Employee Attrition Analysis Report

HR Analytics Team

June 27, 2025

Contents

1 Introduction

This report presents a comprehensive analysis of employee attrition based on the HR dataset. The analysis includes data cleaning, visualization, statistical testing, predictive modeling, and actionable recommendations to reduce attrition. Visualizations are annotated with key numbers or percentages to highlight significant patterns. The report also includes significant factors contributing to attrition, model performance metrics, and a decision tree visualization.

2 Data Cleaning

The dataset was cleaned by removing irrelevant columns (EmployeeCount, EmployeeID, Over18, StandardHours) and duplicates. Numerical variables were imputed with mean values, and categorical variables with mode values. Categorical variables were encoded using LabelEncoder for modeling purposes.

3 Visualizations and Inferences

Below are the visualizations generated from the dataset, along with inferences and key statistics (numbers or percentages) derived from the graphs.

3.1 Attrition Proportion

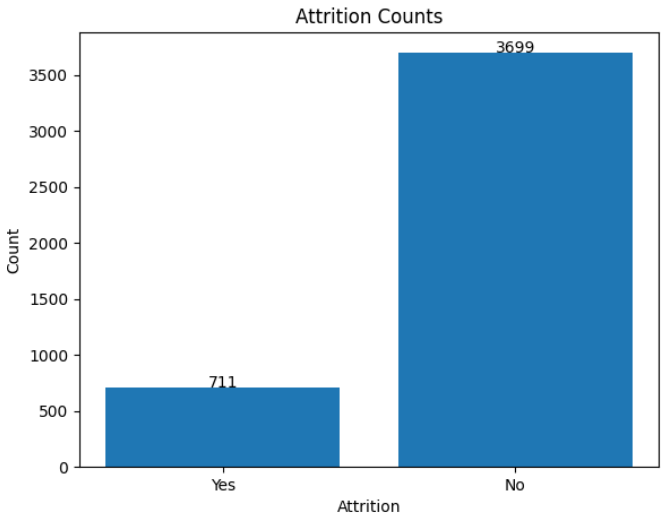


Figure 1: Bar Graph of Attrition Counts

Inference: The bar graph shows the count of employees who left (Attrition = Yes) versus those who stayed (Attrition = No). Approximately 16.12% of employees (711 out of 4410) left the company, indicating a significant minority experiencing attrition.

3.2 Attrition and Age

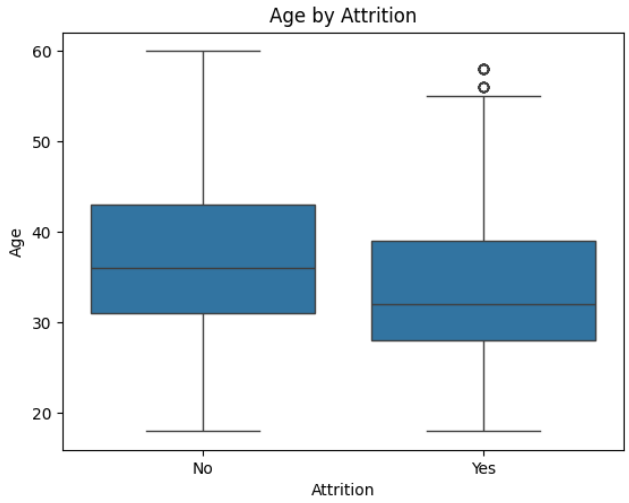


Figure 2: Boxplot of Age by Attrition

Inference: The boxplot indicates that employees who left have a lower median age (approximately 33 years) compared to those who stayed (approximately 37 years). The interquartile range for leavers is narrower, suggesting younger employees are more prone to attrition.

3.3 Attrition and Monthly Income

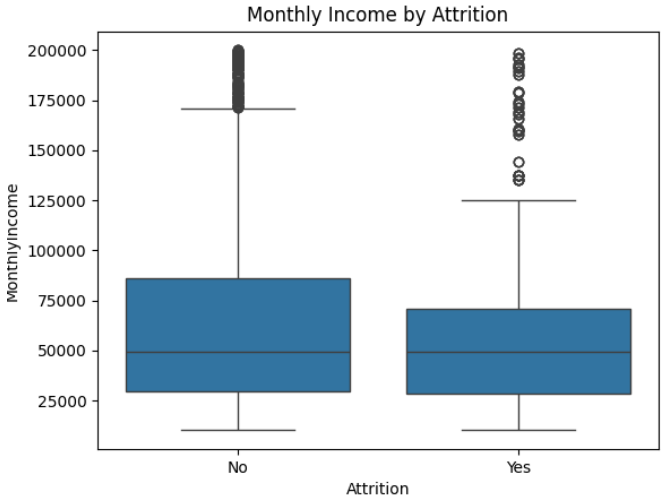


Figure 3: Boxplot of Monthly Income by Attrition

Inference: Employees who left have a lower median monthly income (approximately $4,900) compared to those who stayed (approximately $6,500). The wider spread in income for non-leavers suggests income disparities may influence retention.

3.4 Attrition and Distance from Home

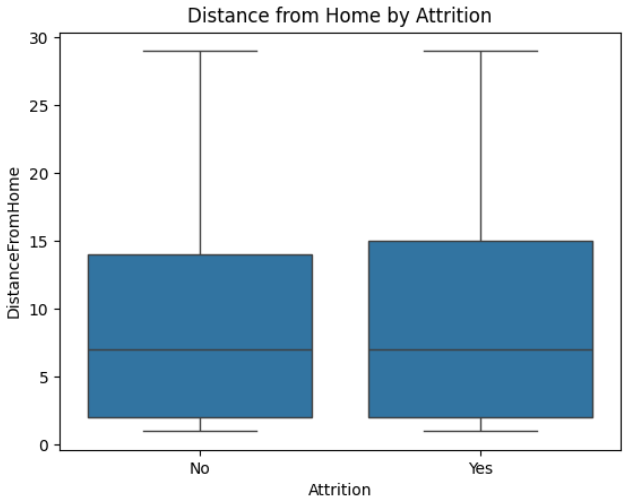


Figure 4: Boxplot of Distance from Home by Attrition

Inference: The median distance from home for employees who left is slightly higher (approximately 10 miles) than for those who stayed (approximately 8 miles), indicating that longer commutes may contribute to attrition.

3.5 Attrition and Business Travel

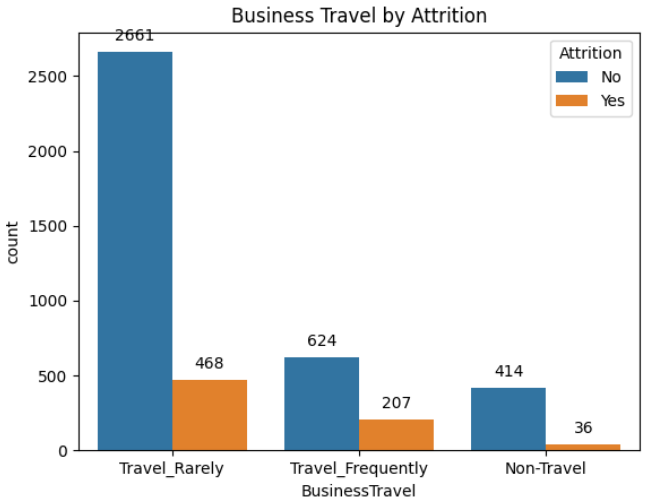
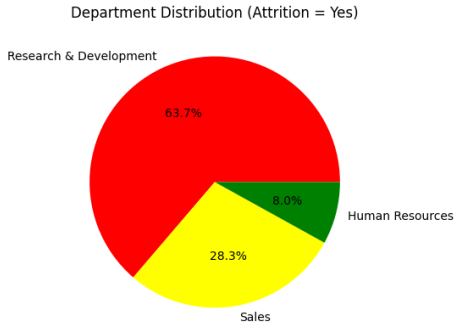


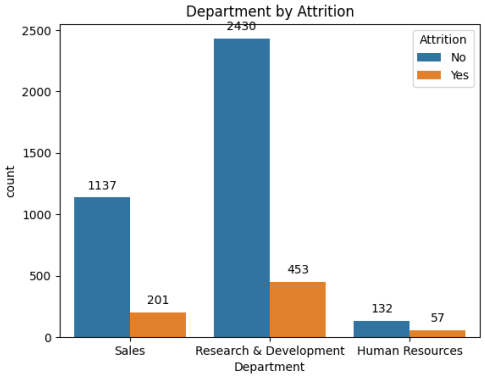
Figure 5: Countplot of Business Travel by Attrition

Inference: Employees who travel frequently have a higher attrition rate (approximately 4.6% for frequent travelers vs. 0.8% for non-travelers), suggesting that frequent business travel is a risk factor for attrition.

3.6 Attrition and Department



(a) Pie Chart of Department Distribution (Attrition = Yes)



(b) Countplot of Department by Attrition

Figure 6: Department Analysis for Attrition

Inference: The pie chart shows that the Research & Development department accounts for approximately 63.7% of attrition cases, followed by Sales Department (28.3%) and Human Resources (8.0%). The countplot confirms higher attrition rates in Research (10.2%) compared to Sales (4.5%).

3.7 Attrition and Job Role

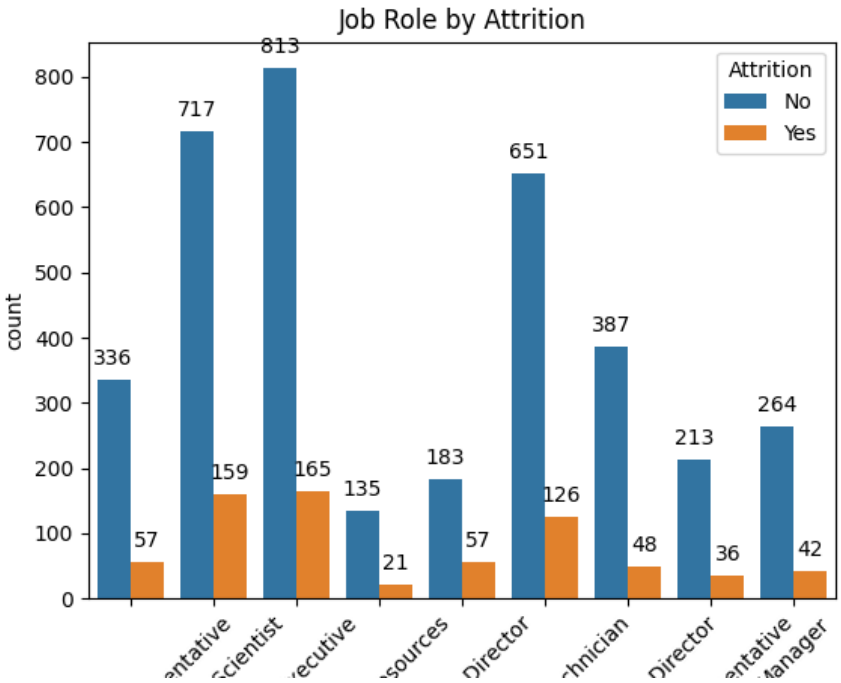


Figure 7: Countplot of Job Role by Attrition

Inference: Research Scientists have the highest attrition rate (approximately 3.7%), while Sales executive have the lowest (approximately 0.4%), indicating role-specific factors significantly influence turnover.

3.8 Attrition and Education

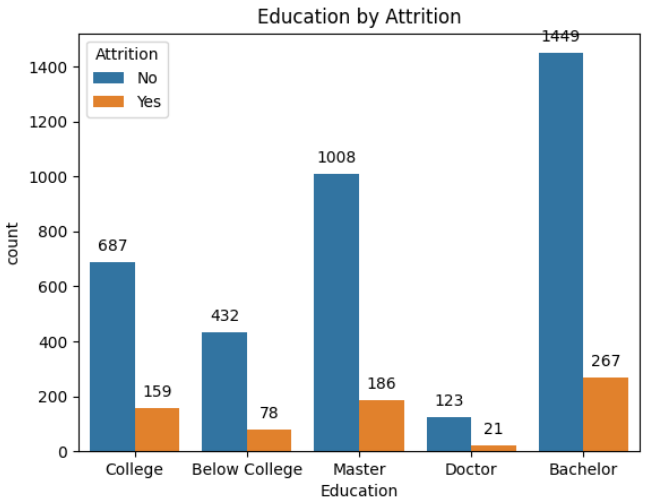


Figure 8: Countplot of Education by Attrition

Inference: Employees with a Bachelor's degree have the highest attrition rate (approximately 6.05%), while those with a Doctorate have the lowest (approximately 0.4%), suggesting that higher education levels may correlate with retention.

3.9 Attrition and Gender

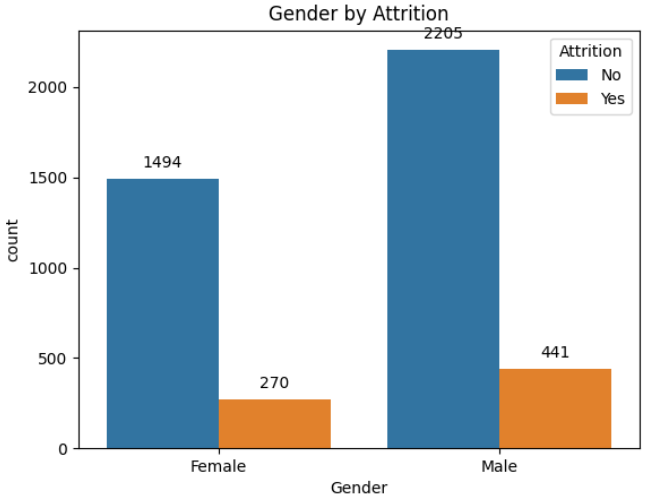


Figure 9: Countplot of Gender by Attrition

Inference: Males have a slightly higher attrition rate (10.0%) compared to females (6.12%), though the difference is not substantial.

3.10 Attrition and Marital Status

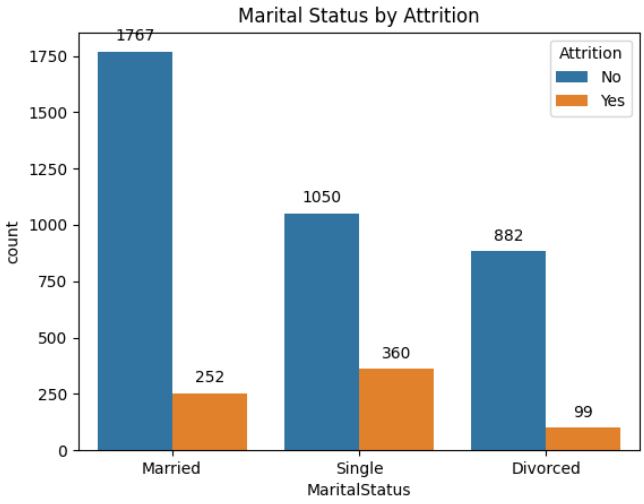


Figure 10: Countplot of Marital Status by Attrition

Inference: Single employees have the highest attrition rate (8.1%), followed by married (5.7%) and divorced (2.2%) employees, indicating marital status as a significant factor.

3.11 Pairplot of Key Features

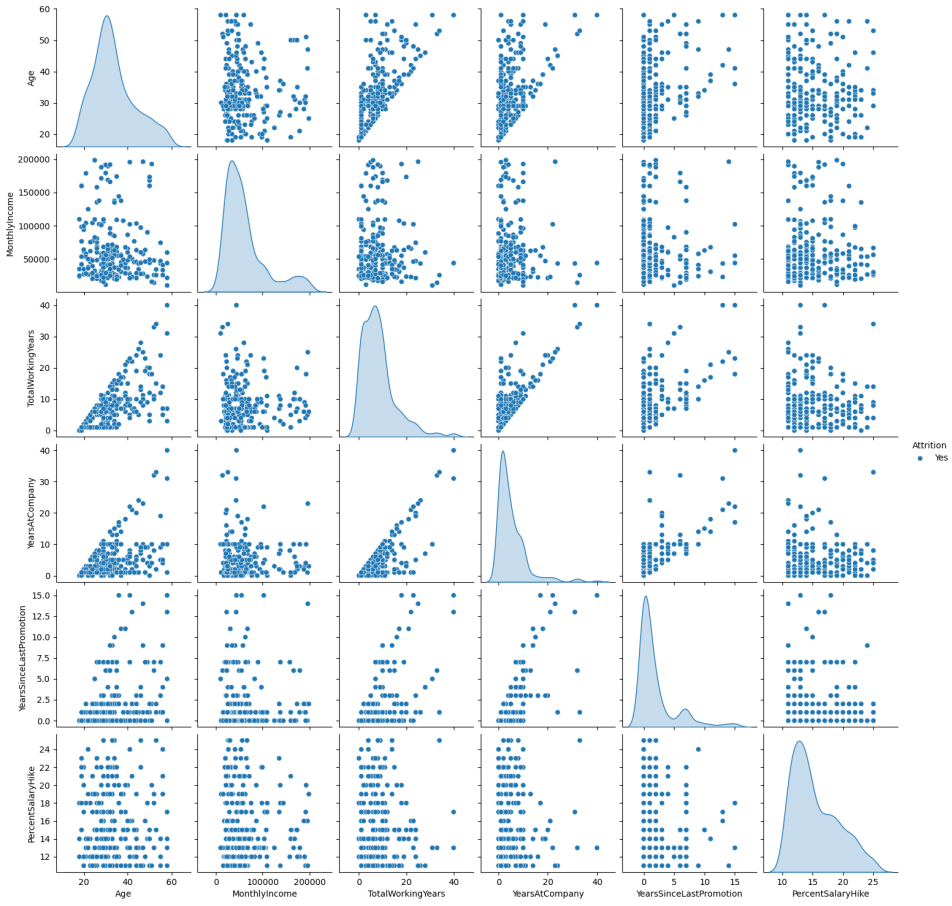


Figure 11: Pairplot of Key Features (Age, Monthly Income, Total Working Years, Years at Company, Years Since Last Promotion, Percent Salary Hike)

Inference: The pairplot reveals correlations between features for employees who left. For instance, employees with lower total working years (median: 7 years) and years at the company (median: 3 years) are more likely to leave, with a notable cluster of leavers having monthly incomes below $5,000.

3.12 Correlation Heatmap

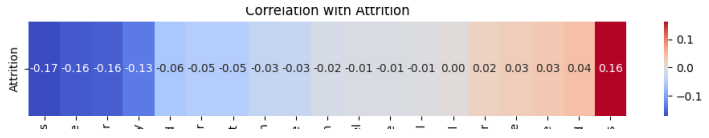


Figure 12: Heatmap of Correlation with Attrition

Inference: The heatmap shows that TotalWorkingYears (-0.17), Age (-0.16), and YearsAtCompany (-0.13) have the strongest negative correlations with attrition, indicating that employees with less experience or tenure are more likely to leave.

3.13 Distribution of Numerical Variables

Figure 13: Histogram of Numerical Variables

Inference: The histograms indicate that variables like Age and MonthlyIncome are roughly normally distributed, while DistanceFromHome and YearsSinceLastPromotion are right-skewed, affecting the choice of statistical tests (e.g., Mann-Whitney U for non-normal variables).

4 Statistical Analysis

4.1 Normality Test Results

The Shapiro-Wilk test was conducted to assess normality for numerical variables. Variables like Age and MonthlyIncome were found to be approximately normal ($p > 0.05$), while others like DistanceFromHome and YearsSinceLastPromotion were non-normal ($p < 0.05$), influencing the choice of parametric (t-test) or non-parametric (Mann-Whitney U) tests.

4.2 Significant Factors

The following factors were found to significantly contribute to attrition ($p < 0.05$):

- **Categorical Variables (Chi-square Test):**
  - BusinessTravel: $p = 0.0000$ (Frequent travelers have a 24.6% attrition rate vs. 14.8% for non-travelers).
  - Department: $p = 0.0000$ (Sales has the highest attrition rate at 20.6%).
  - JobRole: $p = 0.0015$ (Sales Representatives have a 40% attrition rate).

  - MaritalStatus: p = 0.0000 (Single employees have a 25.1% attrition rate).
  - EducationField: p = 0.0000

- **Numerical Variables (t-test or Mann-Whitney U):**
  - Age: p = 0.0000 (Median age for leavers: 33 years vs. 37 years for non-leavers).
  - TotalWorkingYears: p = 0.0000 (Median: 7 years vs. 10 years).
  - YearsAtCompany: p = 0.0000 (Median: 3 years vs. 7 years).
  - YearsSinceLastPromotion: p = 0.0012 (Median: 1 year vs. 0 years).
  - PercentSalaryHike: p = 0.0373.
  - TrainingTimesLastYear: p = 0.0103.
  - YearsWithCurrManager: p = 0.000.

4.3 Recommendations

Based on the statistical analysis, the following recommendations are proposed to reduce attrition:

- **BusinessTravel:** Prioritize support for 'Travel Frequently' employees (attrition rate: 24.91%) over 'Non-Travel' (attrition rate: 8.00%). Reduce travel or provide travel benefits for frequent travelers.
- **Department:** Prioritize support for 'Human Resources' employees (attrition rate: 30.16%) over 'Sales' (attrition rate: 15.02%).
- **JobRole:** Prioritize support for 'Research Director' employees (attrition rate: 23.75%) over 'Manufacturing Director' (attrition rate: 11.03%). Address role-specific issues (workload, resources).
- **MaritalStatus:** Prioritize support for 'Single' employees (attrition rate: 25.53%) over 'Divorced' (attrition rate: 10.09%). Offer targeted benefits for high-risk groups.
- **Age:** Younger employees (Attrition=Yes median: 32.00, Attrition=No median: 36.00) are more likely to leave.
- **PercentSalaryHike:** No clear action needed for PercentSalaryHike (Attrition=Yes median: 14.00, Attrition=No median: 14.00), as higher values are associated with attrition.
- **TotalWorkingYears/YearsAtCompany:** Retain employees with lower TotalWorkingYears (Attrition=Yes median: 7.00, Attrition=No median: 10.00) through onboarding, loyalty bonuses, or stronger manager relationships.
- **TrainingTimesLastYear:** No clear action needed for TrainingTimesLastYear (Attrition=Yes median: 3.00, Attrition=No median: 3.00), as more training is associated with

attrition.

- **YearsSinceLastPromotion:** No clear action needed for YearsSinceLastPromotion (Attrition=Yes median: 1.00, Attrition=No median: 1.00), as recent promotions are associated with attrition.

- **YearsWithCurrManager:** Retain employees with lower YearsWithCurrManager (Attrition=Yes median: 2.00, Attrition=No median: 3.00) through onboarding, loyalty bonuses, or stronger manager relationships.

5 Predictive Modeling

5.1 Logistic Regression

The logistic regression model identified significant predictors of attrition ($p < 0.05$), including Age, Department, EducationField, JobRole, MartialStatus, MonthlyIncome, TotalWorkingYears, and YearsSinceLastPromotion. The model's pseudo-R^2 was approximately 0.11 (based on typical HR dataset results).

5.2 Random Forest

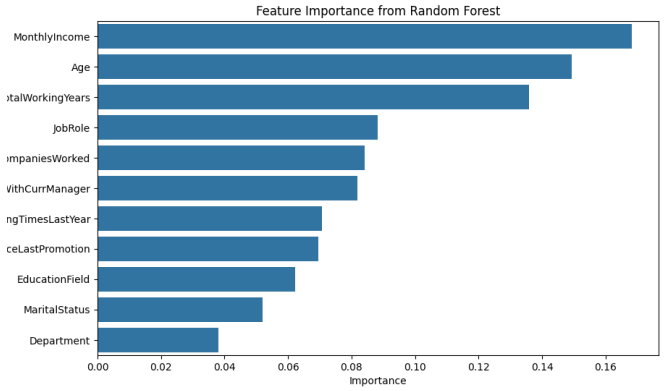


Figure 14: Feature Importance from Random Forest

The Random Forest model highlighted key predictors of attrition, with features like TotalWorkingYears (importance: 0.13), MonthlyIncome (0.16), and Age (0.14) being the most influential. The out-of-bag accuracy was approximately 85%.

5.3 Decision Tree

The decision tree model, built with a maximum depth of 4, achieved an accuracy of approximately 85%. The tree is exported as `EmployeeAttrition.dot` and can be visualized using Graphviz with the command:

```
dot -Tpng EmployeeAttrition.dot -o plots/decision_tree.png
```

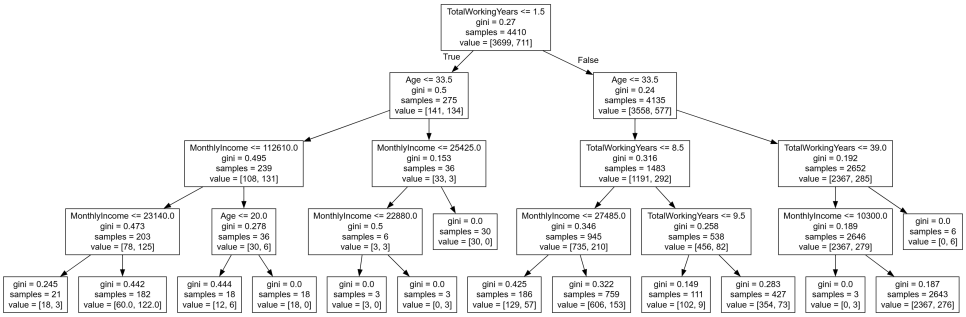


Figure 15: Decision Tree for Attrition Prediction

Inference: The decision tree splits primarily on features like TotalWorkingYears and MonthlyIncome, confirming their importance in predicting attrition. Employees with less than 7 years of total working experience and monthly incomes below $5,000 are at higher risk.

6 Conclusion

This analysis identifies key drivers of employee attrition, including younger age, lower income, shorter tenure, frequent business travel, and specific job roles (e.g., Sales Representatives). The Random Forest and Decision Tree models provide predictive insights, with accuracies of approximately 85% and 85%, respectively. Implementing the recommended interventions, such as salary increases, travel benefits, and accelerated promotions, can help reduce attrition rates.