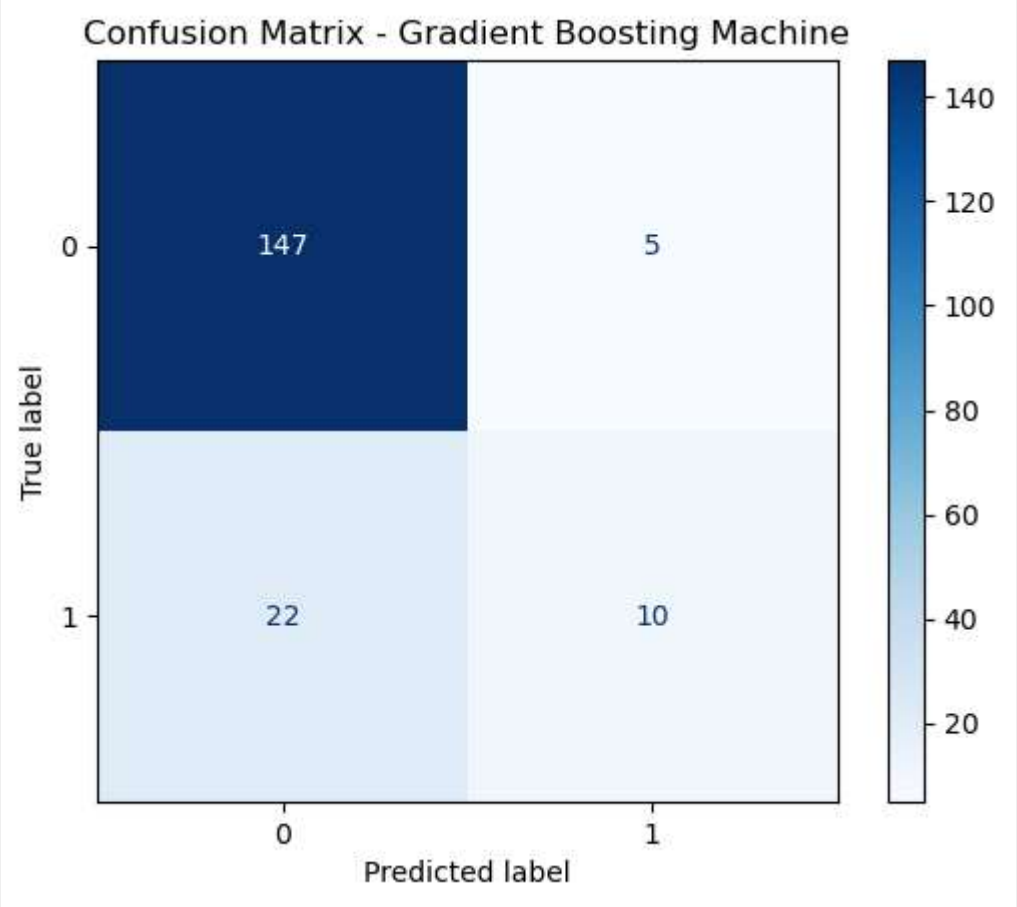
Model	Accuracy	Precision (Attrition=1)	Recall (Attrition=1)	F1-Score (Attrition=1)
Logistic Regression	0.8804	0.78	0.44	0.56
Decision Tree	0.7826	0.38	0.41	0.39
Random Forest	0.8261	0.50	0.09	0.16
K-Nearest Neighbors	0.8315	0.60	0.09	0.16
Gradient Boosting Machine	0.8478	0.62	0.31	0.42

Model Evaluation

- The performance of each model is evaluated using several metrics, including:
- **Accuracy**: The percentage of correctly predicted instances out of the total instances.
- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall**: The proportion of actual positives that were correctly identified.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.
- **Confusion Matrix**: A matrix showing the counts of true positives, true negatives, false positives, and false negatives.

Best Performing Model

- **Gradient Boosting Machine**
  - **Accuracy**: 85%
  - **Precision (Attrition=1)**: 0.67
  - **Recall (Attrition=1)**: 0.31
  - **F1-Score (Attrition=1)**: 0.43
- The model performs well in predicting employees who will stay (high precision and recall for Attrition=0) and shows relatively better performance for predicting attrition compared to other models.



Recommendations

- Focus on understanding and addressing the factors contributing to attrition among the minority group.
- Implement targeted retention strategies for younger employees and those with lower incomes. Revisit overtime policies to prevent burnout and reduce attrition.
- Tailor retention strategies to the most common age groups (30-40), and develop specific programs for younger and older employees.
- Review compensation policies to ensure competitive salaries, especially for roles with higher attrition rates. Implement performance-based incentives and salary increments to retain top talent.
- Improve job satisfaction through recognition programs, career development opportunities, and fostering a positive work environment.
- Focus on employee engagement and career development to encourage long-term tenure. Develop targeted programs to support single employees and manage overtime effectively to reduce burnout.

Thank You