****

**A study on HR Analytics By Predicting Employee Attrition**

**Prepared by:  
Harish Rudraraju  
PGDM – Business Analytics**

**Institute of Public Enterprise  
Shamerpet,Hyderabad**

**Academic Year: 2025 – 2026**

**Objective:**

To understand the main causes of employee resignation and predict future attrition using historical data of employees.

**Abstract :**

This report analyses employee attrition using historical HR data of employees to identify the key factors that contribute to employees’ resignations and to develop a predictive model to forecast future attrition. The dataset contains various employee attributes, including demographics, job roles, compensation, performance ratings, satisfaction levels, and work-life balance indicators Etc.

Exploratory analysis Exposed the factors such as overtime, low job satisfaction, longer distance from home to workplace, fewer years since last promotion, and lower monthly income are strongly correlated with higher attrition rates. Employees in certain departments and job roles, mainly those under high work pressure or lacking growth opportunities, exhibited a higher likelihood of leaving the organization.

To support proactive retention strategies, a machine learning model was developed using features such as job satisfaction, overtime, years at company, and total working years. The model achieved satisfactory predictive performance, allowing HR to identify high-risk employees and implement targeted interventions.

**Dataset overview:**

The dataset which contains information on 1,470 employees, with each row representing an individual employee’s profile. It also includes a total of 35 features covering:

Demographics: Age, Gender, Marital Status, Distance from Home

Job Details: Department, Job Role, Job Level, Years at Company, Years In Current Role

Compensation: Monthly Income, Hourly Rate, Daily Rate, Monthly Rate, Stock Option Level

Performance & Engagement: Performance Rating, Job Involvement, Job Satisfaction, Work Life Balance

Satisfaction & Work Conditions: Environment Satisfaction, Relationship Satisfaction, Overtime, Business Travel

Training & Experience: Education, Education Field, Num of Companies Worked, Total Working Years, Training Times Last Year

Target Variable: Attrition (Yes/No) — Indicates if the employee has left the company

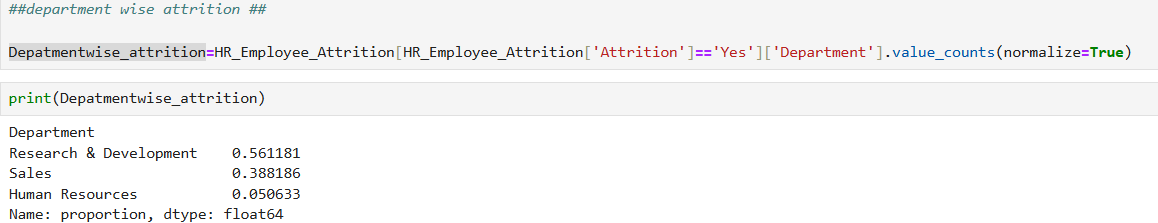
Some columns like Employee Count, Standard Hours, and Over18 are constant and may be dropped during cleaning as they do not contribute to analysis.

**Exploratory Data Analysis:**

Exploratory Data Analysis is a process of examining and summarizing a dataset to understand its main characteristics, often using visual methods. It is typically the first step in any data analysis or machine learning project and helps analysts or data scientists uncover patterns, spot anomalies, test assumptions, and check data quality before applying more formal modelling techniques.

Below are some insights that I have captured from the code:

a) Department Wise Attrition:



1)Research & Development (R&D) has the highest proportion of employees who left (over half of total attritions).

2)Sales shows a significant share of attritions.

3)Human Resources accounts for a very small proportion of attritions.

B) Education Wise Attrition:

A computer screen shot of a computer

AI-generated content may be incorrect.

1)Life Sciences (37.6%): Highest attrition share; likely due to high R&D involvement.

2)Medical (26.6%): Second highest; may indicate mismatch or stress in roles.

3)Marketing & Technical Degrees (~28%): Moderate attrition; performance or growth-related issues possible.

4)Human Resources (3%): Lowest attrition; suggests high job satisfaction and stability.

c)Age Wise Attrition:



1)Highest attrition is among 30–40 and 20–30 age groups (combined ~72%).

2)Attrition drops significantly after age 40.

3)No attrition in the 60+ group (possibly due to retirement or low headcount).

D)Job Role Wise Attrition:



1)Lab Technicians and Sales Executives together account for half of all attritions.

2)Research-based roles like Research Scientist also show significant attrition.

3)Leaders (Managers, Directors) show very low attrition, likely due to job stability and satisfaction.

E)Bar plot for Attrition rate based on Department:



1)Research & Development has the highest attrition rate (~56%).

2)Sales follows with ~39%.

3)Human Resources has a very low attrition rate (~5%).

F) Bar plot for Attrition rate based on Job Role:

A screenshot of a computer

AI-generated content may be incorrect.

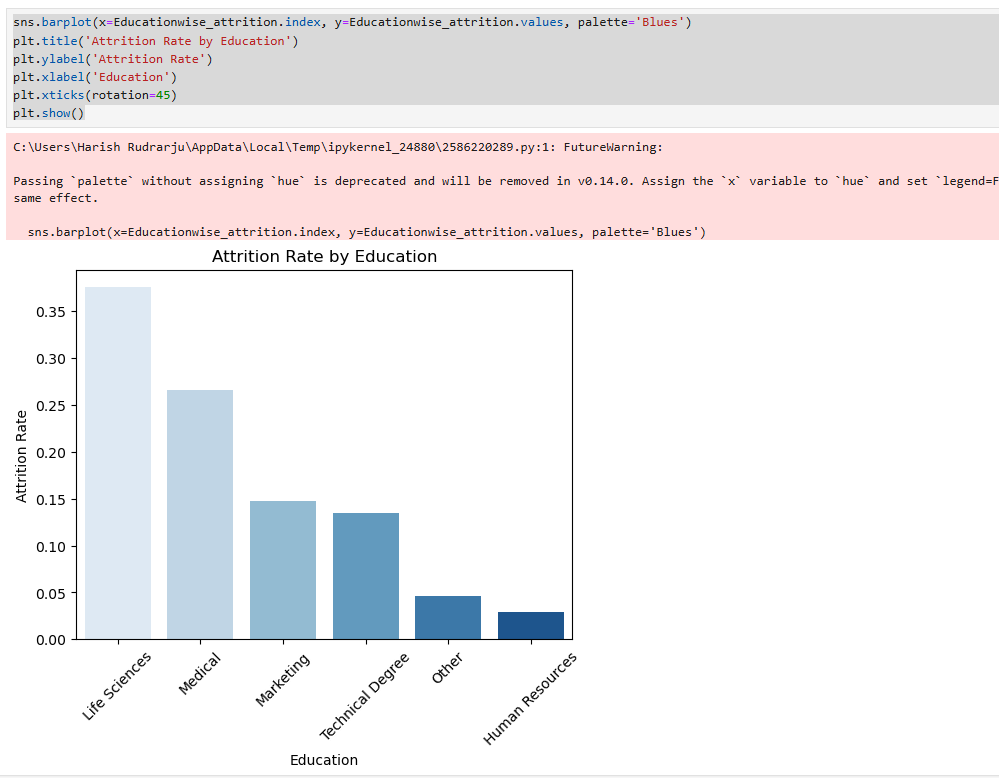
1) Highest Attrition:

* Laboratory Technician (~26%)
* Sales Executive (~24%)
* Research Scientist (~20%)

2) Lowest Attrition:

* Research Director (~1%)
* Manager (~2%)
* Healthcare Rep (~3.7%)

G) Bar plot for Attrition rate based on Education:



1) Highest Attrition:

* Life Sciences (~38%)
* Medical (~27%)

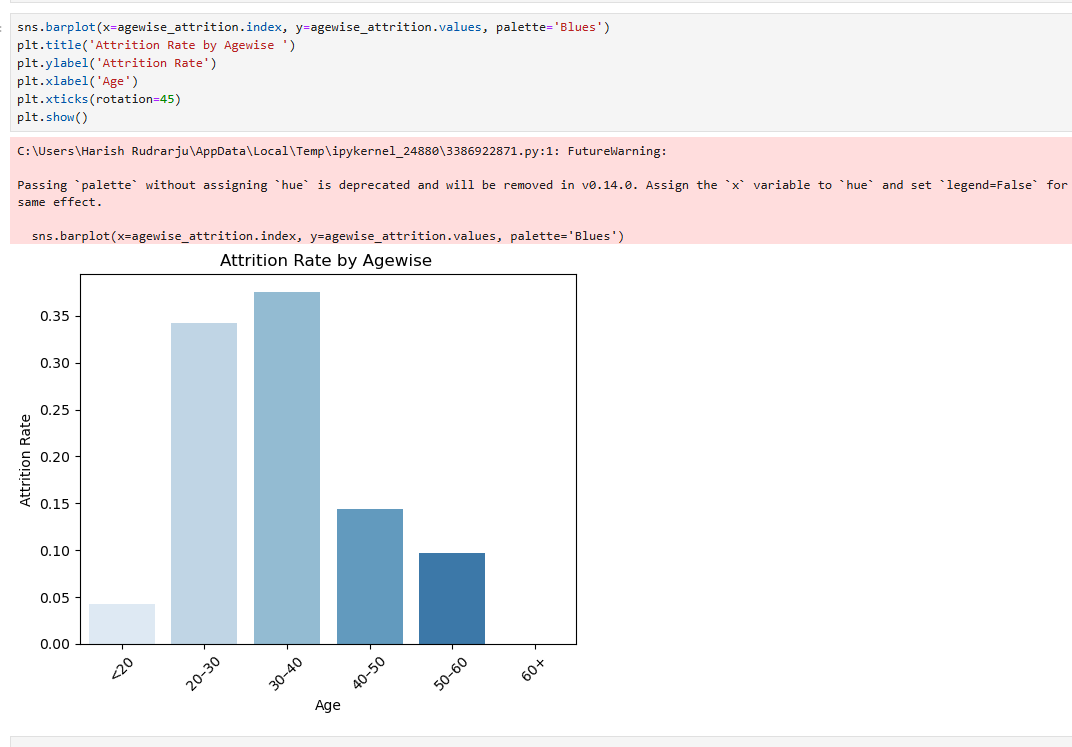
2) Moderate Attrition:

* Marketing (~15%)
* Technical Degree (~13%)

3) Lowest Attrition:

* Other (~5%)
* Human Resources (~3%)

H) Bar plot for Attrition rate based on Age:



1) Highest Attrition:

* 30–40 age group (~37%)
* 20–30 age group (~34%)

2) Moderate Attrition:

* 40–50 age group (~14%)
* 50–60 age group (~9%)

3) Lowest Attrition:

* <20 age group (~4%)
* 60+ age group (0%)

1. corelation between Attrition and Jobsatisfaction:

A screenshot of a computer code

AI-generated content may be incorrect.

* The correlation between AttritionNumeric and JobSatisfaction is -0.103.
* This is a weak negative correlation, indicating that as job satisfaction increases, attrition slightly decreases (employees are a bit less likely to leave).

J) corelation between Attrition and DistanceFromHome:

A screenshot of a computer code

AI-generated content may be incorrect.

* The correlation between AttritionNumeric and DistanceFromHome is **0.0779**.
* This is a very weak positive correlation, which means:
* Employees who live farther from home may be *slightly* more likely to leave.
* But the correlation is too small to be practically significant.

K) corelation between Attrition and PercentSalaryHike:

A screenshot of a computer

AI-generated content may be incorrect.

* Correlation Coefficient: -0.0135
* This is extremely close to 0, meaning:
* There is virtually no linear relationship between percent salary hike and attrition.
* A higher or lower salary hike has negligible impact on whether employees leave.

L) corelation between Attrition and WorkLifeBalance:

A screenshot of a computer code

AI-generated content may be incorrect.

* Correlation coefficient = -0.064
* This is a very weak negative correlation.
* Suggests that better work-life balance is *slightly* associated with lower attrition.
* However, the effect is very minimal — likely not statistically significant on its own.

**Suggestions on Exploratory Data Analysis:**

1. By Department:

* Investigate **workload, stress, or career growth** issues in R&D.
* Offer **incentives, career coaching, or flexible targets** in Sales.
* Learn from HR’s approach – apply similar **employee engagement strategies** elsewhere.

1. By Education:

* Provide **tailored career paths or mentoring** for Life Sciences and Medical graduates.
* Ensure roles **match qualifications and expectations.**
* Analyse what makes HR roles more stable and replicate best practices.

C) By Age Group:

* Focus retention strategies (e.g., development, recognition) on younger professionals.
* Offer career planning support and work-life balance programs.
* Conduct exit interviews to understand why early- to mid-career employees leave.

1. By Job Role:

* Assess **work environment, pay, and career progression** for Lab Technicians and Sales Executives.
* Explore **job redesign or cross-training** to increase satisfaction.
* Leverage **leadership insights** from low-attrition roles to guide improvements in high-risk ones.

1. Improve Job Satisfaction:

* Conduct regular feedback surveys to understand dissatisfaction.
* Recognize and reward employee achievements.
* Provide meaningful roles and career growth opportunities.

1. Enhance Work-Life Balance:

* Offer flexible working hours or remote work options.
* Promote wellness programs (mental health days, ergonomic workspace.
* Avoid excessive overtime and foster a healthy work culture.

1. . Address Commute Issues (DistanceFromHome):

* Consider hybrid or remote roles for employees living far away.
* Offer transport facilities or commute allowances.
* Relocate teams based on geographic clusters if possible

1. Rethink Salary Hike Strategies:

* Instead of only increasing hikes, **align raises with performance and growth**.
* Provide **transparent promotion paths** and upskilling opportunities.
* Use salary as part of a broader **employee value proposition** (culture, purpose, benefits).

**Machine Learning Model Using Python:**

Machine Learning is a branch of artificial intelligence that focuses on developing algorithms that allow computers to learn from and to make predictions or decisions based on given data.

A machine learning model is a mathematical representation that is trained on data set to recognize patterns or relationships. It maps input data i.e. independent features to desired outputs (labels or predictions) i.e. dependent features and can be used to make decisions or predictions on new and unseen data.

There are Three Types of Machine Learning Model

* 1. Supervised Learning Model
  2. Unsupervised Learning Model
  3. Reinforcement Learning Model

**A) Supervised Learning Model :**

* Supervised Learning model is trained on labelled data (input-output pairs).
* Examples: Linear Regression, Decision Trees, Support Vector Machines.

**B) Unsupervised Learning Model:**

* Unsupervised Learning model identifies patterns in data without labelled outputs.
* Examples: K-Means Clustering

**D)Reinforcement Learning Model:**

 Reinforcement Learning Model learns by interacting with an environment and receiving rewards or penalties.

 Example: Q-Learning, Deep Q-Networks.

**Model Applied Based On DataSet:**

For this project, I Was applied a Logistic Regression, a supervised learning algorithm, to the dataset.

Logistic Regression is applied where the target feature is in chatogorical features which is either YES or NO.

**Steps Involved In this Model:**

**1)Data Preprocessing:**

In this I was done with data preprocessing by cleaning the data which was unwanted for prediction available in data set.

**2) Model Training**:

In This I Used a Model that is logistic regression algorithm to train on the labelled dataset.

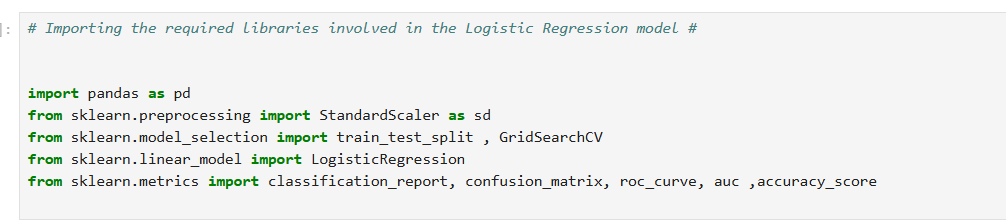
**3)Accuracy of Model:**

Once Model was created the accuracy was checked by using Confusion Matrix, Classification Report, Accuracy Score

**4) Improvement of Accuracy of Model**:

To Improve Accuracy of Model have done cross validation and Hyperparameter Tunning using GridSearchCv().

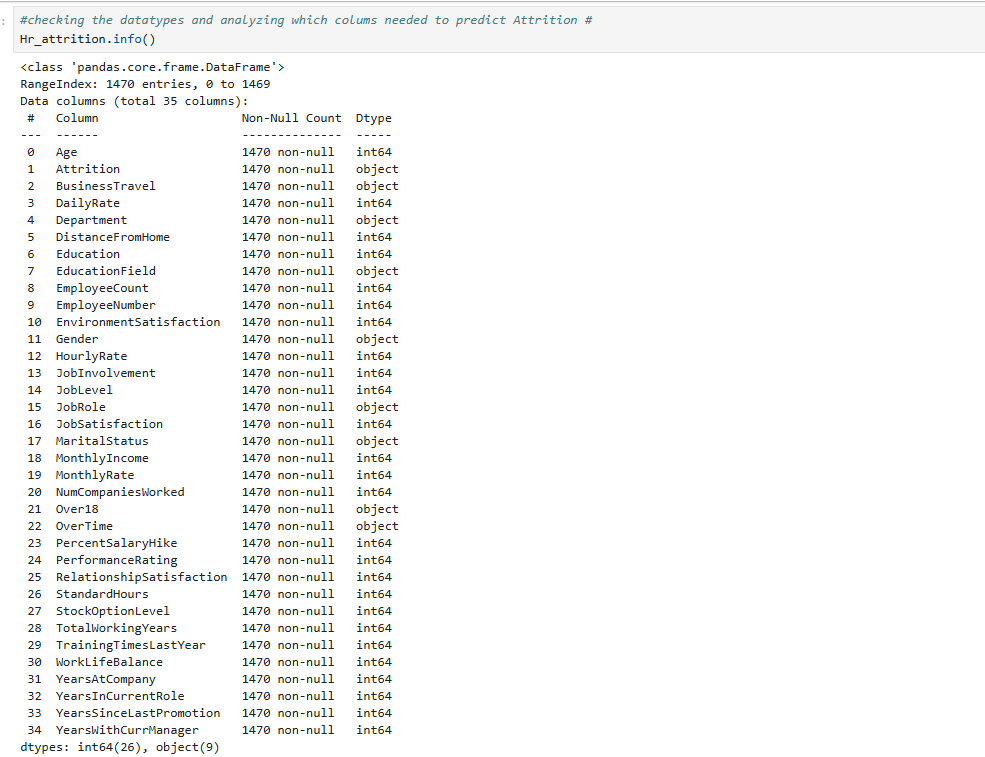
**Machine Learning Model Explanation Using Python In Jupiter Note Book Environment:**

****

In This I Imported All Required Libraries required for the model to Build a Logistic Regression.



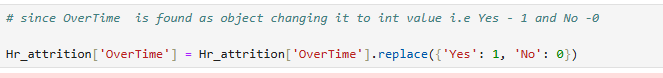
Once All the Required Libraries are imported I Loaded the Historical Dataset of Hr\_Attrition into environment.



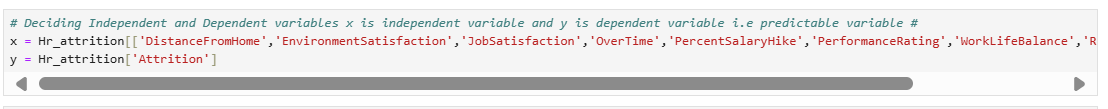
I Checked the all the attributes in the data set to check Datatype and also to decide the independent variable and dependent variable to develop the model.



I was done with data cleaning by dropping the unwanted columns and as I checked the duplicates values and empty values in EDA I did not went through since there were no issues with them.



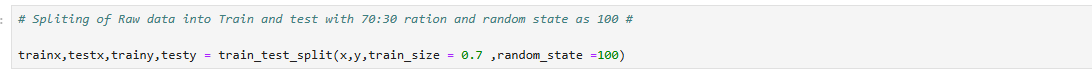
As I decided the independent variables in data set I observed that overtime is in categorical feature since the independent variables should be in numeric I converted into numeric.



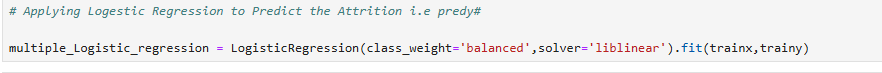
Now I decide the Dependent feature as Attrition and Remaining independent features.



I applied Standardsscaller to make all the variables in one format and fitted.



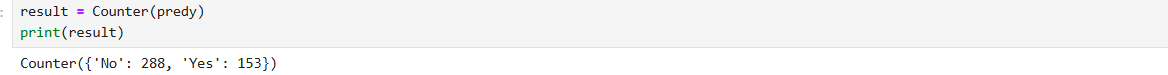
Here the data was splitted into train and test that is to develop the model and evaluate the model. Here I applied 70:30 in ratio.



Applied Logistic Regression once train and test were splitted to predict the future value of Attrition.



In Above result Yes indicates that predicted attrition i.e the employee may leave and NO indicates retention i.e the employee is predicted to stay.



This would mean 288 employees are predicted to stay, and 153 are predicted to leave.

A screenshot of a computer

AI-generated content may be incorrect.

|  | **Predicted No** | **Predicted Yes** |
| --- | --- | --- |

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual No** | 261 (True Neg) |  | 110 (False Pos) |

|  |  |  |
| --- | --- | --- |
| **Actual Yes** | 27 (False Neg) | 43 (True Pos) |
|  |  |  |
| **Accuracy:**  **TP+TN/Total = 261 + 43 /441 = 0.689 89** |  |  |

From The three I can conclude that the Accuracy was approx. 69 percentage , so I order to improve the accuracy of the model I used some techniques.

A computer code with text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

From the above results I can say the accuracy of model was improved to 70 percentage which was decent.

**Observation:**

**A)Predicted Results**:

* The majority of predictions were 'NO', indicating low attrition, both **before** and **after** tuning.
* However, **after tuning**, a few additional 'Yes' predictions were made—capturing more potential attrition cases.

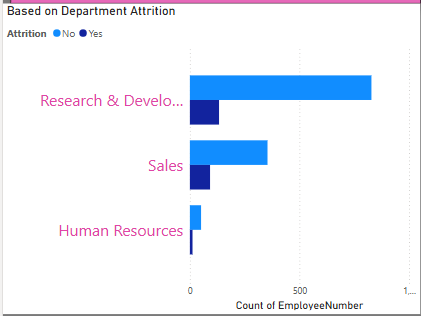
**B)Model Improvement**:

* Hyperparameter tuning with GridSearchCV resulted in a **1% accuracy improvement**.
* While this numerical gain is small, the model's **sensitivity to "Yes" class (attrition cases)** improved slightly, which is valuable in HR contexts.

**Conclusion:**

Hence, The Logistic Regression model provides a reasonable baseline for predicting employee attrition, with overall accuracy of 70% after tuning. While most predictions indicate employees are likely to stay, the tuned model improves the detection of those at risk of leaving.

**Visualize attrition factors using Power BI :**

****

**Interpretation:**

* Research & Development is the most affected numerically, but not necessarily by percentage.
* Sales department has relatively higher attrition rate, considering its total size.
* Human Resources has minimal attrition, indicating stability.

A blue line with black text

AI-generated content may be incorrect.

**Interpretation:**

* Employees who left (Attrition = Yes) generally earn less in total monthly income.
* Indicates a strong correlation between low income and higher attrition.

A graph with blue bars

AI-generated content may be incorrect.

**Interpretation:**

* The 25–35 age group represents the biggest retention risk.
* Attrition tends to decline with age, suggesting stability increases over time.

**A graph of blue rectangular bars

AI-generated content may be incorrect.**

**Interpretation:**

1)Employees with lower job satisfaction (levels 1 & 2) are more likely to leave.

2)As satisfaction improves, attrition decreases noticeably

A graph of blue rectangular bars

AI-generated content may be incorrect.

**Interpretation:**

* There is a negative correlation between Environment Satisfaction and Attrition.
* As environment satisfaction increases, employee attrition decreases.

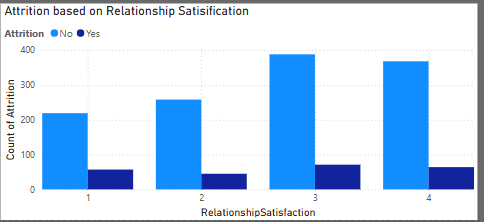
A graph of blue rectangular bars

AI-generated content may be incorrect.

**Interpretation:**

There is a clear trend showing that better work-life balance is associated with lower attrition.

* Employees with poor work-life balance (levels 1 & 2) are more likely to leave, while those with a balanced life (especially level 3 or 4) tend to stay longer.



**Interpretation:**

* There is a negative relationship between relationship satisfaction and attrition.
* As relationship satisfaction increases, the likelihood of attrition decreases.
* Fostering better interpersonal relationships within the organization can help enhance employee retention.

**Conclusion :**

Employee attrition is multi factorial, driven largely by job satisfaction, compensation, age, and work-life dynamics.  
To reduce attrition and improve retention, organizations should focus on:

* Enhancing employee satisfaction
* Offering competitive pay
* Creating a supportive work environment
* Balancing workloads
* Encouraging meaningful workplace relationships