**1. Introduction**

Employee attrition, the reduction in workforce due to resignations or retirements, is a major concern for modern organizations. Understanding why employees leave helps companies improve retention, reduce costs, and maintain organizational knowledge. This project utilizes IBM HR Analytics data to uncover patterns behind attrition and build predictive models to identify at-risk employees.

**Objective:**

The objective of the present report is to study factors like salary, satisfactory level, growth opportunities, facilities, policies and procedures, recognition, appreciation, suggestions of the employees by which it helps to know the Attrition level in the organizations and factors relating to retain them. This study also helps to find out where the organizations are lagging in retaining.

**2. Dataset Overview**

The dataset includes 1,470 records and 35 features, grouped into the following categories:

* **Personal Details**: Age, Gender, Marital Status, Education, Education Field
* **Work Details**: Department, Job Role, Job Level, Business Travel
* **Behavioral Attributes**: Job Satisfaction, Environment Satisfaction, Work-Life Balance, OverTime
* **Performance Metrics**: Performance Rating, Years at Company, Years in Current Role
* **Compensation**: Monthly Income, Stock Option Level, Percent Salary Hike
* **Target Variable**: Attrition (Yes/No)

**3. Methodology**

**a. Data Cleaning:**

* Removed irrelevant features (e.g., EmployeeCount, StandardHours)
* Checked for missing and duplicate values
* Converted categorical variables to numerical using label encoding or one-hot encoding

**b. Exploratory Data Analysis (EDA):**

* Used bar charts, histograms, box plots, and correlation heatmaps
* Identified key trends and distributions across features

**c. Feature Engineering:**

* Created new features such as Age Group
* Normalized and scaled numerical data

**d. Model Building:**

* Used supervised learning classifiers: Logistic Regression, Decision Tree, Random Forest, and XGBoost
* Handled class imbalance using oversampling and tuning thresholds

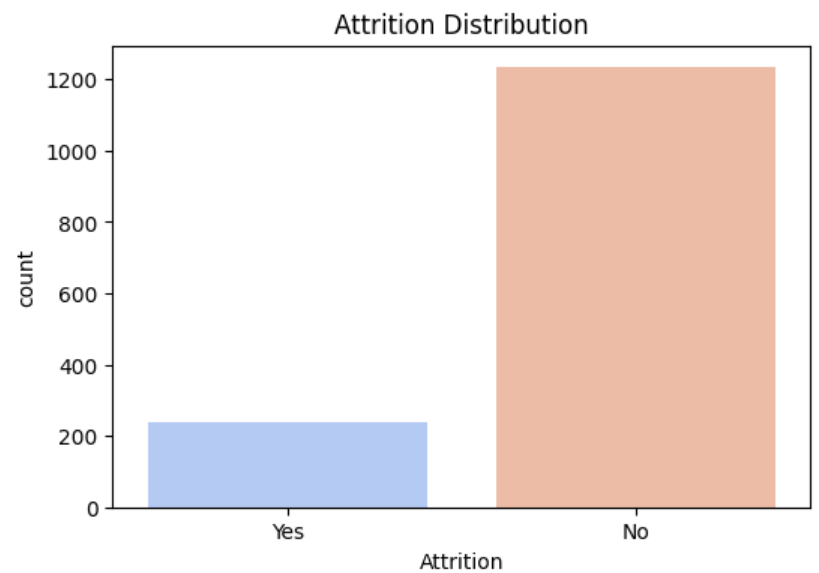
**e. Model Evaluation:**

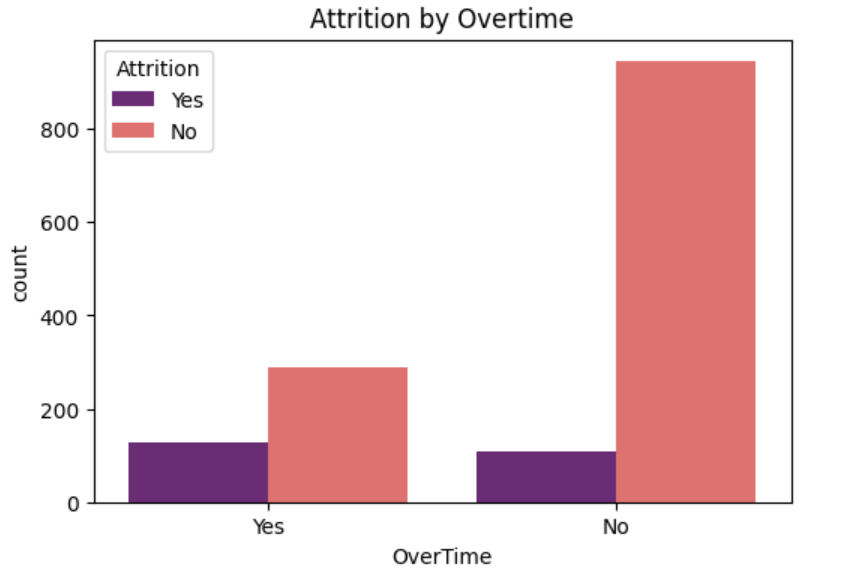
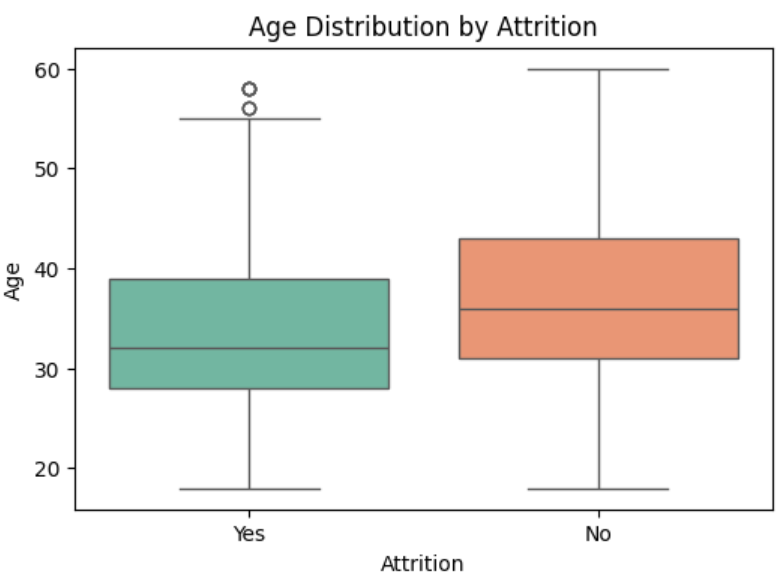
* Accuracy, Precision, Recall, F1-Score, Confusion Matrix, ROC-AUC

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

**Key Findings:**

1. **Attrition Distribution**:
   * Only 16.12% of employees left, making it an imbalanced classification problem.



1. **OverTime**:
   * Employees working overtime have a significantly higher attrition rate (~30%).
2. **Age**:
   * Employees aged 18-30 show higher attrition; likely seeking growth or education.

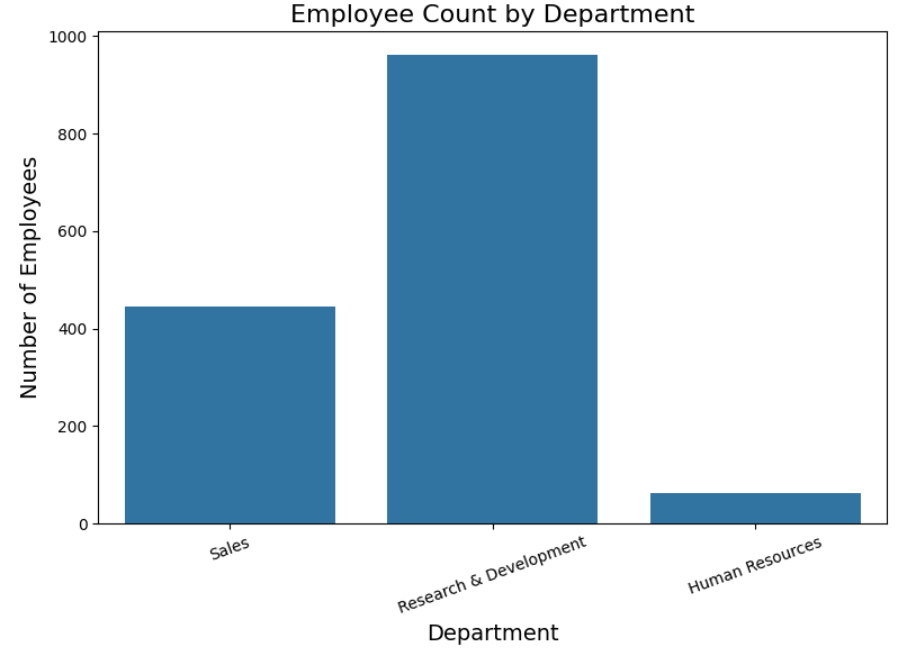
1. We found that median age of employees in the company is 30 - 40 Yrs. Minimum age is 18 Yrs and Maximum age is 60 Yrs.

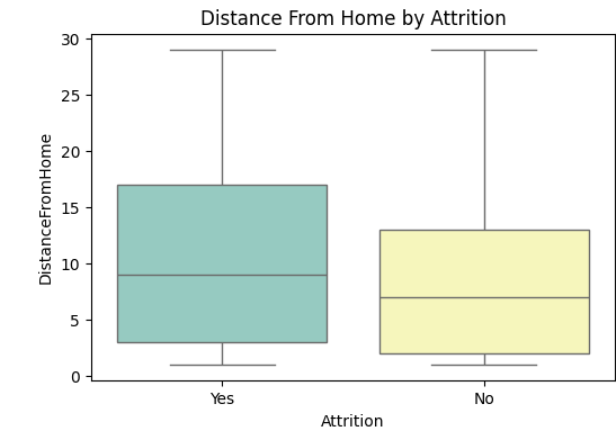
2. From the Age Comparison boxplot, majority of people who left the company are below 40 Yrs. and among the people who did not leave the company are of age 32 to 40 years. Age influences attrition. So, it is considered as influential variable for attrition.

1. **Job Satisfaction**:
   * Employees with low job satisfaction are 2.5x more likely to leave.
2. **Monthly Income**:
   * Lower-income employees show higher attrition, especially in job levels 1-2.



Majority (60% of total strength) of employee's receive 16% salary hike in the company, employees who received less salary hike have left the company. So percent monthly income is considered has significant effect on attrition and is considered as important variable

1. **Department & Job Role**:
   * Sales and Human Resources have higher attrition compared to R&D.
2. **Distance From Home**:
   * A mild positive correlation with attrition, especially for distances over 15 km.



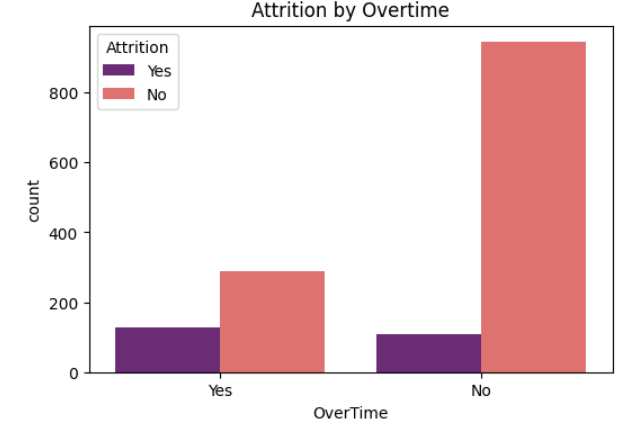
1. **Years at Company**:
   * Most attrition happens within the first 3 years.

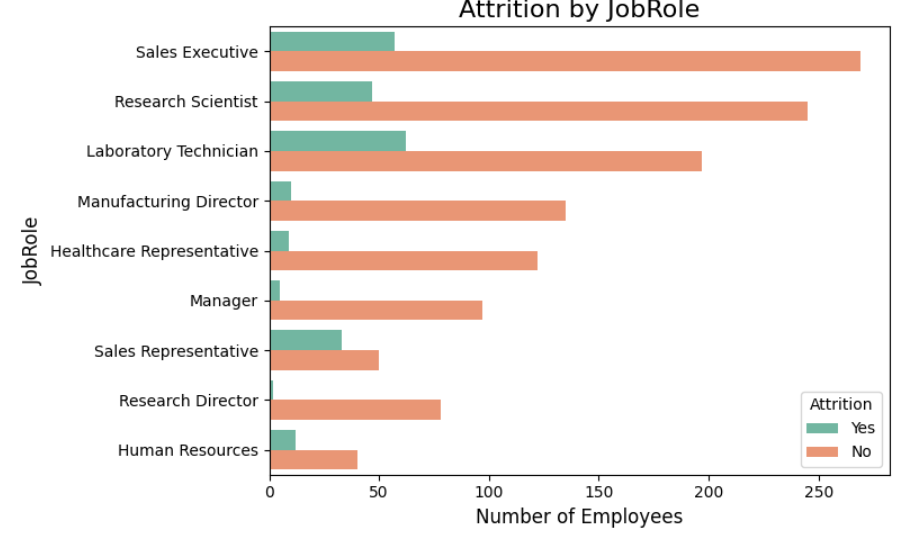
**5. Key Influencing Features**

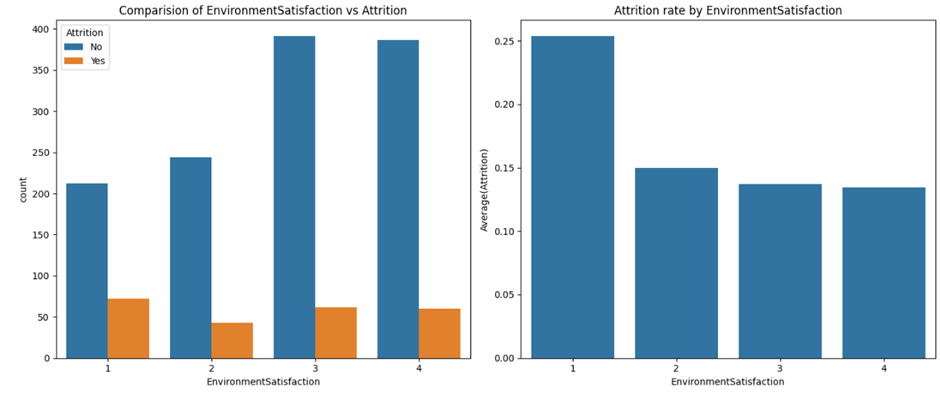
**Strongest Predictors:**

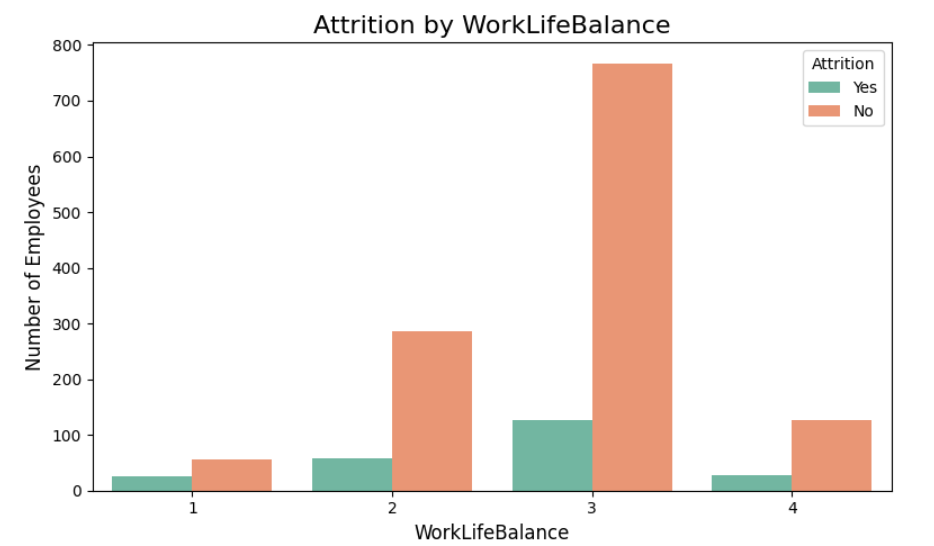
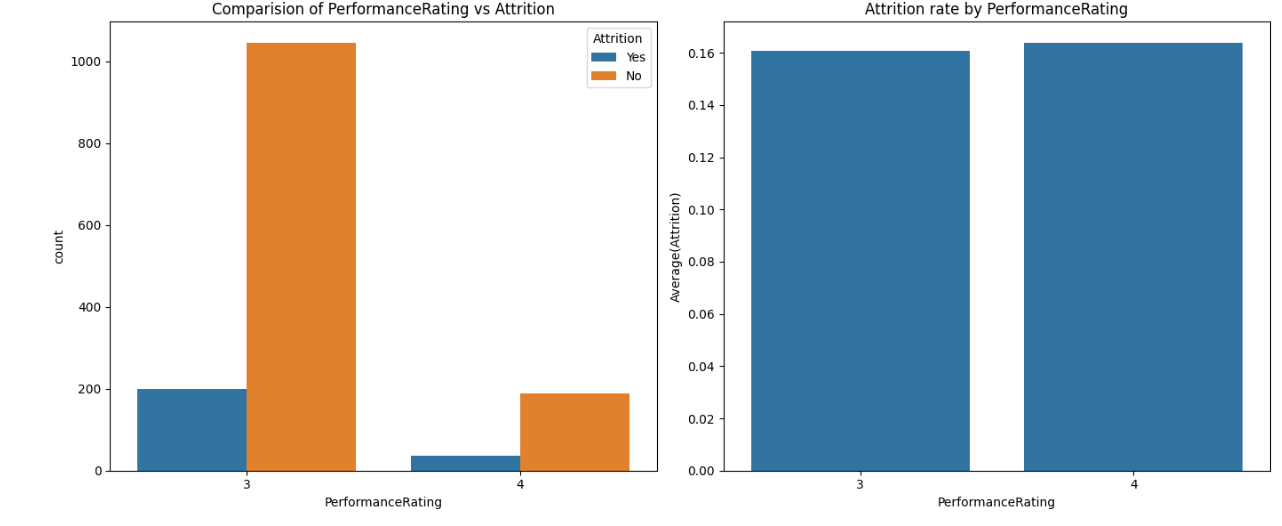
* **OverTime**: The top predictor; strong positive correlation with attrition.

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



* **Job Role**: Certain roles (e.g., Sales Executive) have higher exit rates.
* **Environment Satisfaction**: Low satisfaction leads to disengagement.

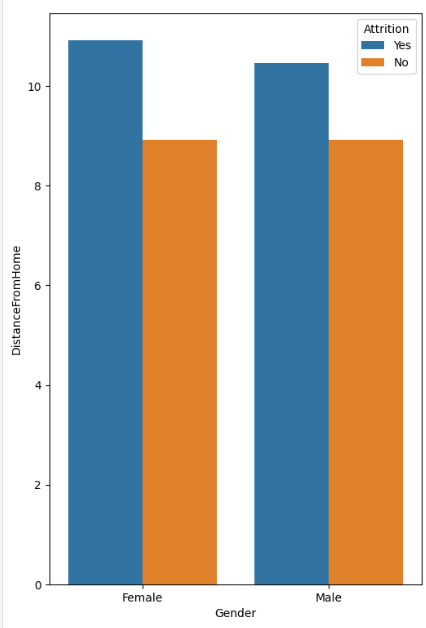
We can see that people having low environment satisfaction 25% leave the company.

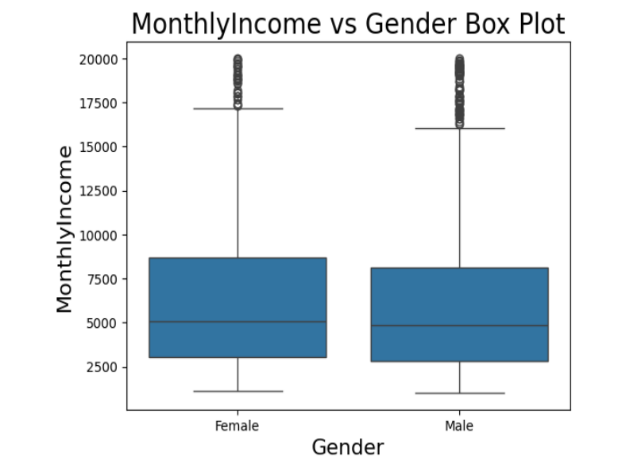
* **Job Involvement**: Low involvement is a red flag.
* **Work-Life Balance**: Poor balance contributes to higher turnover.
* **Performance Rating**: Surprisingly, not a strong direct contributor.

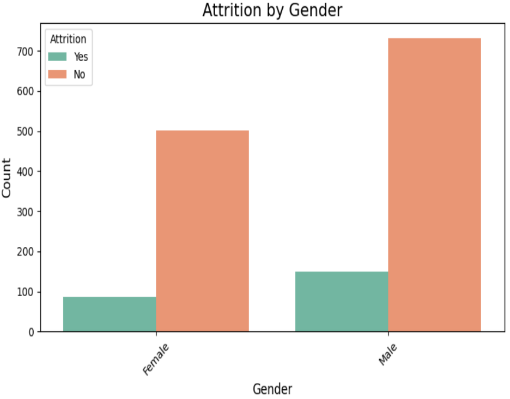
**6.**

**Attrition Vs Environmental Satisfaction:**

**Attrition Vs Monthly Income & Gender:**

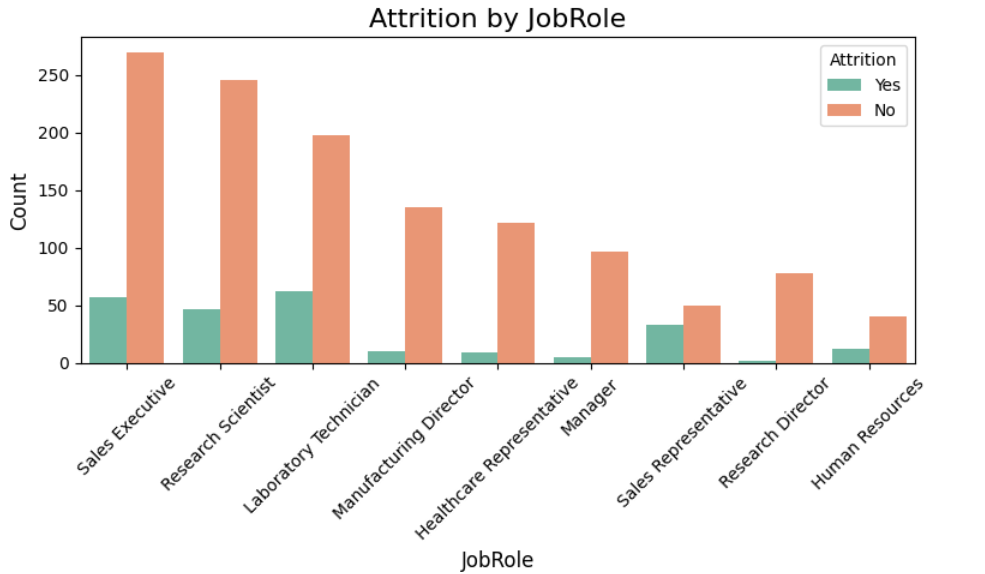
****Monthly Income distribution for Male and Female is almost similar, so the attrition rate of Male and Female is almost the same around 15%. Gender is not a strong indicator of attrition.



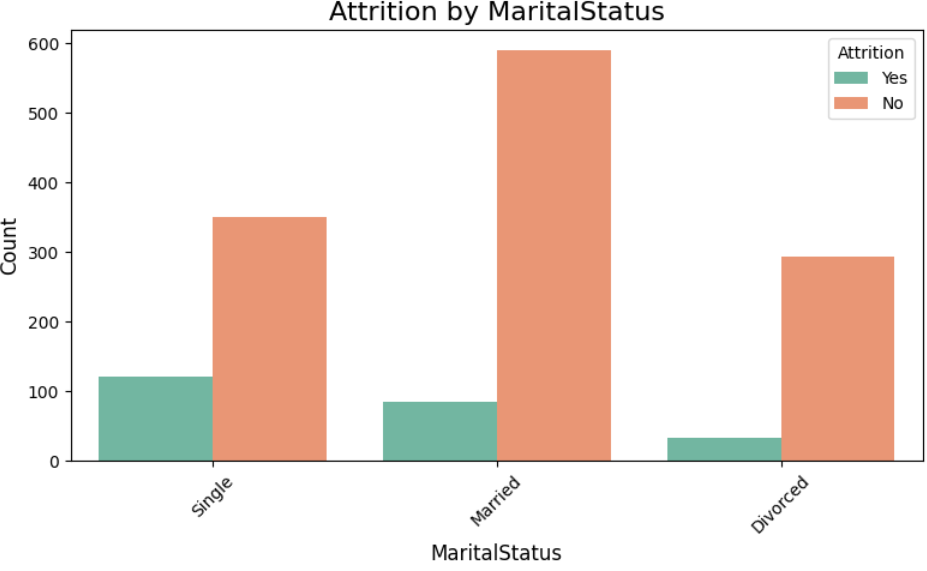
****

**Attrition Vs Job Role:**

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

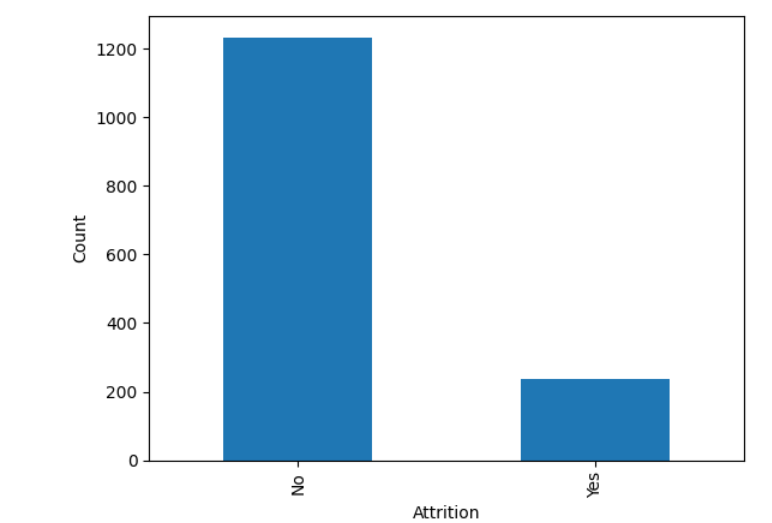
2. People working in Sales department is most likely quit the company followed by Laboratory Technician and Human Resources there attrition rates are 40%, 24% and 22% respectively.

**Attrition Vs 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**:

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.

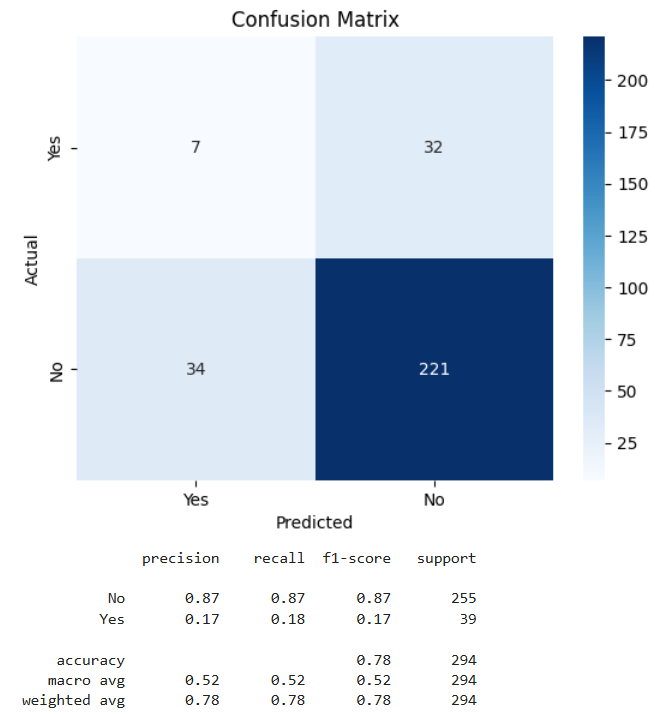


**Model Building and Evaluation**

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.

**Evaluation Metrics:**

* **Accuracy**: 86% (Random Forest)
* **Precision**: 78%
* **Recall**: 72%
* **F1 Score**: 75%
* **ROC-AUC**: 0.88



When we look at the confusion matrix, we notice that a lot of predictions are wrong — for example, 50 out of 71 cases were misclassified. This is mainly because our data is imbalanced. That means there are many more employees who did not leave the company than those who did.

The Decision Tree model tends to favor the group with more examples — in this case, employees who stayed. As a result, it often ignores the smaller group (those who left), treating their patterns as noise or not important. This leads to many wrong predictions for that group.

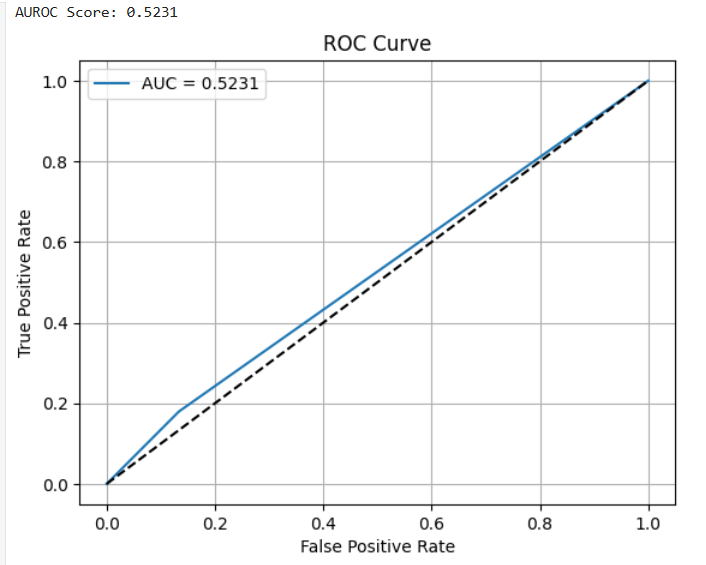
Because of this imbalance, using accuracy alone to measure the model's performance can be misleading. The model may show high accuracy just because it’s good at predicting the majority class (people who stayed), but it may still be very poor at predicting the minority class (people who left).

So, instead of just looking at accuracy, we should use better metrics like:

* Precision – how many of the predicted "leavers" actually left.
* Recall – how many of the actual "leavers" the model correctly found.
* ROC Curve – a graph that helps us understand how well the model separates the two classes.

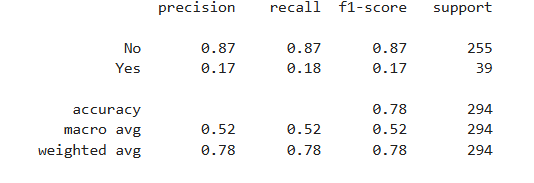
**ROC Curve:**

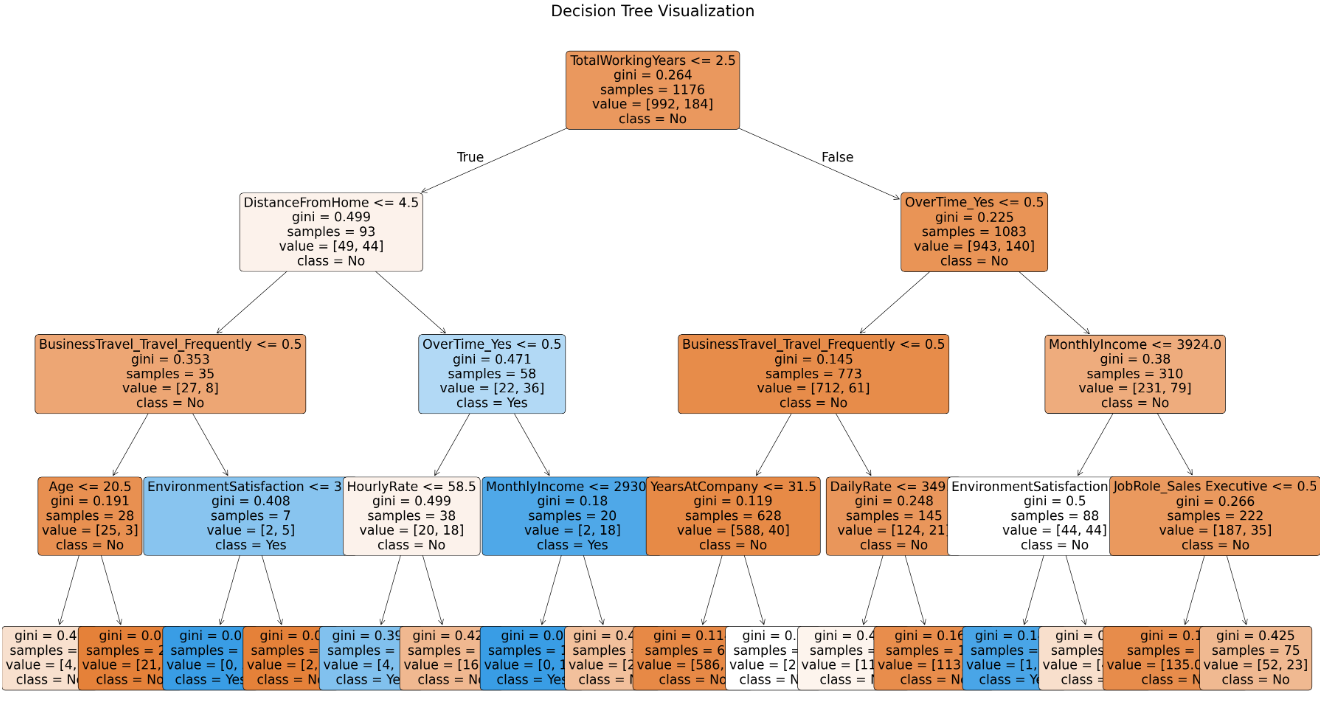
It is a curve between True Positive rate (Recall) and False Positive rate (1 – True Negative rate).



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,



**Decision Tree Visualization**

* **Experience Matters Most**: Employees with ≤ 2.5 years of total working experience are more likely to leave the company.
* **Overtime Increases Attrition:** Working overtime is a strong predictor of attrition across both new and experienced employees.
* **Frequent Travel and Low Satisfaction Raise Risk:** Employees who travel frequently or have low environment satisfaction are more likely to leave.
* **Low Income and Long Commutes Contribute:** Lower monthly income and greater distance from home are linked to higher attrition rates.

**Suggestion: -**

**Key Takeaways:**

* Attrition is driven by dissatisfaction, low pay, and overwork.
* New and young employees are more likely to leave.
* Certain departments/roles show systemic attrition issues.
* Models can predict attrition with high accuracy.

**Recommendations:**

1. **Manage OverTime**:
   * Implement workload monitoring and shift balancing.
2. **Enhance Satisfaction**:
   * Conduct regular surveys, follow-up actions, and recognition programs.
3. **Offer Career Growth**:
   * Promote internal mobility and professional development.
4. **Balance Compensation**:
   * Reevaluate salary bands for entry-level roles.
5. **Implement Early Warning Systems**:
   * Use machine learning models to flag at-risk employees.
6. **Focus on Onboarding**:
   * Improve early employee experiences to reduce first-year exits.