UNIFIED MENTOR PVT LTD
Haryana, India, 125033

INTERNSHIP PROJECT REPORT ON

IBM HR Analytics Employee Attrition & Performance

UNDER THE GUIDANCE OF
UNIFIED MENTOR PVT LTD
Haryana, India, 125033

Summer Internship Program:

Data Analytics
(Duration: [15th April 2025 & 15th July 2025])

ID: UMID05042528011
[**UNIFIED MENTOR PVT LTD**]

SUBMITTED BY
SURYAVARDHAN MAGATHALA
CSB22072

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
TEZPUR UNIVERSITY
TEZPUR, ASSAM, 784028

Table of Contents

1. Introduction

Employee attrition, also known as employee turnover, refers to the gradual loss of employees over time due to resignation, termination, or retirement. In many organizations, particularly large enterprises, attrition presents a serious challenge. It affects workforce stability, increases recruitment and training costs, and disrupts team performance. Therefore, understanding the key drivers of attrition is essential for improving employee retention and organizational efficiency.

This report focuses on analyzing employee attrition using the **IBM HR Analytics Employee Attrition & Performance** dataset. The dataset contains detailed records of employees, including demographic data, job roles, compensation, satisfaction levels, commute distance, and work-related behaviors such as overtime and business travel.

The purpose of this analysis is to identify the most significant factors that influence employee attrition. The approach includes:

- **Exploratory Data Analysis (EDA):** to investigate patterns and trends within the data.
- **Structured Query Language (SQL):** to extract aggregated insights using relational queries.
- **Machine Learning (Decision Tree Classifier):** to model and predict employee attrition based on various features.
- **Visualization tools (Seaborn, Matplotlib):** to represent findings in a clear and interpretable manner.

The results of this study aim to support Human Resource (HR) departments and management teams in making informed, data-driven decisions that can help reduce attrition and improve workforce engagement.

2. Objective

The primary objective of this project is to analyze the factors contributing to employee attrition using the IBM HR Analytics dataset and to derive actionable insights that can assist in improving employee retention strategies.

Specifically, the goals of this project include:

- To identify demographic, organizational, and behavioral patterns associated with employee attrition.
- To perform Exploratory Data Analysis (EDA) to uncover trends, outliers, and relationships between features.
- To extract key statistical insights using SQL queries within Python.
- To develop a machine learning model using Decision Tree classification to predict the likelihood of employee attrition.
- To visualize and interpret the model structure to understand the most influential factors.
- To present findings in a structured format that enables Human Resource (HR) professionals to take proactive steps in reducing attrition.

This analysis ultimately seeks to enhance understanding of attrition risks and support data-driven decision-making in HR planning and policy formulation.

3. Tools and Technologies Used

The analysis of employee attrition in this project has been carried out using a combination of tools and technologies that support data processing, analysis, visualization, querying, and modeling. Each tool was chosen based on its suitability for handling specific tasks involved in the project workflow.

Tool / Technology	Purpose
Python	Used as the primary programming language for data analysis, machine learning, and visualization.
Pandas	Utilized for data manipulation and preprocessing.
NumPy	Used for numerical operations during data preparation.
Matplotlib & Seaborn	Used for generating visualizations to support Exploratory Data Analysis (EDA).
scikit-learn	Applied for building and evaluating the Decision Tree classifier.
SQLite (via sqlite3)	Used for running SQL queries on the dataset within the Python environment.
pydotplus & Graphviz	Used to generate and visualize the decision tree model.
Excel	Used to preview and initially inspect the dataset.

These tools collectively enabled efficient data processing, analysis, interpretation, and presentation of results throughout the project.

4. Dataset Description

The dataset used for this project is the **IBM HR Analytics Employee Attrition & Performance** dataset. It contains detailed information about 1,470 employees across various job roles, departments, and experience levels within the organization. The dataset is publicly available and widely used for exploring HR-related analytics and predictive modeling.

Key Details:

- **Source:** IBM
- **Total Records:** 1,470 employees
- **Total Features:** 35 attributes (including target label)

Target Variable:

- **Attrition** — Indicates whether an employee has left the company ("Yes") or stayed ("No").

Types of Features:

Feature Type	Examples
Demographic	Age, Gender, MaritalStatus, Education
Job-related	JobRole, Department, BusinessTravel, Overtime
Performance	JobInvolvement, PerformanceRating, WorkLifeBalance
Compensation	MonthlyIncome, PercentSalaryHike, StockOptionLevel
Tenure	TotalWorkingYears, YearsAtCompany, YearsInRole

Data Quality:

- The dataset does not contain missing values.
- Several categorical columns required encoding before machine learning could be applied.
- Class imbalance is present in the target variable: **Attrition = 'No'** is significantly more frequent than **Attrition = 'Yes'**.

This structured and clean dataset provides a suitable foundation for both descriptive and predictive analytics.

5. Exploratory Data Analysis

Analyizing the variables

- Numerical Variables

5.1 Attrition vs Age

- **Hypothesis**: Younger employees are more likely to leave the organization.
- **Observation**:
  The box plot shows that the median age of employees who left ("Yes") is lower than that of those who stayed ("No"). Attrition is concentrated between ages 25 and 35. Very few employees older than 45 left the organization.

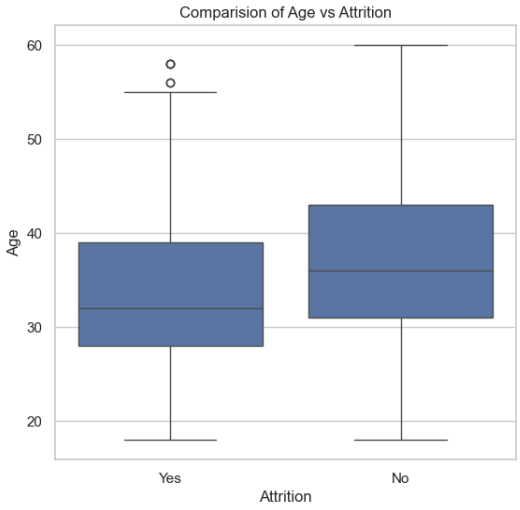
The histogram on the right illustrates that the overall age distribution is skewed toward employees aged between 28 and 40, with peak frequencies in the early 30s.

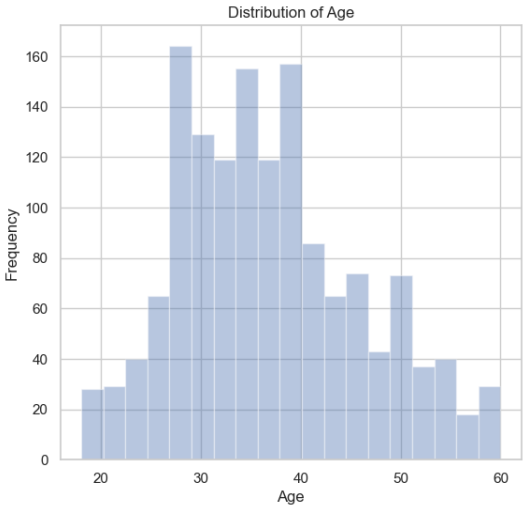
1. We found that median age of employee's in the company is 30 - 40 Yrs. Minimum age is 18 Yrs. and Maximum age is 60 Yrs.

2. From the Age Comparison boxplot, majority of people who left the company are below 40 Yrs. and among the people who didn't left the company are of age 32 to 40 years.

Age has an effect on attrition. So it is considered as influential variable for attrition.

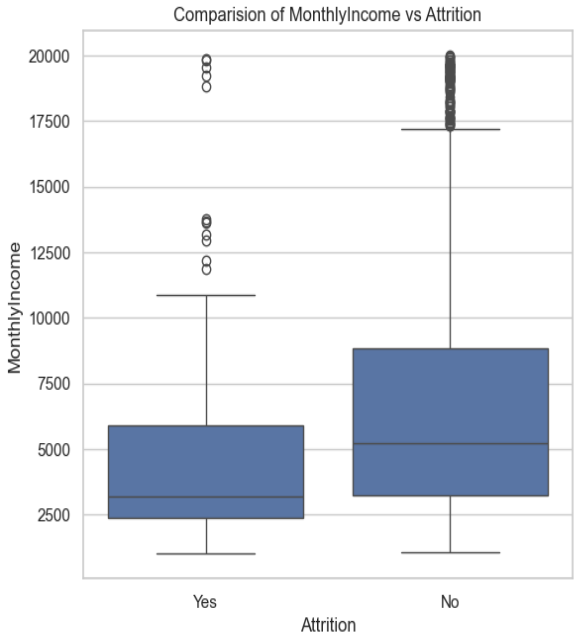
- **Conclusion**:
  Age is an influential factor in employee attrition. Younger employees, particularly those under 35, are more prone to resignation. This may be due to job switching trends early in careers or lower organizational commitment in the younger age group.
- **Visual Reference**:

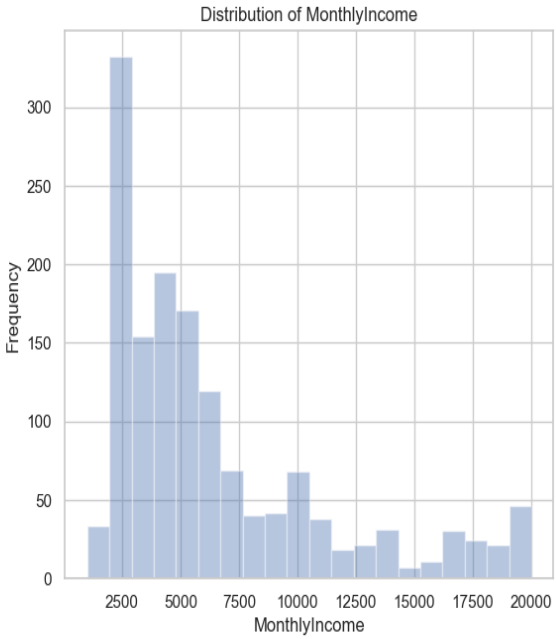




5.2 Attrition vs Monthly Income

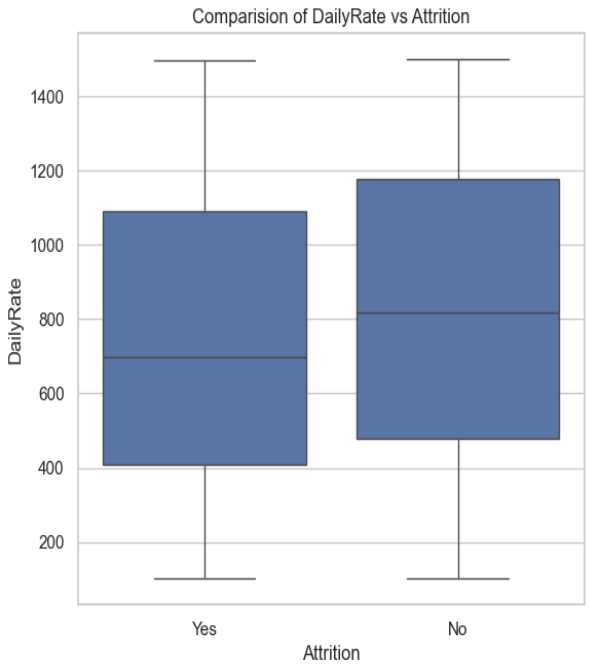
- **Hypothesis**: Employees with lower income are more likely to leave.
- **Observation**: Employees earning below $5,000/month have the highest attrition. Higher income brackets show stronger retention.
- **Conclusion**: Monthly income is a strong indicator of attrition risk.
- **Visual Reference**

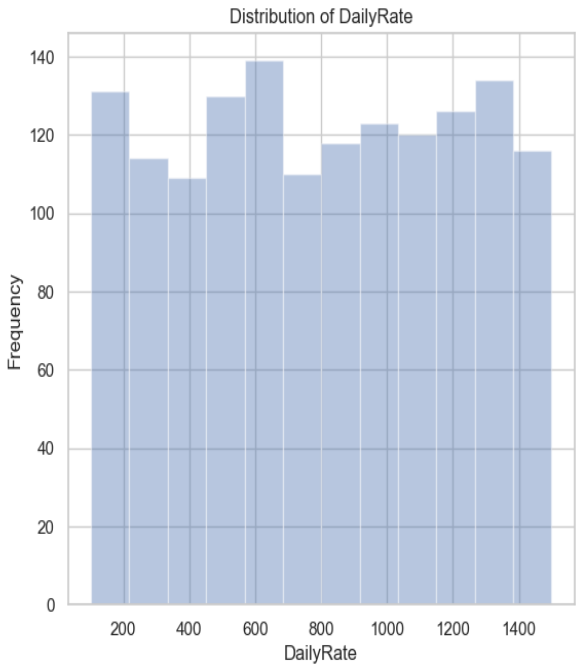




5.3 Attrition vs Daily Rate

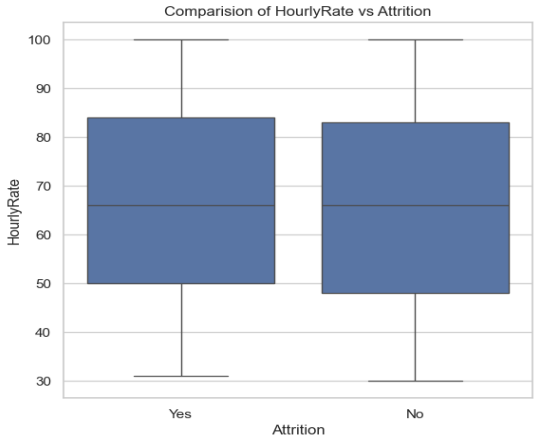
- **Hypothesis**: Employees with lower daily earnings might have higher attrition.
- **Observation**: There is no significant difference in daily rate distribution between employees who stayed and those who left.
- **Conclusion**: Daily rate does not appear to have a strong influence on attrition.
- **Visual Reference**:

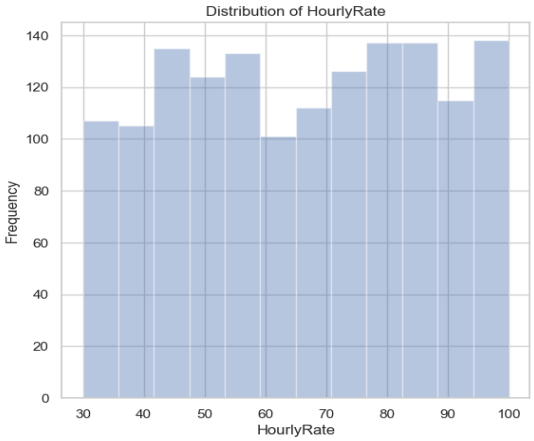




5.4 Attrition vs Hourly Rate

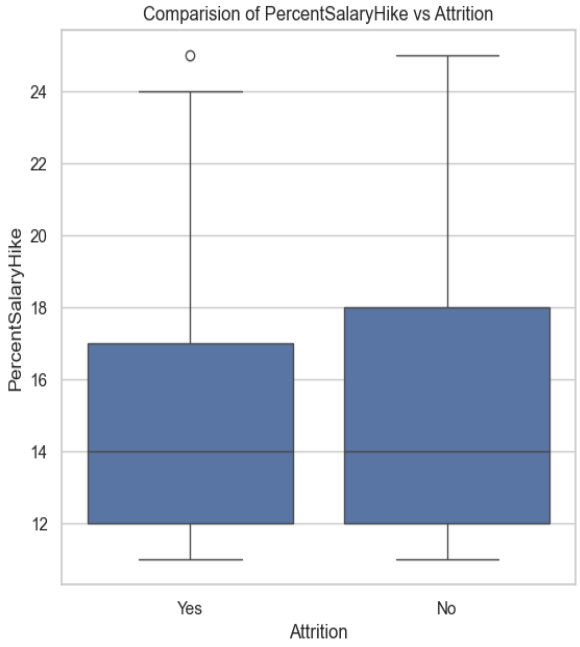
- **Hypothesis**: Hourly earnings may impact job satisfaction and retention.
- **Observation**: The distribution of hourly rates is nearly identical for both attrition categories.
- **Conclusion**: Hourly rate is not a significant predictor of attrition.
- **Visual Reference**:

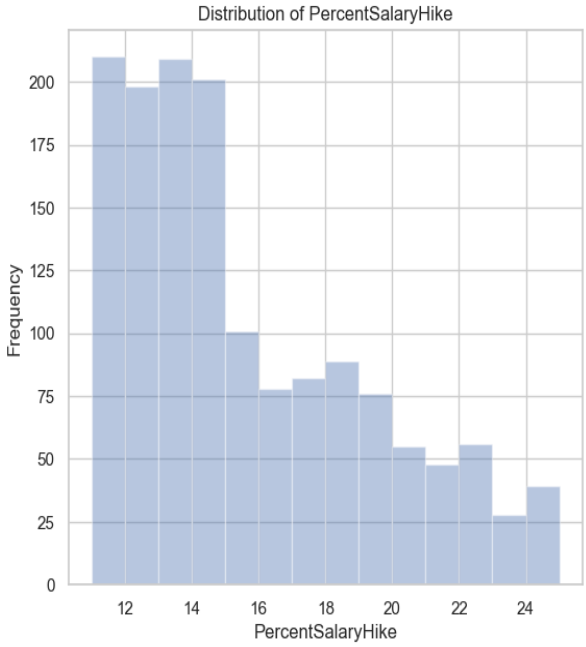




5.5 Attrition vs Percent Salary Hike

- **Hypothesis**: Employees who receive lower salary hikes may be more likely to leave.
- **Observation**: There is no major difference in salary hike distribution between the two groups.
- **Conclusion**: Percent salary hike alone is not a strong factor in attrition.
- **Visual Reference**:

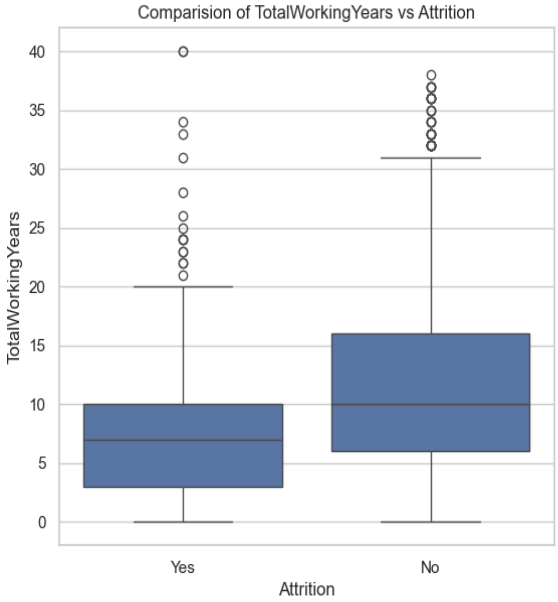


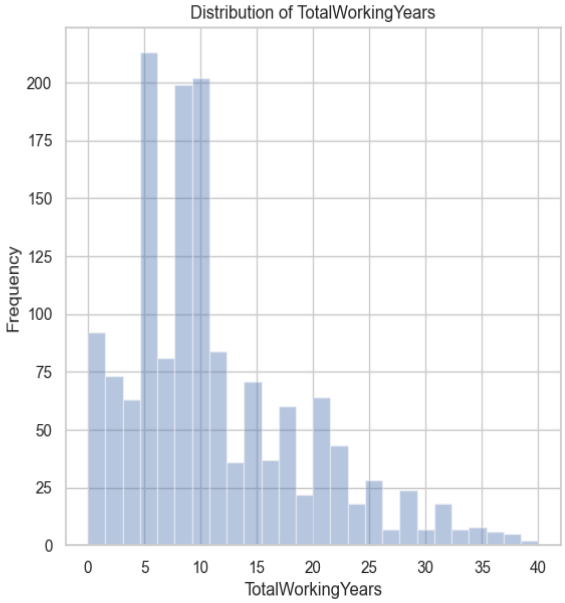


- Majority (60% of total strength) of employee's receive 16% salary hike in the company, employee's who received less salary hike have left the company.

5.6 Attrition vs Total Working Years

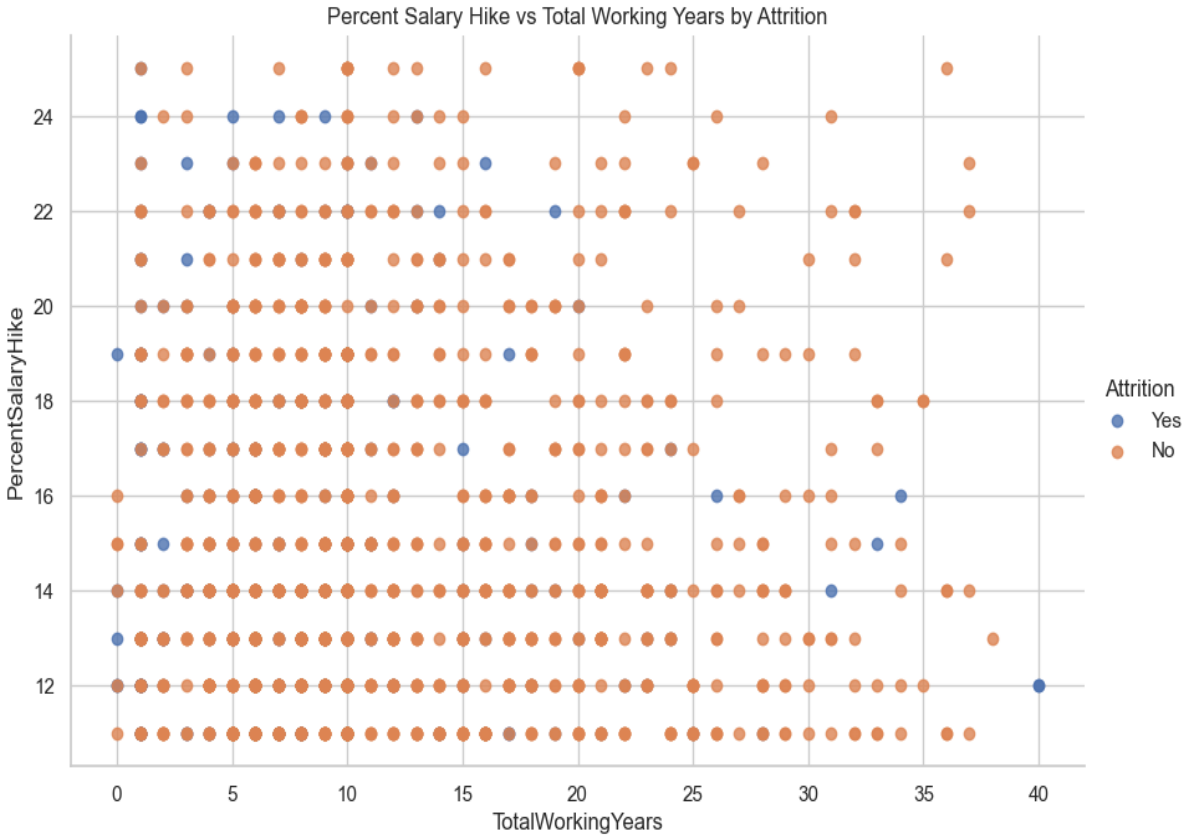
- **Hypothesis**: Employees with fewer years of experience may have higher attrition.
- **Observation**: Employees with less than 10 years of experience show significantly higher attrition.
- **Conclusion**: Total working years is a strong predictor; less experienced employees are more likely to leave.
- **Visual Reference**:





5.7 Percent Salary Hike vs Total Working Years (Colored by Attrition)

- **Purpose**: To analyze the interaction between experience, salary hikes, and attrition.
- **Observation**: Employees with fewer working years and low-to-moderate hikes show more attrition (blue dots).
- **Conclusion**: Salary hikes may matter more for early-career employees.
- **Visual Reference**:



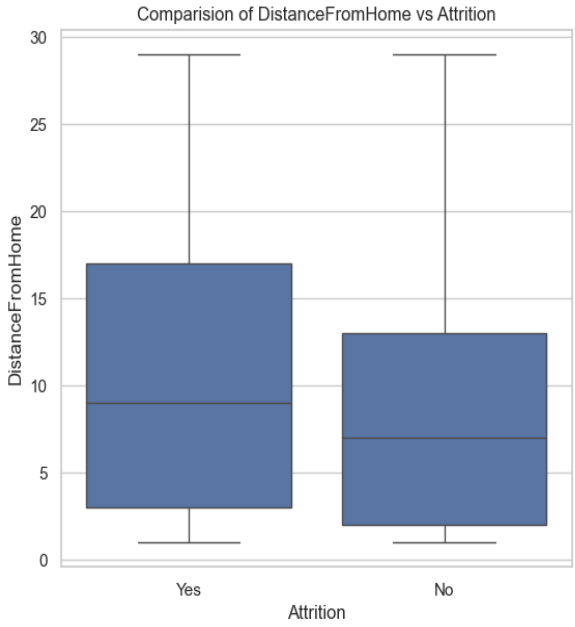
- Employee's with less working years have received 25% Salary hike when they switch to another company, but there is no linear relationship between working years and salary hike.

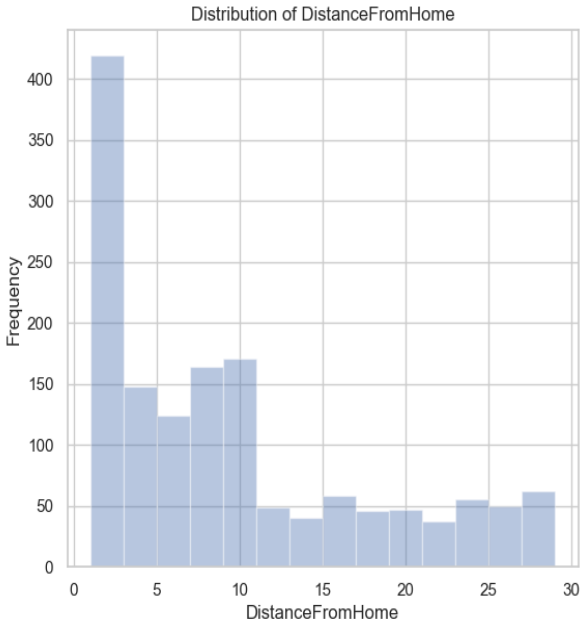
- Attrition is not seen amomg the employee's having more than 20 years of experience if their salary hike is more than 20%, even if the salary hike is below 20% attrition rate among the employee's is very low.

- Employee's with lesser years of experience are prone to leave the company in search of better pay, irrespective of salary hike

5.8 Attrition vs Distance From Home

- **Hypothesis**: Long commuting distance influences resignation.
- **Observation**: Employees living more than 10 km from work are more likely to leave. However, short-distance commuters also show some attrition.
- **Conclusion**: Distance from home moderately affects attrition.
- **Visual Reference**:



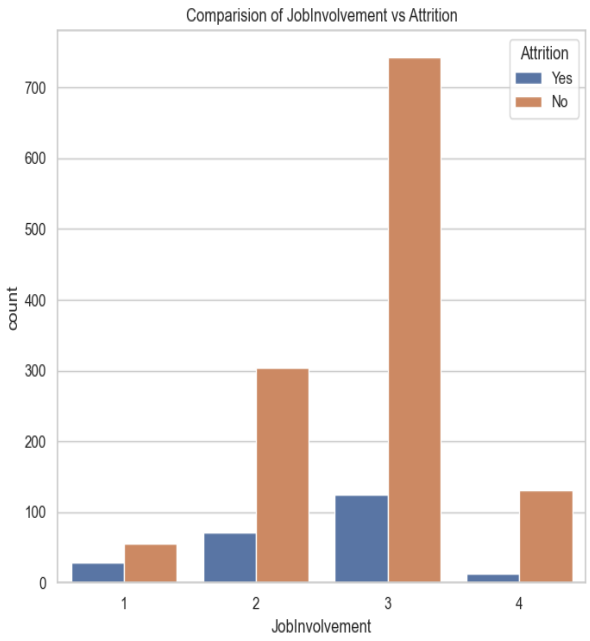


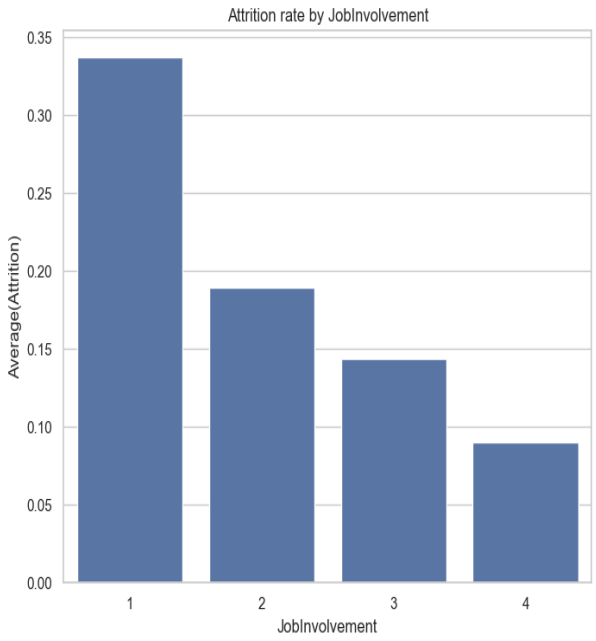
- There is a higher number of people who reside near to offices and hence the attrition levels are lower for distance less than 10. With increase in distance from home, attrition rate also increases

Analyzing the variables

- Categorical Variables

Job Involvement



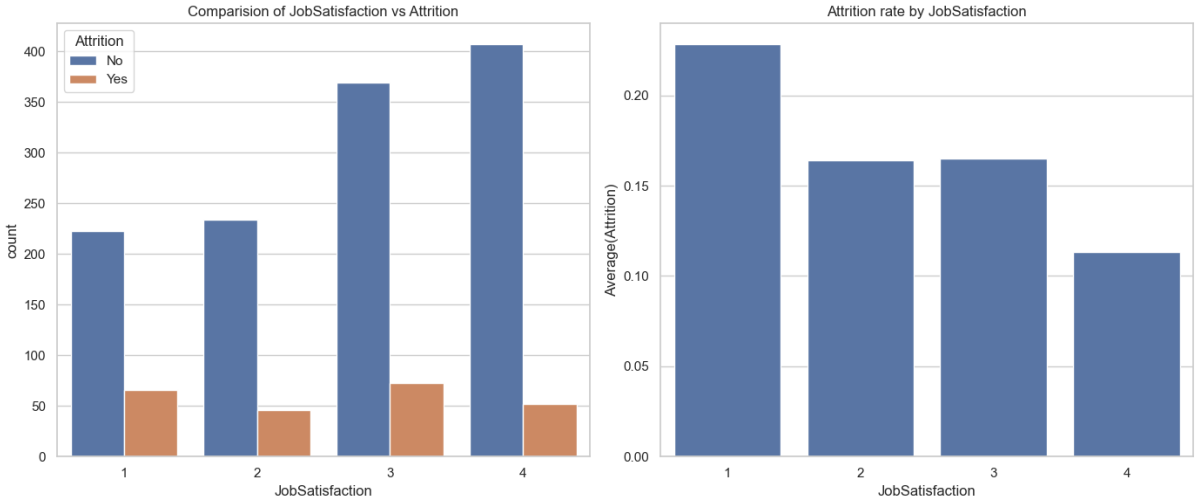


1. In the total data set, 59% have high job involvement whereas 25% have medium involvement rate

2. From above plot we can observe that round 50% of people in low job involvement (level 1 & 2) have left the company.

3. Even the people who have high job involmenent have higher attrition rate around 15% in that category have left company

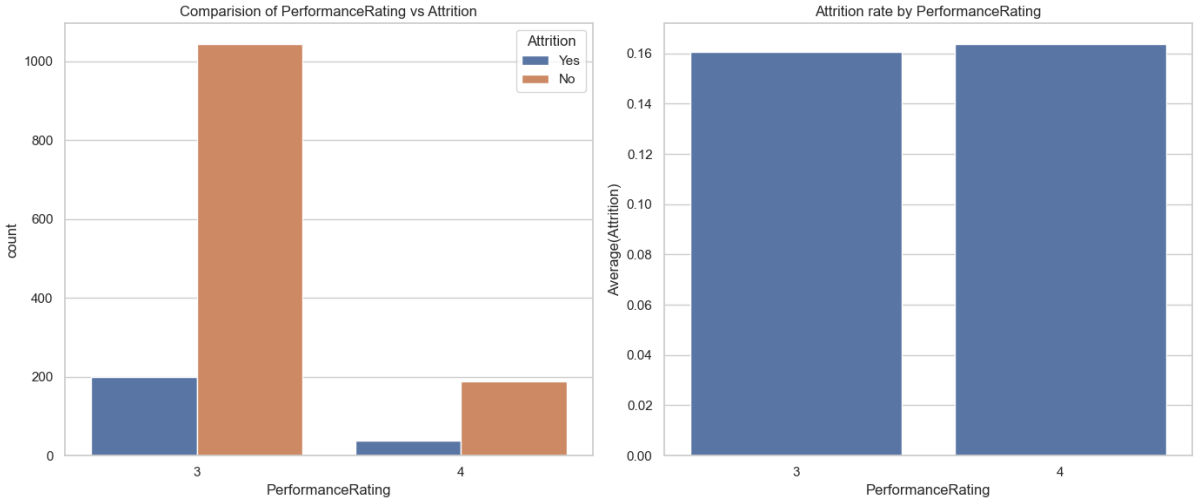
JobSatisfaction



As expected, people with low satisfaction have left the company around 23% in that category. what surprising is out of the people who rated medium and high job satisfaction around 32% has left the company. There should be some other factor which triggers their exit from the company

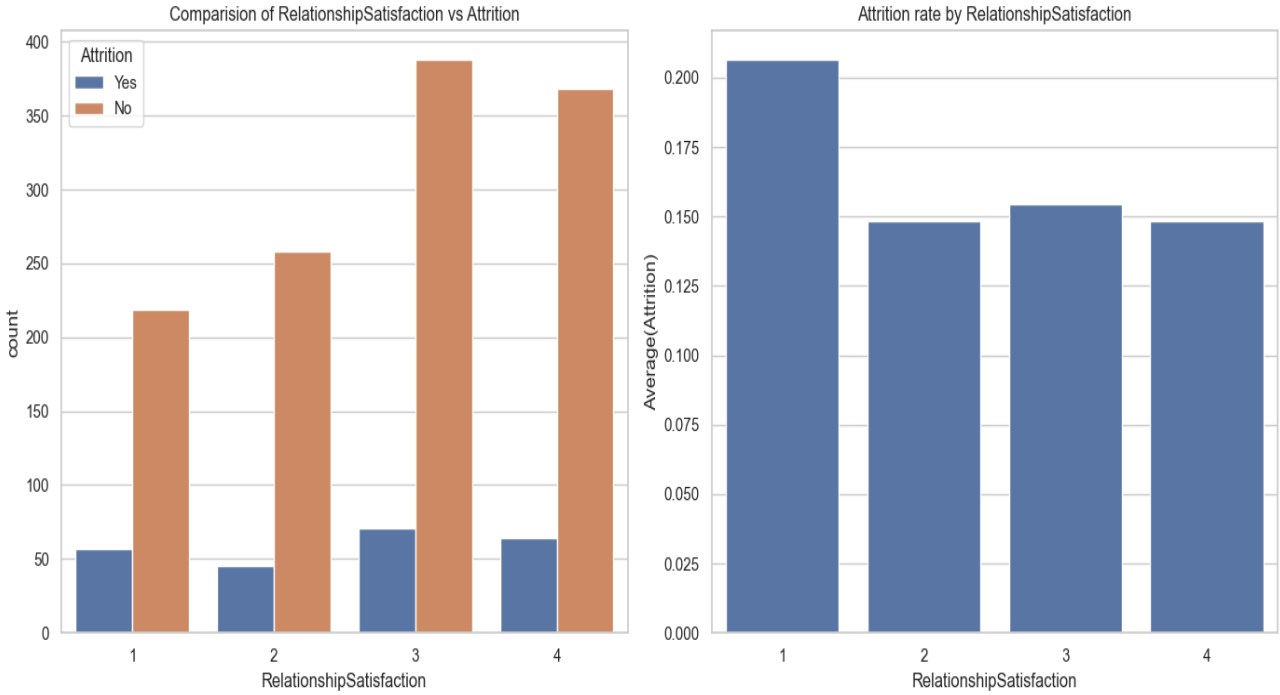
Performance Rating

Around 85% of people in the company rated as Excellent and remaining 15% rated as Outstanding



Contrary to normal belief that employee's having higher rating will not leave the company. It may be seen that there is no significant difference between the performance rating and Attrition Rate.

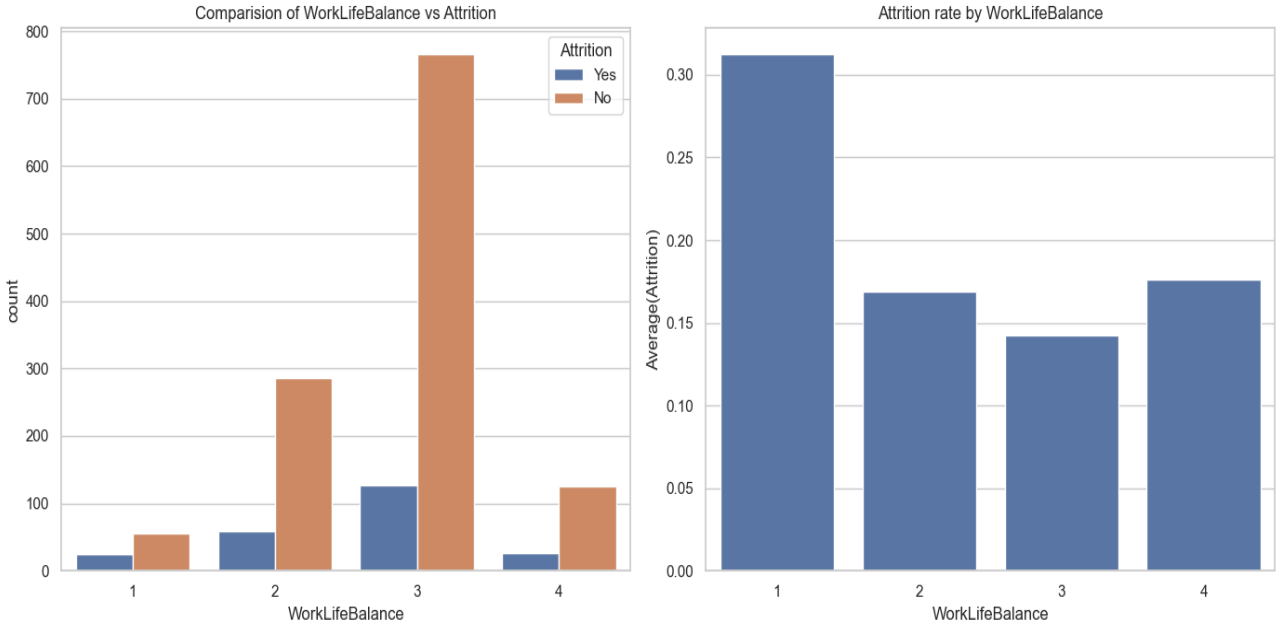
RelationshipSatisfaction



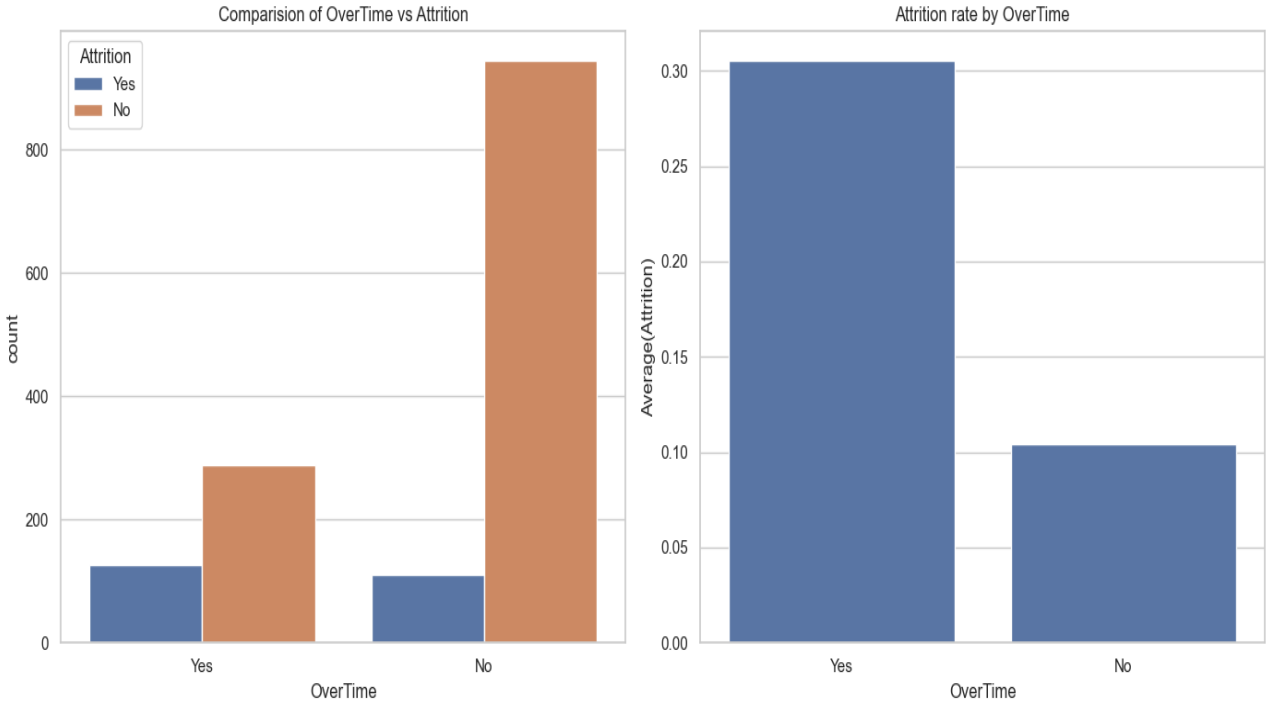
In this too, we found that almost 30% of employees with high and very high RelationshipSatisfaction have left the company. Here also there is no visible trend among the relationshipsatisfaction and attrition rate.

WorkLifeBalance

More than 60% of the employee's rated that they have Better worklife balance and 10% rated for Best worklife balance.

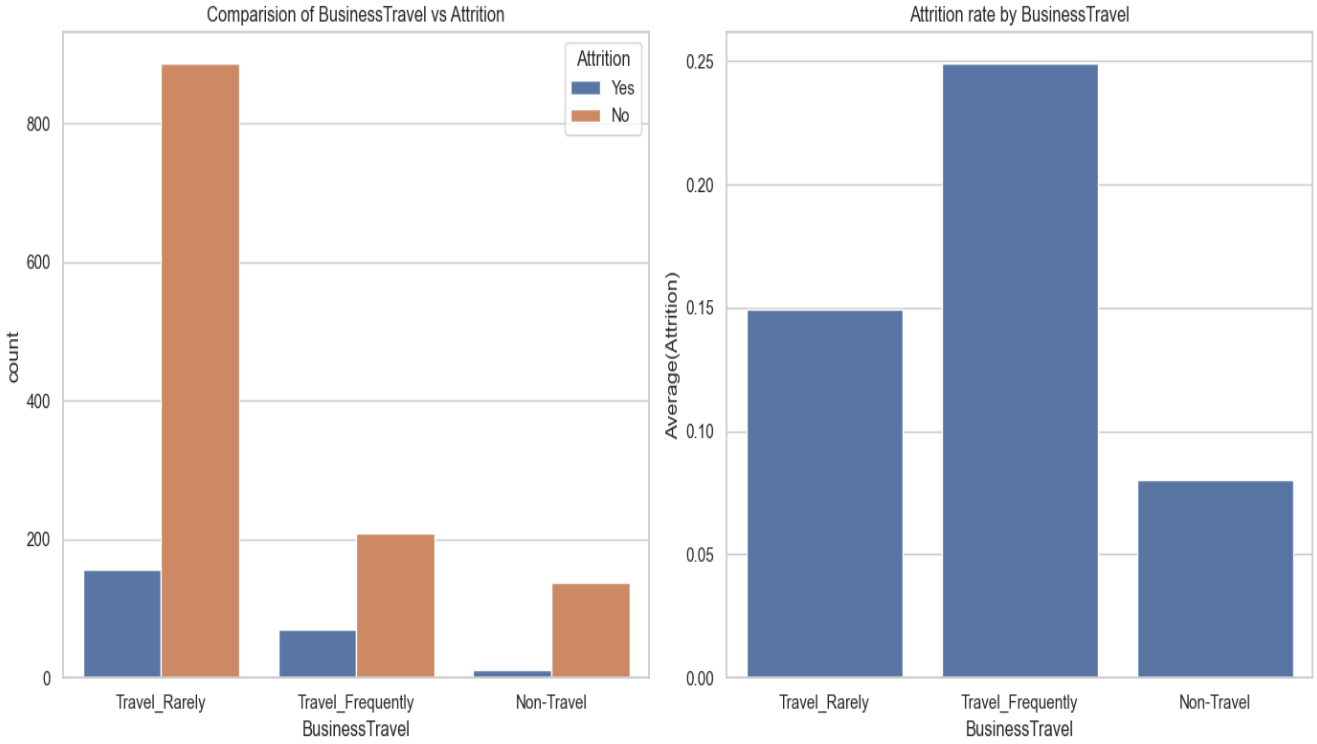


- As expected more than 30% of the people who rated as Bad WorkLifeBalance have left the company and around 15% of the people who rated for Best WorkLifeBalance also left the company.



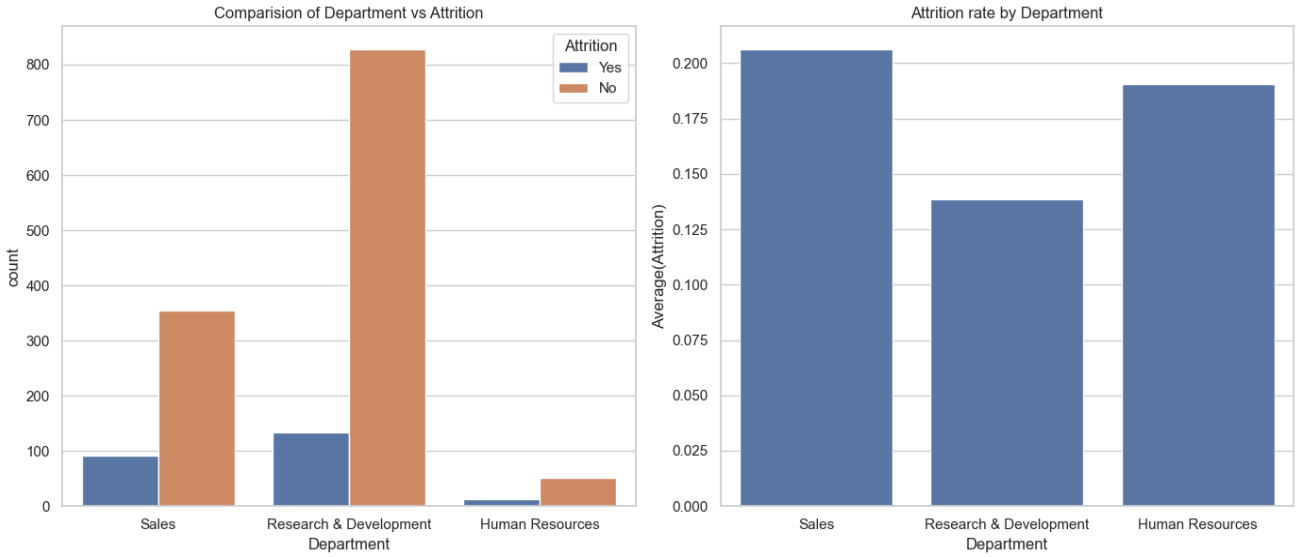
More than 30% of employee's who worked overtime has left the company, where as 90% of employee's who have not experienced overtime has not left the company. Therefore overtime is a strong indicator of attrition.

BusinessTravel



- There are more people who travel rarely compared to people who travel frequently. In case of people who travel Frequently around 25% of people have left the company and in other cases attrition rate doesn't vary significantly on travel.

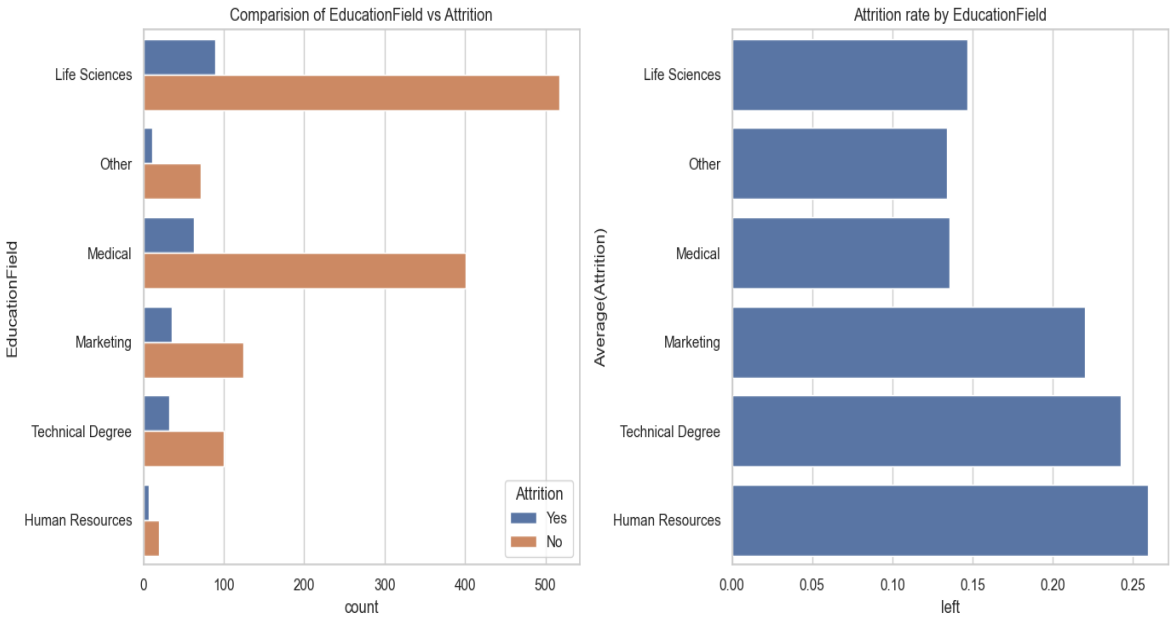
Department

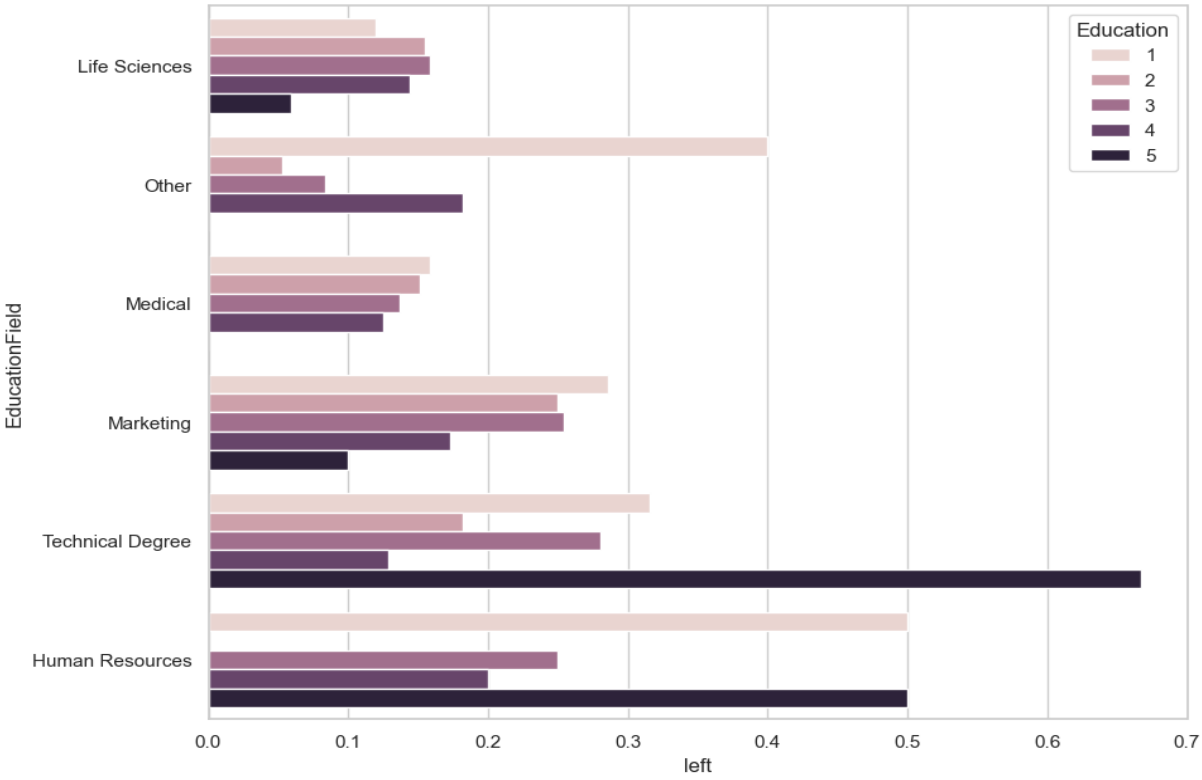


- On comparing departmentwise,we can conclude that HR has seen only a marginal high in turnover rates whereas the numbers are significant in sales department with turnover rates of 39 %. The attrition levels are not appreciable in R & D where 67 % have recorded no attrition.

- Sales has seen higher attrition levels about 20.6% followed by HR around 18%

EducationField



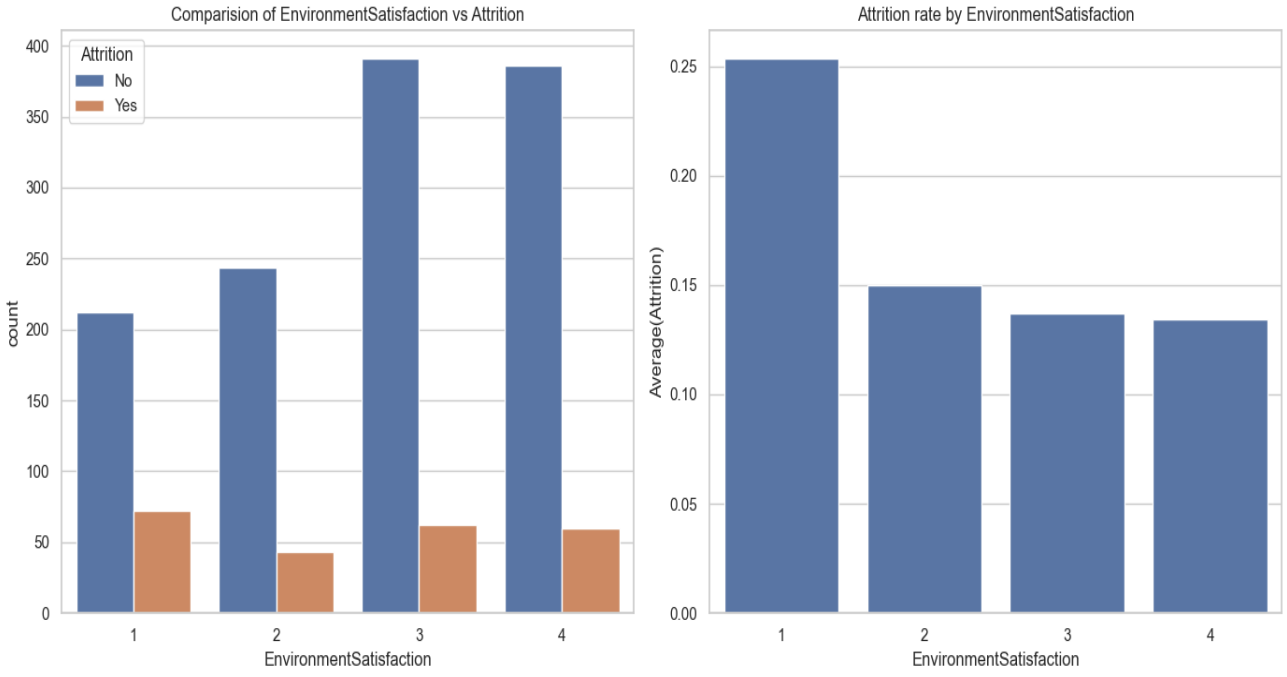


- There are more people with a Life sciences followed by medical and marketing.

- Employee's in the EducationField of Human Resources and Technical Degree have highest attrition levels around 26% and 23% respectively.

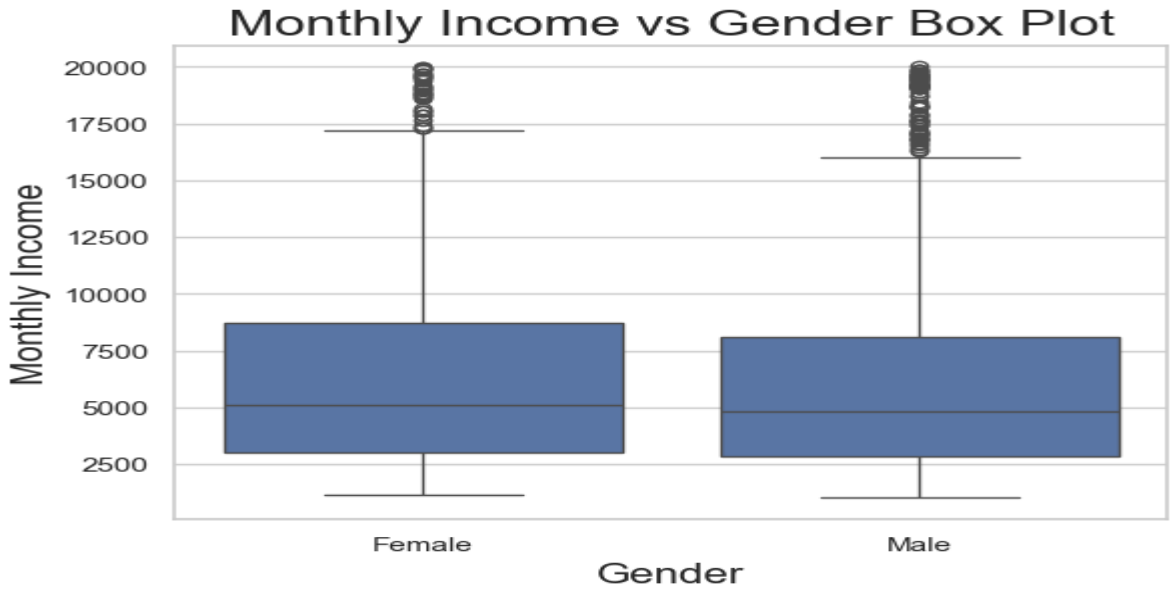
- When compared with Education level, we have observed that employees in the highest level of education in there field of study have left the company. We can conclude that EducationField is a strong indicator of attrition.

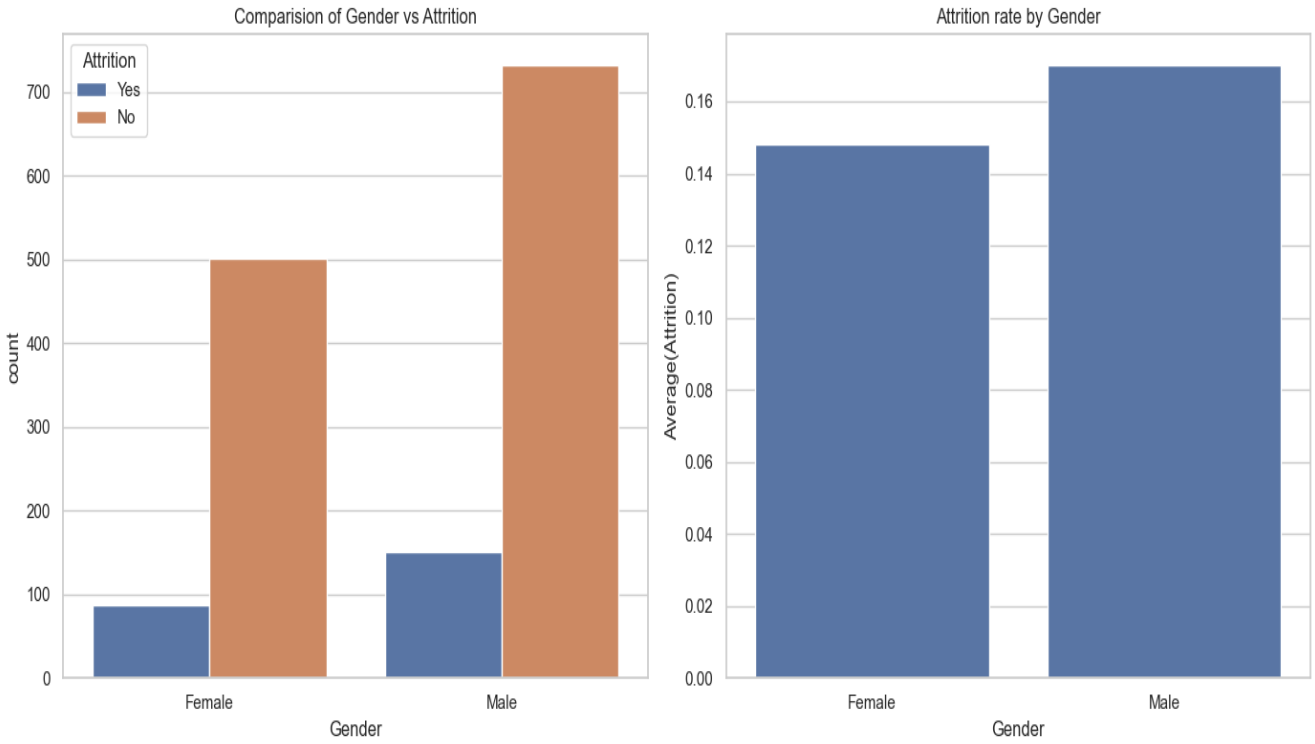
EnvironmentSatisfaction



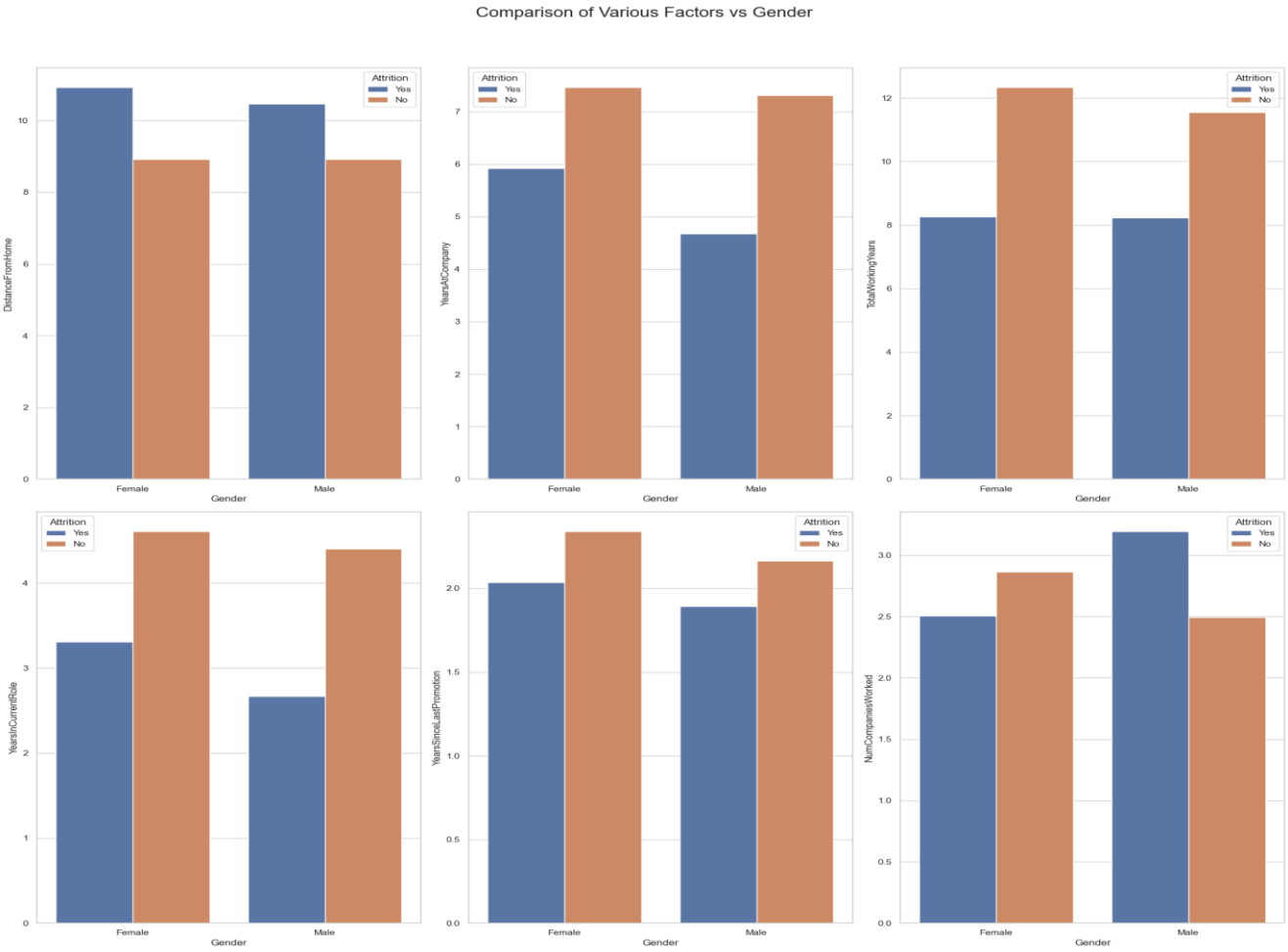
we can see that people having low environment satisfaction 25% leave the company.

Gender Vs Attrition





- Monthly Income distribution for Male and Female is almost similar, so the attrition rate of Male and Female is almost the same around 15%. Gender is not a strong indicator of attrition

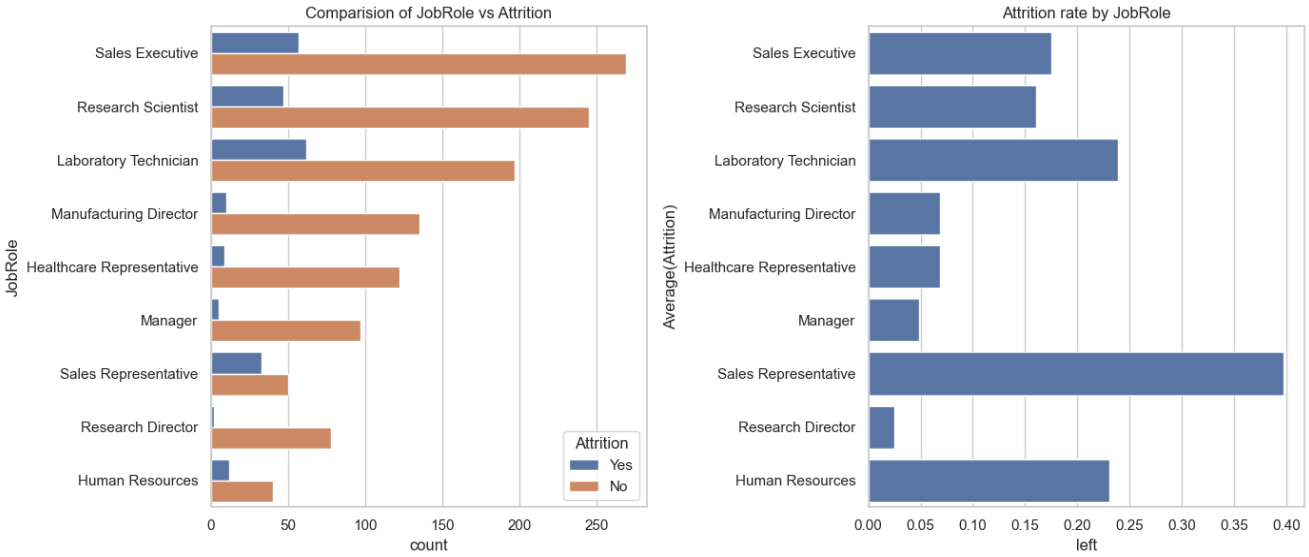


1. Distance from home matters to women employees more than men.

2. Female employes are spending more years in one company compare to their counterpart.

3. Female employes spending more years in current company are more inclined to switch.

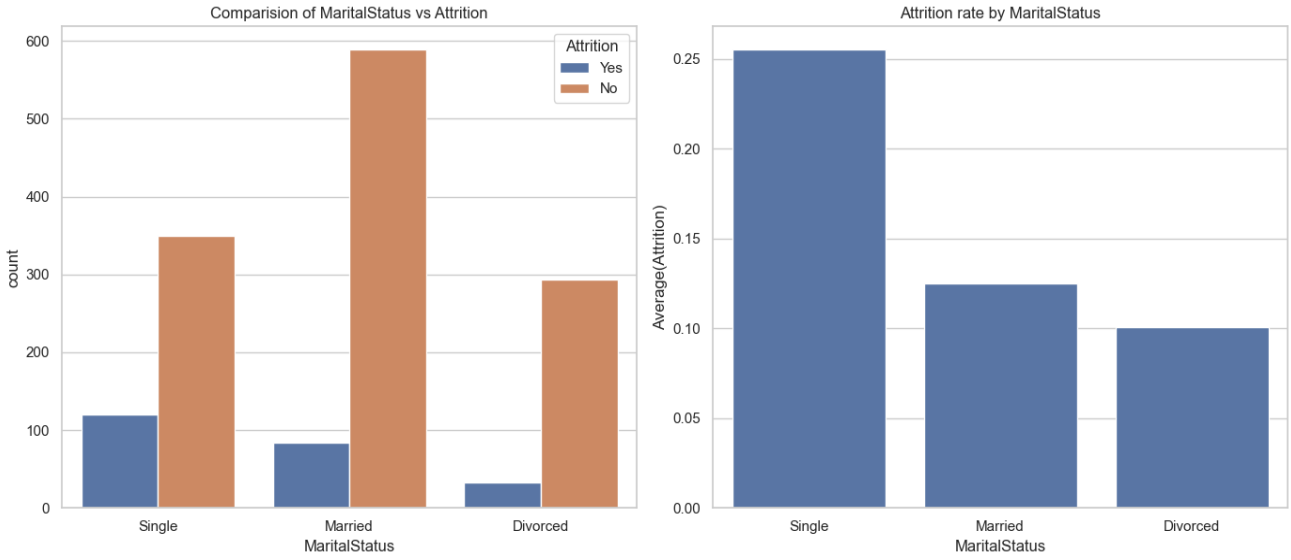
Job Role



1. Jobs held by the employee is maximum in Sales Executive, then R&D , then Laboratory Technician

2. People working in Sales department is most likely quit the company followed by Laboratory Technician and Human Resources there attrition rates are 40%, 24% and 22% respectively

Marital Status



From the plot,it is understood that irrespective of the marital status,there are large people who stay with the company and do not leave.Therefore,marital status is a weak predictor of attrition.

Decision Tree Modelling:

I have used Decision tree to create model. Decision Tree is a greedy algorithm it searches the entire space for possible decision trees. so we need to find an optimum parameter(s) or criteria for stopping the decision tree at some point. We use the hyperparameters to prune the decision tree. By using grid search best parameters was found to be –

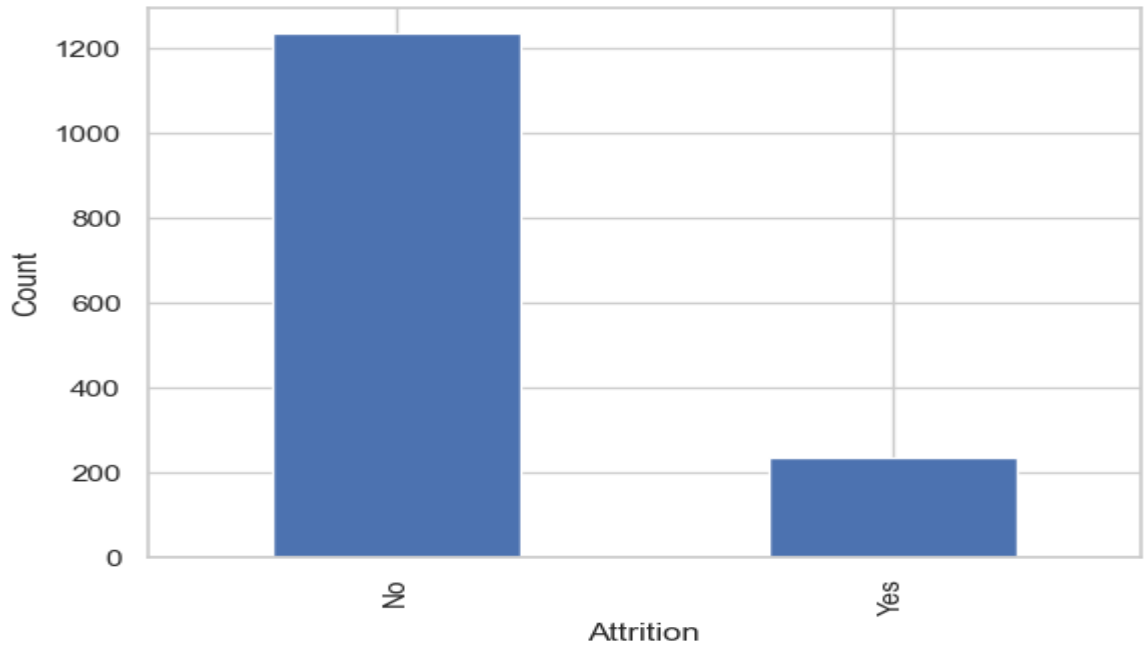
{'decisiontreeclassifier__max_depth': 3,

'decisiontreeclassifier__max_features': 4,

'decisiontreeclassifier__min_samples_leaf': 1,

'decisiontreeclassifier__min_samples_split': 2}

The major hurdle was we had 1223 unlabelled (No in Attrition) and 237 labelled (Yes in Attrition), which is a highly imbalanced data. So I used stratified sampling based on proportion of attrition in overall data.

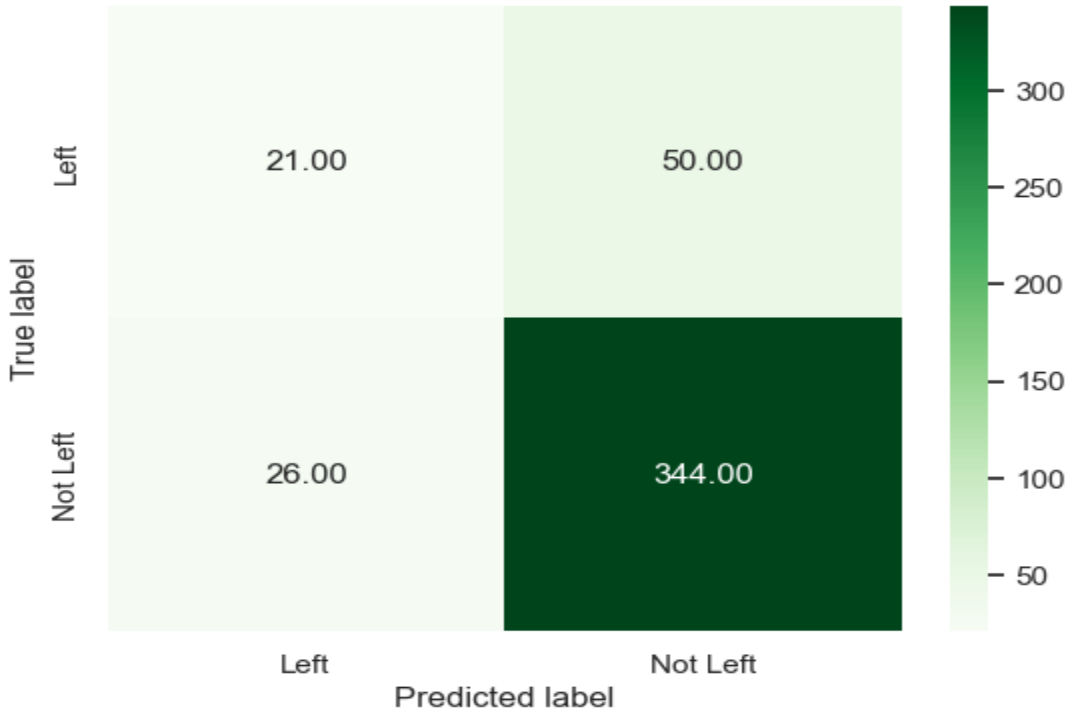


Model Evaluation:

I have used the k fold cross validation technique for assessing how the results of a model will generalize to an independent test data set. I used k =5 i.e.. 5 fold cross validation. Model was fit on the stratified sample data and tested on unmarked dataset.

Confusion Matrix:

The confusion matrix is a way of tabulating the number of misclassifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

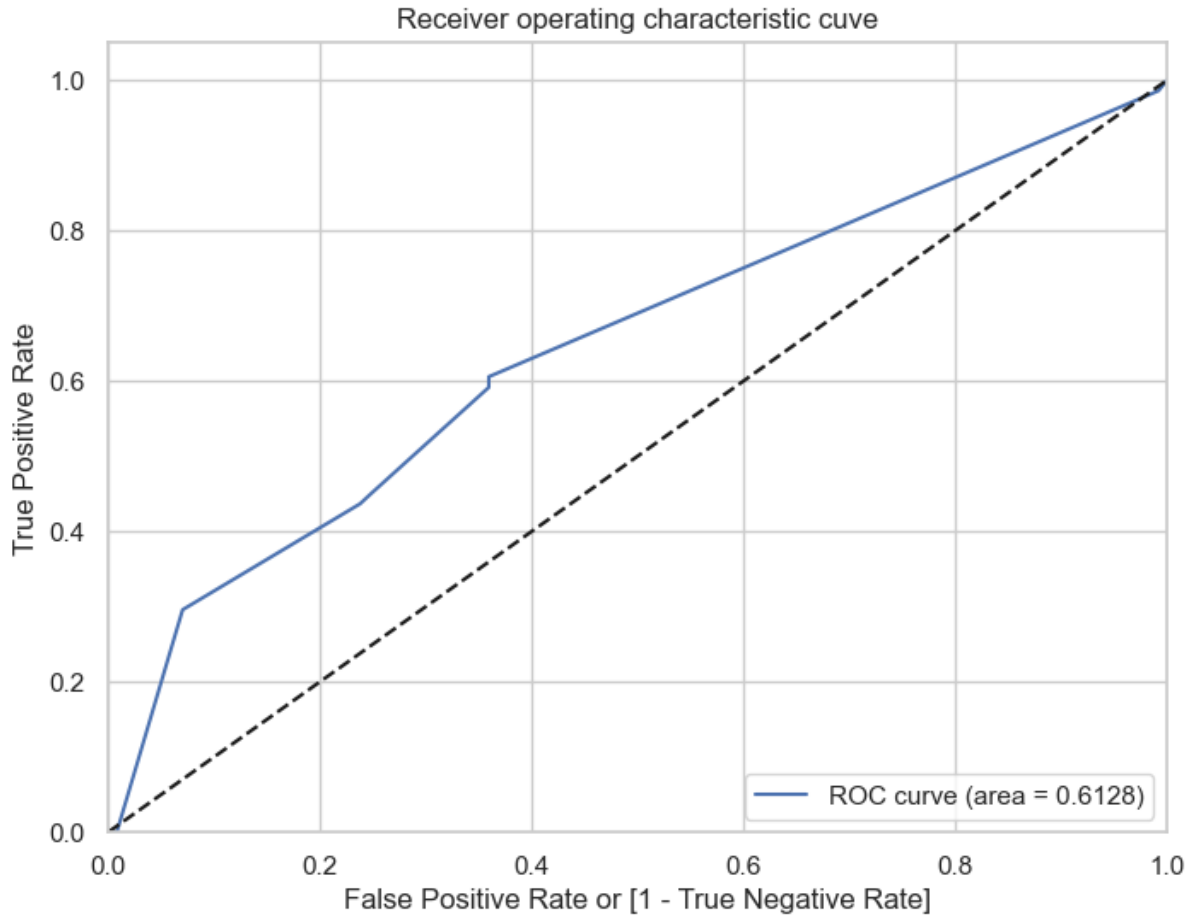


From the confusion matrix we have seen there are lot of misclassifications i.e.. 50 out of 71 were misclassified, this is due the fact that the data is highly imbalanced. Decision Tree algorithm is bias towards classes which have number of instances, here in this dataset we have more number of employees didn't leave the company. So the decision tree algorithm tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored.

Thus, there is a high probability of misclassification of the minority class as compared to the majority class. In the cases like this accuracy measures tell the story that you might have excellent accuracy but the accuracy is only reflecting the overall accuracy of the data distribution. Instead of using accuracy as a performance metric we will use 'Precision', 'Recall' and ROC Curve.

ROC Curve:

It is a curve between True Positive rate (Recall) and False Positive rate (1 – True Negative rate).



The ROC curve is a simple plot that shows the trade-off between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold. From the ROC Curve, we have a choice to make depending on the value we place on true positive and tolerance for false positive rate.

If we wish to find the more people who are leaving, we could increase the true positive rate by adjusting the probability cut-off for classification.

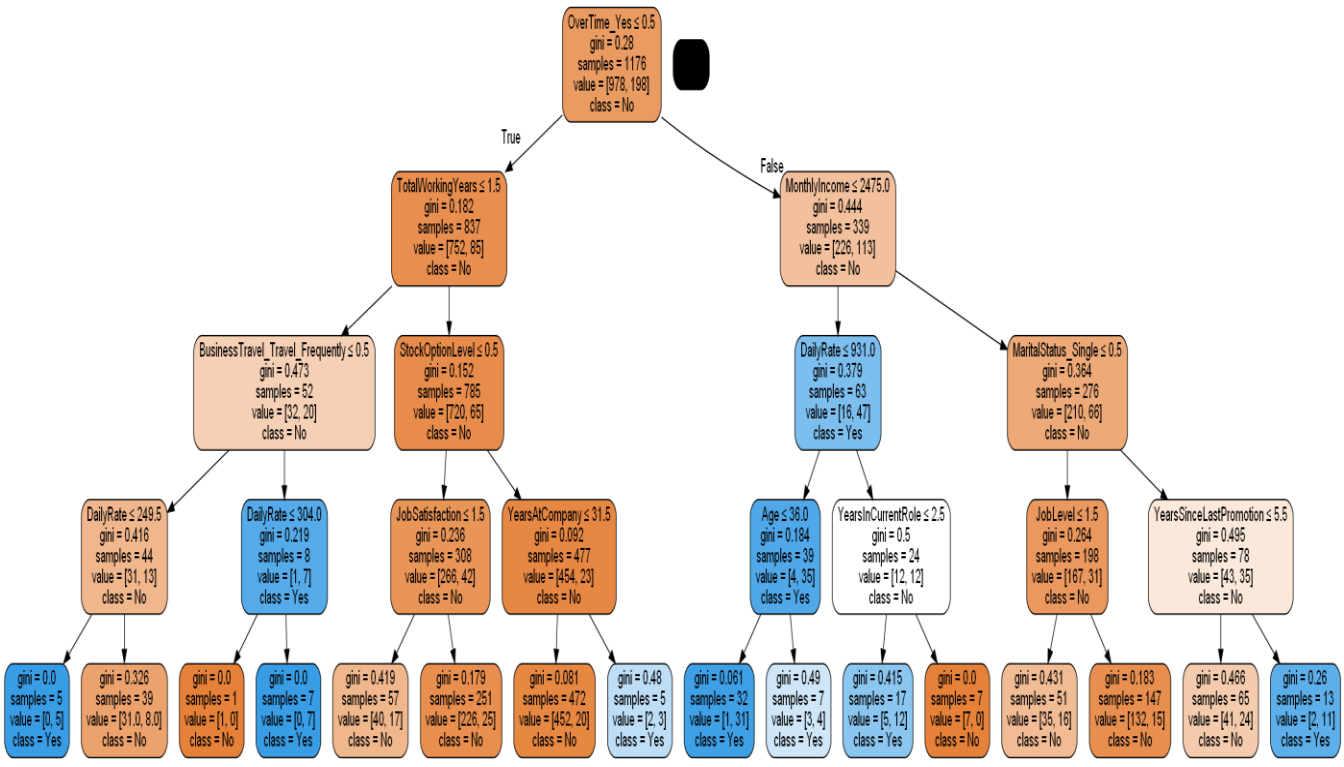
However by doing so would also increase the false positive rate. we need to find the optimum value of cut-off for classification.

From the classification report,

	precision	recall	f1-score	support
0	**0.87**	**0.93**	**0.90**	**370**
1	**0.45**	**0.30**	**0.36**	**71**
accuracy			**0.83**	**441**
macro avg	**0.66**	**0.61**	**0.63**	**441**
weighted avg	**0.80**	**0.83**	**0.81**	**441**

Because of imbalanced data we got less f1 score.

6. Decision Tree Visualization



• The dependent variable for the decision tree is a OverTime which has two classes Yes or No.

• The most influential attribute to determine how to classify employee will leave or not is the Monthly Income variable.

• Even though the employee's having the average monthly income if their job involvement is average or less than average, those employees will most likely leave the company.

• Employee's whose monthly income is less than 2000 approximately will leave the company even if their total working years is more than 6 months.

Suggested Actions:

It's not sensible to focus on every employee who wants to leave because it costs time and energy for human resources management department. HR department need to focus on:

• Improving the work conditions Provide an option for the employee's to work from home, on a flexible schedule, or in an office with an ergonomic workspace, they will be more satisfied with their work and more likely to achieve a healthy work-life balance.

• Offer modest salaries and perks To maintain the critical employee's company need's to offer equitable and modest salaries. You can also give added perks like flexible schedules, travel discounts etc.

• Employee Engagement When you have talented employee's we need to find ways that you can help expand the employee's skill set, so that their involvement in the job increases. If their involvement is low, they will get bored and think that they are not growing within the organization

7. SQL Queries

```
import pandas as pd

import sqlite3


#  Load the dataset

employee_data = pd.read_excel("Attrition.xlsx")  # Or use read_csv if CSV is used


#  Create in-memory SQLite DB

conn = sqlite3.connect(":memory:")

employee_data.to_sql("employee_data", conn, index=False, if_exists="replace")


#  1. Count of employees by Gender

query1 = """

SELECT Gender, COUNT(*) AS Count
```

```
FROM employee_data
GROUP BY Gender;
"""
print("\n1. Employee Count by Gender:")
print(pd.read_sql_query(query1, conn))

#  2. Average Monthly Income by Attrition
query2 = """
SELECT Attrition, AVG(MonthlyIncome) AS Avg_Income
FROM employee_data
GROUP BY Attrition;
"""
print("\n2. Average Monthly Income by Attrition:")
print(pd.read_sql_query(query2, conn))

#  3. Department-wise Attrition Count
query3 = """
SELECT Department, Attrition, COUNT(*) AS Count
FROM employee_data
GROUP BY Department, Attrition;
"""
print("\n3. Department-wise Attrition Count:")
print(pd.read_sql_query(query3, conn))

#  4. High-Income Employees Who Left
query4 = """
SELECT EmployeeNumber, MonthlyIncome
FROM employee_data
WHERE MonthlyIncome > 15000 AND Attrition = 'Yes';
"""
print("\n4. High-Income Employees Who Left:")
```

```
print(pd.read_sql_query(query4, conn))

#  5. Attrition by Job Role
query5 = """
SELECT JobRole, Attrition, COUNT(*) AS Count
FROM employee_data
GROUP BY JobRole, Attrition;
"""
print("\n5. Attrition by Job Role:")
print(pd.read_sql_query(query5, conn))

# 6. Average Age and Distance from Home by Gender
query6 = """
SELECT Gender, AVG(Age) AS Avg_Age, AVG(DistanceFromHome) AS Avg_Distance
FROM employee_data
GROUP BY Gender;
"""
print("\n6. Avg Age & Distance by Gender:")
print(pd.read_sql_query(query6, conn))

#  7. Average Job Satisfaction by Department
query7 = """
SELECT Department, AVG(JobSatisfaction) AS Avg_Satisfaction
FROM employee_data
GROUP BY Department;
"""
print("\n7. Avg Job Satisfaction by Department:")
print(pd.read_sql_query(query7, conn))

#  8. OverTime vs Attrition
query8 = """
```

```
SELECT OverTime, Attrition, COUNT(*) AS Count
FROM employee_data
GROUP BY OverTime, Attrition;
"""
print("\n8. OverTime vs Attrition:")
print(pd.read_sql_query(query8, conn))

#  9. Experienced Employees Who Left (>10 years)
query9 = """
SELECT EmployeeNumber, TotalWorkingYears, Attrition
FROM employee_data
WHERE TotalWorkingYears > 10 AND Attrition = 'Yes';
"""
print("\n9. Experienced Employees Who Left:")
print(pd.read_sql_query(query9, conn))

#  10. Income by Education Field & Attrition
query10 = """
SELECT EducationField, Attrition, AVG(MonthlyIncome) AS Avg_Income
FROM employee_data
GROUP BY EducationField, Attrition;
"""
print("\n10. Income by Education Field & Attrition:")
print(pd.read_sql_query(query10, conn))

#  11. Attrition by Business Travel Category
query11 = """
SELECT BusinessTravel, Attrition, COUNT(*) AS Count
FROM employee_data
GROUP BY BusinessTravel, Attrition;
"""
```

```
print("\n11. Attrition by Business Travel Category:")
print(pd.read_sql_query(query11, conn))

#  Close the connection
conn.close()
```

output:

1. Employee Count by Gender:

	Gender	Count
0	Female	588
1	Male	882

2. Average Monthly Income by Attrition:

	Attrition	Avg_Income
0	No	6832.739659
1	Yes	4787.092827

3. Department-wise Attrition Count:

	Department	Attrition	Count
0	Human Resources	No	51
1	Human Resources	Yes	12
2	Research & Development	No	828
3	Research & Development	Yes	133
4	Sales	No	354
5	Sales	Yes	92

4. High-Income Employees Who Left:

	EmployeeNumber	MonthlyIncome

0	58	19545
1	787	19859
2	825	19246
3	1038	19845
4	1277	18824

5. Attrition by Job Role:

	JobRole	Attrition	Count
0	Healthcare Representative	No	122
1	Healthcare Representative	Yes	9
2	Human Resources	No	40
3	Human Resources	Yes	12
4	Laboratory Technician	No	197
5	Laboratory Technician	Yes	62
6	Manager	No	97
7	Manager	Yes	5
8	Manufacturing Director	No	135
9	Manufacturing Director	Yes	10
10	Research Director	No	78
11	Research Director	Yes	2
12	Research Scientist	No	245
13	Research Scientist	Yes	47
14	Sales Executive	No	269
15	Sales Executive	Yes	57
16	Sales Representative	No	50
17	Sales Representative	Yes	33

6. Avg Age & Distance by Gender:

	Gender	Avg_Age	Avg_Distance
0	Female	37.329932	9.210884
1	Male	36.653061	9.180272

7. Avg Job Satisfaction by Department:

	Department	Avg_Satisfaction
0	Human Resources	2.603175
1	Research & Development	2.726327
2	Sales	2.751121

8. OverTime vs Attrition:

	OverTime	Attrition	Count
0	No	No	944
1	No	Yes	110
2	Yes	No	289
3	Yes	Yes	127

9. Experienced Employees Who Left:

	EmployeeNumber	TotalWorkingYears	Attrition
0	42	19	Yes
1	58	23	Yes
2	64	23	Yes
3	163	12	Yes
4	165	40	Yes
5	179	18	Yes
6	282	17	Yes
7	291	14	Yes
8	328	13	Yes
9	342	17	Yes
10	392	19	Yes
11	433	11	Yes
12	445	12	Yes
13	582	15	Yes
14	587	12	Yes

15	590	11	Yes
16	723	18	Yes
17	787	24	Yes
18	825	40	Yes
19	842	12	Yes
20	967	16	Yes
21	970	17	Yes
22	986	22	Yes
23	1033	15	Yes
24	1038	33	Yes
25	1042	18	Yes
26	1081	18	Yes
27	1098	24	Yes
28	1101	14	Yes
29	1127	21	Yes
30	1165	11	Yes
31	1167	22	Yes
32	1277	26	Yes
33	1360	31	Yes
34	1372	24	Yes
35	1420	20	Yes
36	1457	28	Yes
37	1458	11	Yes
38	1489	16	Yes
39	1522	11	Yes
40	1534	15	Yes
41	1572	34	Yes
42	1639	15	Yes
43	1667	13	Yes
44	1691	19	Yes
45	1716	25	Yes

46	1733	16	Yes
47	1758	11	Yes
48	1807	11	Yes
49	1821	13	Yes
50	1869	14	Yes
51	1968	15	Yes
52	2032	14	Yes
53	2044	12	Yes
54	2055	20	Yes

10. Income by Education Field & Attrition:

	EducationField	Attrition	Avg_Income
0	Human Resources	No	8579.950000
1	Human Resources	Yes	3416.000000
2	Life Sciences	No	6775.437137
3	Life Sciences	Yes	4650.022472
4	Marketing	No	7569.774194
5	Marketing	Yes	6564.942857
6	Medical	No	6800.805486
7	Medical	Yes	4659.269841
8	Other	No	6422.704225
9	Other	Yes	3805.000000
10	Technical Degree	No	6284.810000
11	Technical Degree	Yes	4112.968750

11. Attrition by Business Travel Category:

	BusinessTravel	Attrition	Count
0	Non-Travel	No	138
1	Non-Travel	Yes	12
2	Travel_Frequently	No	208
3	Travel_Frequently	Yes	69

4 Travel_Rarely No 887

5 Travel_Rarely Yes 156

8. Machine Learning Model Development

This section focuses on building a predictive model using supervised learning to classify whether an employee is likely to leave the company (attrition = Yes or No) based on HR-related features.

8.1 Objective

Build a classification model to predict employee attrition using historical data and extract important influencing features.

8.2 Data Preprocessing

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder

# Load data
df = pd.read_csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")

# Encode categorical features
le = LabelEncoder()
for column in df.select_dtypes(include=['object']):
    df[column] = le.fit_transform(df[column])

# Define features and target
X = df.drop(columns=['Attrition'])
y = df['Attrition']

# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

8.3 Model Building (Decision Tree Classifier)

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Build and train model
clf = DecisionTreeClassifier(max_depth=4, random_state=42)
clf.fit(X_train, y_train)

# Predictions
y_pred = clf.predict(X_test)

# Evaluation
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
print("Accuracy:", accuracy_score(y_test, y_pred))
```

8.4 Feature Importance

```
import matplotlib.pyplot as plt

# Plot feature importances
feat_importances = pd.Series(clf.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh', title='Top 10 Important Features')
plt.xlabel("Feature Importance Score")
plt.show()
```

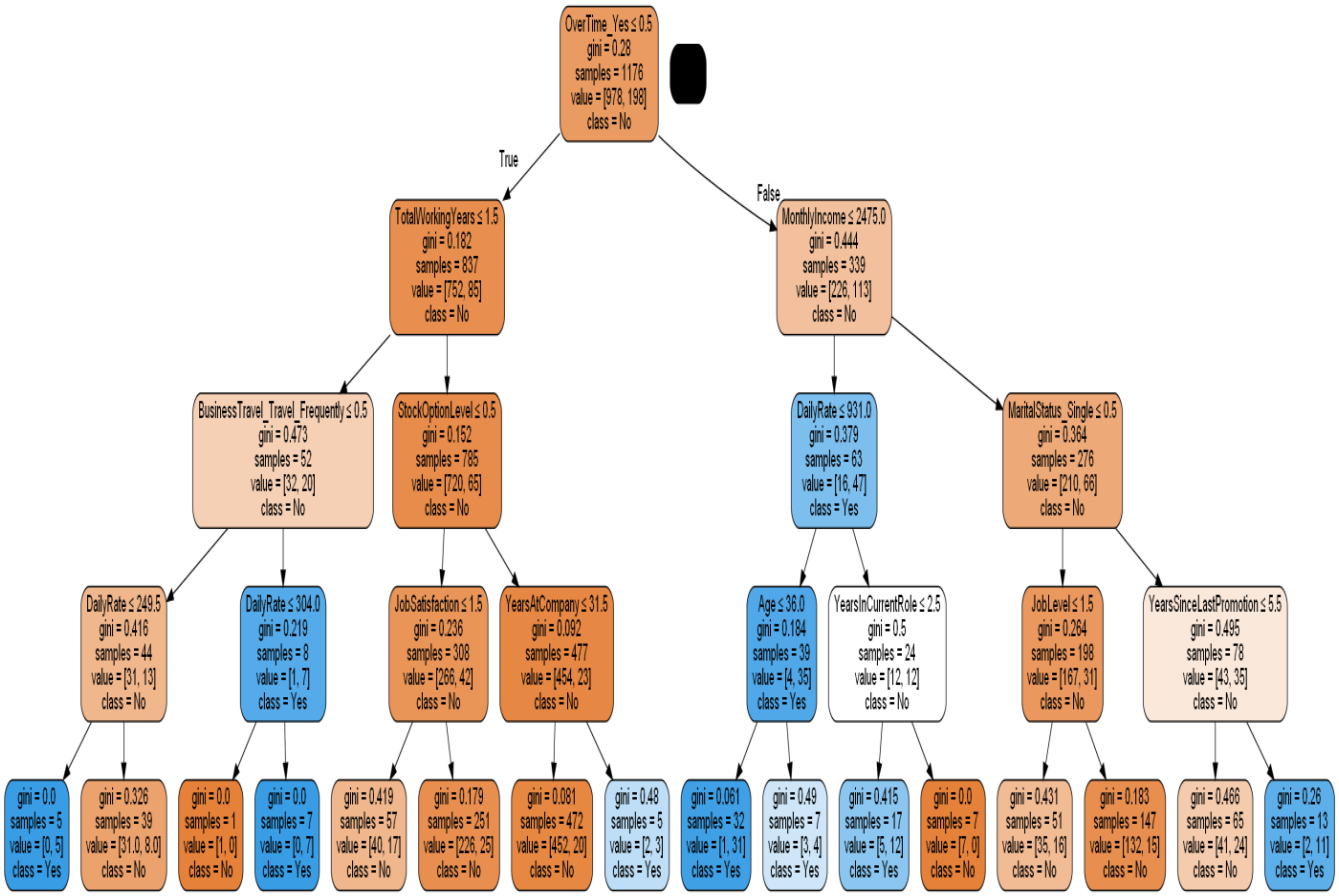
8.5 Decision Tree Visualization

```
from sklearn.tree import export_graphviz
import graphviz

# Export tree
dot_data = export_graphviz(clf, out_file=None,
                           feature_names=X.columns,
```

```
                class_names=["No", "Yes"],
                filled=True, rounded=True,
                special_characters=True)

# Visualize
graph = graphviz.Source(dot_data)
graph.render("decision_tree")
graph.view()
```



9. Conclusion

The analysis conducted in this project has successfully identified several key factors that influence employee attrition within an organization. Using exploratory data analysis (EDA), various trends and patterns were revealed, such as a higher attrition rate among employees with lower monthly income, those working overtime, and individuals with longer distances from home to the workplace. Categorical factors like job role, department, and education field also showed significant associations with attrition.

A Decision Tree model was developed to predict employee attrition, achieving moderate performance. While it effectively captured key decision splits, its ROC curve indicated scope for further optimization using advanced models.

Overall, this project demonstrates the importance of data-driven approaches in Human Resource Management. By identifying attrition triggers, organizations can adopt targeted retention strategies, improve employee engagement, and minimize the cost of turnover.

10. References

1. IBM HR Analytics Employee Attrition & Performance Dataset
   Kaggle Source
2. **Python Libraries Used:**
   - Pandas Documentation: https://pandas.pydata.org
   - Seaborn Documentation: https://seaborn.pydata.org
   - Matplotlib Documentation: https://matplotlib.org
   - Scikit-learn Documentation: https://scikit-learn.org
3. Graphviz & pydotplus:
   - https://graphviz.org/
   - https://pypi.org/project/pydotplus/
4. Machine Learning Concepts:
   - ROC Curve and Confusion Matrix: Scikit-learn official examples
5. SQL Integration with Python:
   - SQLite3: https://docs.python.org/3/library/sqlite3.html