Project On

IBM HR Analytics – Employee Attrition & Performance

By

Durga Prasad Govindas

UMID05102560279

UNIFIED MENTOR PRIVATE LIMITED

During The Period

October 2025 - December 2025

Tools: Python, Pandas, Matplotlib, Seaborn, Scikit-learn, Power BI

Section-1: Executive Summary

This project, **IBM HR Analytics – Employee Attrition & Performance**, focuses on identifying the key factors influencing employee turnover through data-driven analysis. Using detailed HR data on demographics, job satisfaction, compensation, and work-life balance, the study explores patterns that lead to attrition and develops predictive insights to support workforce stability.

The analysis revealed that **workload, salary disparity, job satisfaction, and tenure** are major contributors to **employee attrition**. Employees experiencing frequent overtime, lower income, or reduced satisfaction are more likely to leave.

By combining analytics and machine learning, the project provides actionable insights that can guide HR teams in designing effective retention strategies and promoting a healthier, more engaged workforce.

Section-2: Problem Statement

High employee attrition increases recruitment and training costs while reducing organizational stability. The key problem is to understand **why** employees leave and to **predict** who is likely to leave.

This project aims to:

- Identify the major factors influencing attrition.
- Build a predictive model to classify at-risk employees.
- Provide actionable insights to improve retention.

By applying data-driven analysis to HR records, the project supports strategic workforce planning and fosters long-term employee engagement.

Section-3: Objectives

The primary goal of this project is to analyze employee data to understand and predict attrition within the organization. By using data analytics and machine learning techniques, the project seeks to transform HR data into actionable insights that support better decision-making.

Specific Objectives

- Analyze employee demographics and job characteristics to identify patterns linked to attrition.
- Determine key factors such as job satisfaction, overtime, and income that influence employee turnover.
- Develop and evaluate predictive models to classify employees likely to leave.
- Provide data-driven recommendations to help HR implement targeted retention strategies.

Section:4 - Data Description

The dataset used in this project is the **IBM HR Analytics Employee Attrition & Performance** dataset, which contains detailed information about employee demographics, job roles, compensation, satisfaction levels, and performance. The data consists of **1,470 records and 35 variables,** with the target variable **Attrition** indicating whether an employee has left the company **("Yes") or stayed ("No")**.

4.1 Variable Overview

Feature Name	Description	Data Type / Levels
Age	Employee's age (years)	Numeric
Attrition	Whether the employee left the company	Yes / No (Target Variable)
BusinessTravel	Frequency of work-related travel	Non-Travel, Travel_Rarely, Travel_Frequently
DailyRate	Employee's daily salary rate	Numeric
Department	Department of work	Sales, Research & Development, Human Resources
DistanceFromHome	Distance from home to workplace (km)	Numeric
Education	Highest education level	1 = Below College, 2 = College, 3 = Bachelor, 4 = Master, 5 = Doctor
EducationField	Field of education	Life Sciences, Medical, Marketing, Technical Degree, Human Resources, Other
EmployeeCount	Number of employees (constant = 1)	Dropped – static value
EmployeeNumber	Unique ID for each employee	Dropped – identifier only
EnvironmentSatisfaction	Satisfaction with work environment	1 = Low, 2 = Medium, 3 = High, 4 = Very High
Gender	Employee gender	Male, Female
HourlyRate	Hourly pay rate	Numeric
JobInvolvement	Level of involvement in job	1 = Low, 2 = Medium, 3 = High, 4 = Very High
JobLevel	Job level within the organization	Numeric (1–5)
JobRole	Employee's job title	e.g., Sales Executive, Research Scientist, Laboratory Technician
JobSatisfaction	Satisfaction with the job	1 = Low, 2 = Medium, 3 = High, 4 = Very High
MaritalStatus	Marital status	Single, Married, Divorced
MonthlyIncome	Monthly salary	Numeric
MonthlyRate	Monthly payment rate	Numeric
NumCompaniesWorked	Number of previous companies worked	Numeric
Over18	Whether the employee is above 18 years old (constant = "Y")	Dropped – static value

OverTime	Whether employee works overtime	Yes, No
PercentSalaryHike	Percent increase in salary compared to last year	Numeric
PerformanceRating	Employee performance evaluation	1 = Low, 2 = Good, 3 = Excellent, 4 = Outstanding
RelationshipSatisfaction	Relationship satisfaction with colleagues/managers	1 = Low, 2 = Medium, 3 = High, 4 = Very High
StandardHours	Standard working hours (constant = 80)	Dropped – static value
StockOptionLevel	Level of stock option granted	0–3
TotalWorkingYears	Total years of professional experience	Numeric
TrainingTimesLastYear	Number of trainings attended last year	Numeric
WorkLifeBalance	Work-life balance perception	1 = Bad, 2 = Good, 3 = Better, 4 = Best
YearsAtCompany	Years spent at the current company	Numeric
YearsInCurrentRole	Years spent in the current role	Numeric
YearsSinceLastPromotion	Years since last promotion	Numeric
YearsWithCurrManager	Years spent with current manager	Numeric

4.2 Featurs Excluded from Analysis

The following columns were removed as they **do not provide meaningful variation or have static values** across all records:

- EmployeeCount
- EmployeeNumber
- Over18
- StandardHours

These columns do not influence attrition outcomes and are excluded from further analysis to maintain model efficiency and interpretability.

4.3 Importance of Variables for Analysis

Several variables hold strong analytical significance in understanding employee attrition:

- OverTime: Strong indicator of workload stress; employees with overtime show higher attrition rates.
- JobSatisfaction & EnvironmentSatisfaction: Directly impact employee morale and retention.
- MonthlyIncome: Reflects compensation fairness and is inversely correlated with attrition.
- WorkLifeBalance: Highlights the employee's equilibrium between personal and professional life.
- YearsAtCompany & YearsSinceLastPromotion: Indicate career growth and stability; shorter tenures and long gaps between promotions are linked to higher turnover.

4.4 Data Quality Summary

- Total Records: 1,470
- Total Variables: 35
- Categorical Variables: 9
- Numeric Variables: 26
- Missing Values: 0
- Duplicates: 0
- Target Class Imbalance: Attrition "Yes" = 16.1%, "No" = 83.9%

Section:5 - Exploratory Data Analysis (EDA)

5.1 Descriptive Statistics

To understand the overall characteristics of the dataset, descriptive statistical measures such as mean, standard deviation, minimum, maximum, and quartile values were computed for all numerical variables.

Metric	Description
Count	Number of valid observations for each variable (all 1,470 — no missing values)
Mean	Average value representing the central tendency of the variable
Standard Deviation (std)	Measures the spread or variability of the data
Min / Max	Lowest and highest observed values
25%, 50%, 75%	Quartile values indicating data distribution spread

The table below summarizes the key descriptive statistics of the dataset:

Variable	count	mean	std	min	0.25	0.5	0.75	max
Age	1470	36.92381	9.135373	18	30	36	43	60
DailyRate	1470	802.485714	403.5091	102	465	802	1157	1499
DistanceFromHome	1470	9.192517	8.106864	1	2	7	14	29
Education	1470	2.912925	1.024165	1	2	3	4	5
EnvironmentSatisfaction	1470	2.721769	1.093082	1	2	3	4	4
HourlyRate	1470	65.891156	20.329428	30	48	66	83.75	100
JobInvolvement	1470	2.729932	0.711561	1	2	3	3	4
JobLevel	1470	2.063946	1.10694	1	1	2	3	5
JobSatisfaction	1470	2.728571	1.102846	1	2	3	4	4
MonthlyIncome	1470	6502.931293	4707.956783	1009	2911	4919	8379	19999
MonthlyRate	1470	14313.1034	7117.786044	2094	8047	14235.5	20461.5	26999
NumCompaniesWorked	1470	2.693197	2.498009	0	1	2	4	9
PercentSalaryHike	1470	15.209524	3.659938	11	12	14	18	25
PerformanceRating	1470	3.153741	0.360824	3	3	3	3	4
RelationshipSatisfaction	1470	2.712245	1.081209	1	2	3	4	4
StockOptionLevel	1470	0.793878	0.852077	0	0	1	1	3
TotalWorkingYears	1470	11.279592	7.780782	0	6	10	15	40
TrainingTimesLastYear	1470	2.79932	1.289271	0	2	3	3	6
WorkLifeBalance	1470	2.761224	0.706476	1	2	3	3	4
YearsAtCompany	1470	7.008163	6.126525	0	3	5	9	40
YearsInCurrentRole	1470	4.229252	3.623137	0	2	3	7	18
YearsSinceLastPromotion	1470	2.187755	3.22243	0	0	1	3	15
YearsWithCurrManager	1470	4.123129	3.568136	0	2	3	7	17

5.2 Descriptive Analysis and Key Observations

The dataset represents 1,470 employees and includes demographic, satisfaction, performance, and compensation-related variables.

The descriptive summary highlights important patterns and distributions, as discussed below:

1. Demographic Insights

- The **average age** of employees is **37 years**, ranging from **18 to 60**. This indicates **a mid-career workforce** with a mix of younger and experienced employees.
- The **Distance From Home** averages **9 km**, but with a high standard deviation (8.1), showing that some employees commute significantly longer distances. Commuting distance can affect job satisfaction and may contribute to turnover.

2. Career and Experience Variables

- Employees have **an average of 11.3 years of total work experience**, with **7 years at the current company** and **4 years in their current role**. This suggests **moderate retention, but also possible stagnation for those without promotions.**
- The **Years Since Last Promotion averages 2.18 years**, and **longer gaps may lead to dissatisfaction or job changes.**
- The **Number of Companies Worked mean is 2.69**, indicating varied employment histories — **some employees are loyal, while others frequently change jobs.**

3. Compensation Variables

- The **average monthly income is around ₹6,502**, but with **a high standard deviation (₹4,707)**, showing **large salary differences among employees**.
- This variation reflects **differences in job level, experience, and performance** — **employees in higher job levels (3–5) earn significantly more than those in entry-level roles.**
- The Monthly Rate also shows **large variation** (mean = ₹14,313; std = ₹7,117), confirming **income diversity across departments**.
- **Percent Salary Hike ranges between 11–25%**, with most employees receiving **moderate annual increases**, indicating a **consistent compensation policy.**

4. Satisfaction and Engagement Factors (Scale 1–4)

- Job Satisfaction, Environment Satisfaction, and Relationship Satisfaction **all have mean values between 2.7–2.8**, implying **moderate satisfaction levels.**
- **Work Life Balance averages 2.76,** suggesting a generally **positive work-life perception,** though employees with lower ratings may be more likely to leave.
- **Job Involvement (mean = 2.73)** indicates that **most employees show medium to high engagement with their roles.**

5. Performance and Training

- **Performance Rating averages 3.15,** with **most employees rated as Excellent or Outstanding**, indicating overall strong performance levels.
- **Training Times Last Year has a mean of 2.79**, showing that most employees **attended about 2–3 training programs**, representing **moderate emphasis on skill development.**

6. Key Interpretations

- The **standard deviation for Monthly Income and Monthly Rate exceeds 50% of their mean**, showing high variability — **expected due to diverse job levels and experience.**
- Variables such as **Distance From Home and Number of Companies Worked also show high dispersion, reflecting lifestyle and career path diversity.**
- Most **satisfaction-related variables** show **moderate averages**, suggesting that while **employees are not dissatisfied,** there is room for **improving engagement and job satisfaction.**

7. Analytical Implications

- Despite demonstrating **high performance ratings and active participation in training programs**, many employees report only **moderate levels of job satisfaction.** This suggests that strong performers who invest in skill development and show high engagement may still **feel undervalued or inadequately compensated.** Such a mismatch between contribution and reward can **lead to hidden dissatisfaction, making these high-performing employees more likely to consider leaving the organization.**
- High dispersion in salary-related and experience-based variables suggests that **these features will play a major role** in predicting attrition.
- Moderate satisfaction levels and varying work-life balance scores hint that **non-financial factors (like culture and management)** are equally important in retention.
- These statistical patterns set the foundation for deeper exploration through **visual analysis** and **machine learning modeling** in subsequent sections.

5.3 Visual Statistics and Analysis

This section visually explores the patterns and relationships identified in the dataset to better understand the factors influencing employee attrition.

Each visualization highlights a specific dimension of the data — demographic, career-related, satisfaction, and performance — to identify key drivers of employee turnover.

5.3.1 Overall Attrition Rate

Let’s plot a bar chart to show the proportion of employees who left the company versus those who stayed.

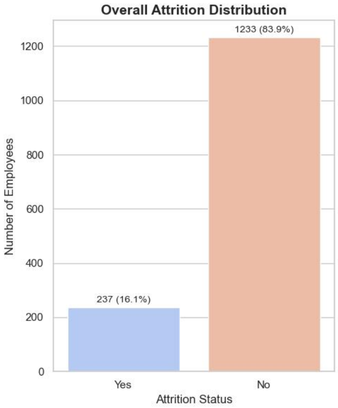

Figure - 1: Overall Attrition

Approximately **16% of employees have left the organization.** This indicates a **moderate attrition rate,** suggesting that while retention is stable, there is still room for improvement in engagement and satisfaction.

5.3.2 Employee Attrition by Demographic Patterns (by Gender, Department, and Marital Status) and Age

Let’s plot a clustered bar chart to analyze how demographic and departmental characteristics — specifically **Gender, Department, and Marital Status** — influence **employee attrition rates**. Understanding these aspects helps identify which workforce groups are more prone to turnover and guides HR in designing targeted retention strategies.

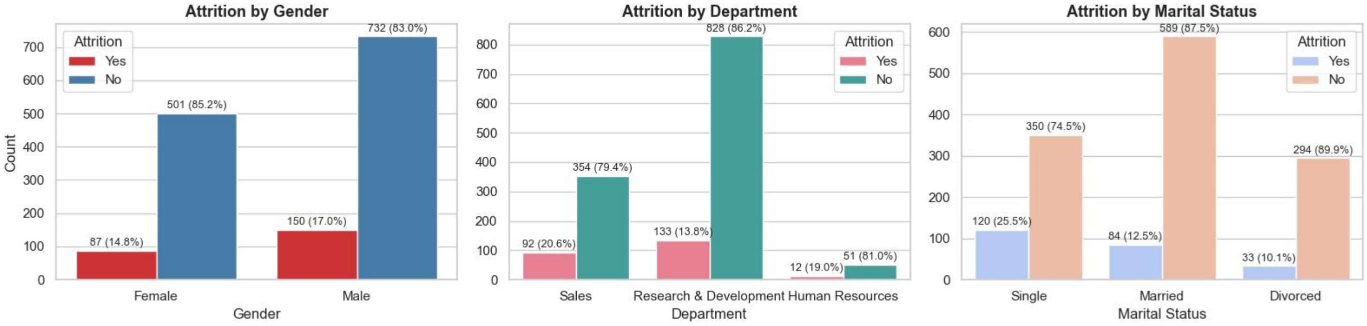

Figure-2: Employee’s Attrition by Demographic Patterns

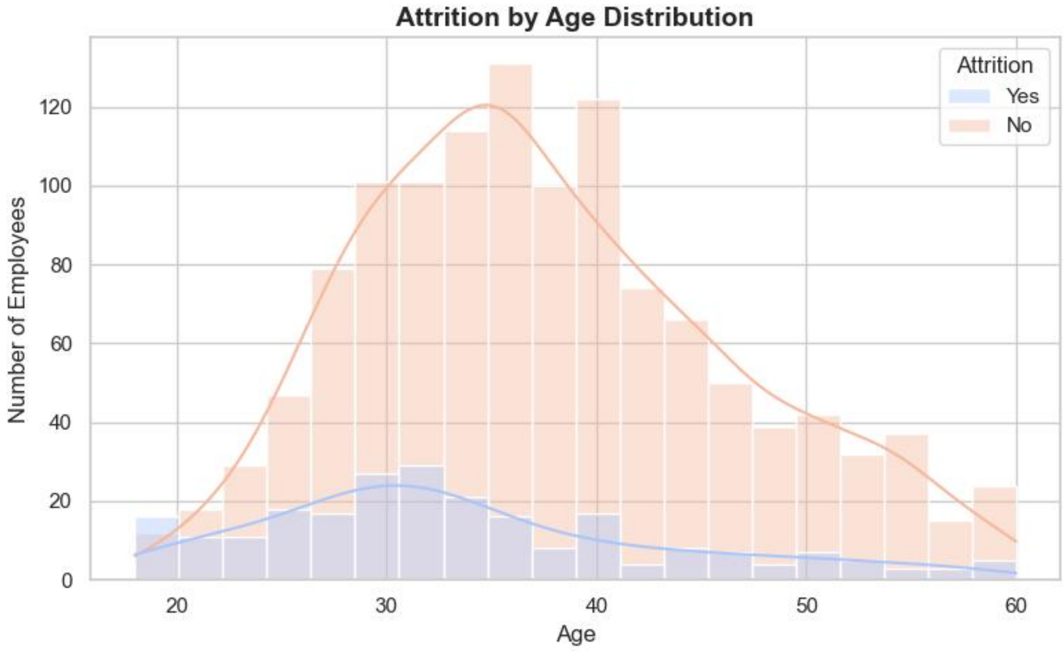

Figure-3: Employee’s Attrition by Age Distribution

The combined demographic trends suggest that **young, single employees in high-pressure departments** such as **Sales** represent the group **most at risk of attrition.** Tailored retention strategies focusing on **career progression, mentorship, and improved role satisfaction** for this segment could help **reduce turnover** and **maintain organizational stability.**

5.3.3 Attrition by Over Time and Work Life Balance

Let’s plot a clustered bar plot to check whether working over time and work life balance effects employee attrition.

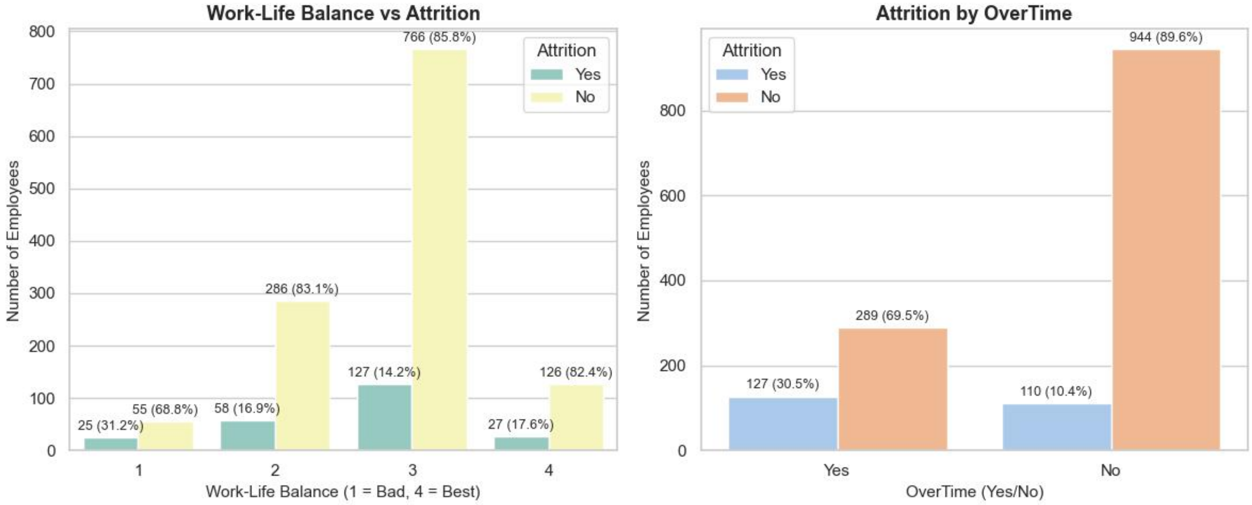

Figure-4: Employee’s Attrition by Over Time and Work-Life Balance

Employees experiencing **poor work-life balance and frequent overtime** are far more likely to leave the organization, regardless of their performance or role. These results highlight the need for HR policies that promote **balanced workloads, reasonable working hours, and mental well-being** initiatives. Encouraging flexible schedules, ensuring adequate rest periods, and monitoring overtime trends can significantly enhance employee retention and satisfaction.

5.3.4 Relationship Between Employee Performance, Compensation, and Attrition

Let’s Check whether employee Performance Rating and Monthly Income effects employee’s Attrition.

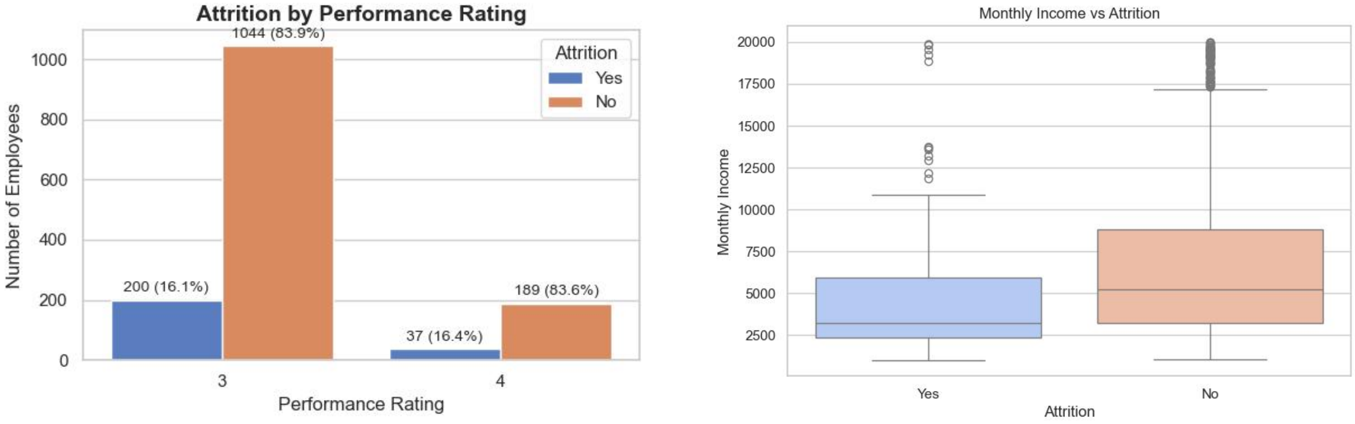

Figure-5: Employee’s Attrition by Performance Rating and Monthly income

These results reveal a **misalignment between employee contribution and reward structure.** High-performing employees with moderate income and limited recognition are more likely to leave, seeking better pay or advancement elsewhere. Addressing this requires implementing **performance-based incentives, clear promotion pathways**, and **transparent compensation policies** to improve both satisfaction and retention.

5.3.5. Correlation Heatmap of Numerical Features

Let’s draw a correlation Heatmap to observe interrelations between numerical variables.

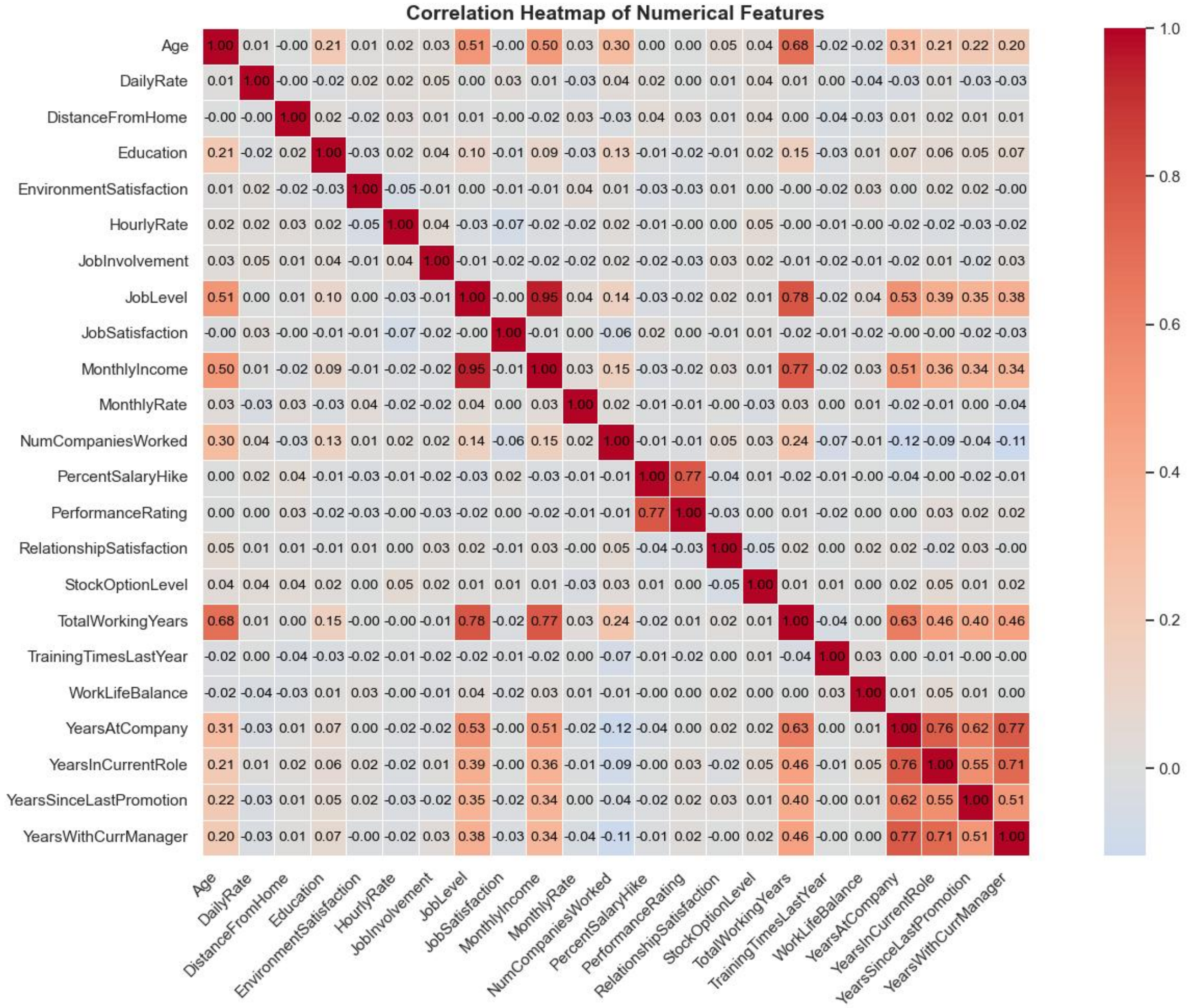

Figure-6: Correlation Heatmap for Numerical variables

The correlation structure reveals that **income, experience, and job level** are tightly interconnected — changes in one strongly influence the others. Meanwhile, **satisfaction-related factors** remain largely independent, meaning they could provide unique predictive power in attrition modeling.

This combination of **quantitative (salary, experience)** and **qualitative (satisfaction, work-life balance)** variables offers a balanced feature base for machine learning, where both career growth and emotional engagement drive employee retention outcomes.

5.4 Transition from EDA to Predictive Modeling

The exploratory data analysis revealed that employee attrition is strongly associated with factors such as low income, frequent overtime, moderate satisfaction, and shorter tenure. Correlation analysis showed high interdependence among Job Level, Monthly Income, and Total Working Years, while engagement metrics like Work Life Balance and Environment Satisfaction remained largely independent.

These insights guided the selection of features for the modeling phase. Highly correlated variables were streamlined to prevent redundancy, and satisfaction-based indicators were retained for their unique predictive value. The next section focuses on developing machine learning models to predict attrition likelihood and identify key drivers influencing employee turnover.

Section-6: Predictive Modeling and Feature Engineering

6.1 Objective

The primary goal of this phase is to develop a predictive model that can accurately classify whether an employee is likely to leave the organization. The model leverages the insights from the exploratory analysis to identify and quantify the influence of key factors such as job satisfaction, compensation, tenure, and workload on employee attrition.

6.2 Data Preprocessing

Before modeling, several preprocessing steps were applied to ensure data quality and compatibility with machine learning algorithms:

- **Handling Missing Values:** The dataset contained no missing values; hence, no imputation was required.
- **Encoding Categorical Variables:** Variables such as Gender, Marital Status, Business Travel, and Department were converted into numerical form using Label Encoding or One-Hot Encoding.
- **Feature Scaling:** Numerical features (e.g., Monthly Income, Age, Years At Company) were standardized using Min–Max Scaling to maintain consistency across features.
- **Feature Selection:** Based on correlation analysis, redundant features such as Job Level (highly correlated with Monthly Income) were excluded to avoid multicollinearity.
- **Target Variable:** The dependent variable Attrition was binary encoded — Yes = 1, No = 0.

6.3 Model Selection

To identify the most suitable predictive model, multiple supervised learning algorithms were evaluated:

- **Logistic Regression** – for interpretability and baseline performance.
- **Random Forest Classifier** – for capturing non-linear relationships and feature importance.
- **Decision Tree Classifier** – for visualization and explainability.
- **Support Vector Machine (SVM)** – for boundary-based classification.

The dataset was divided into training (80%) and testing (20%) subsets to validate model generalization.

6.4 Model Evaluation

Each model was assessed using standard performance metrics:

- Accuracy: Measures overall correctness.
- Precision and Recall: Evaluate the model’s ability to correctly predict attrition cases.
- F1-Score: Balances precision and recall for imbalanced classes.
- ROC–AUC Score: Reflects model’s discriminatory power between attrition and retention classes.

A confusion matrix was also analyzed to visualize the true positive, true negative, false positive, and false negative predictions.

6.5 Key Results

After performing the four supervised learning algorithms, the results are as follows:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**87.41**	**69.23**	**38.3**	**49.32**	**80.57**
SVM	85.37	66.67	17.02	27.12	80.91
Random Forest	84.01	50	12.77	20.34	79.27
Decision Tree	78.23	31.91	31.91	31.91	59.48

Table-1: Comparison of Supervised Learning Model Performance

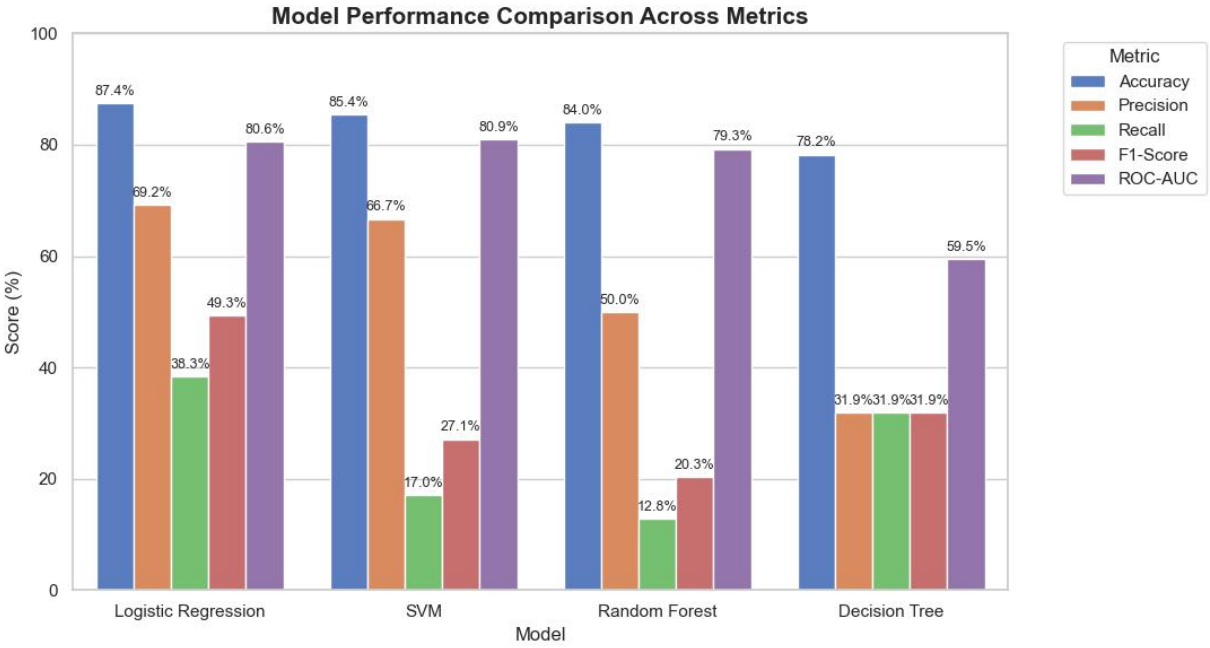

Figure-7: Model Performance Comparison Across Metrics

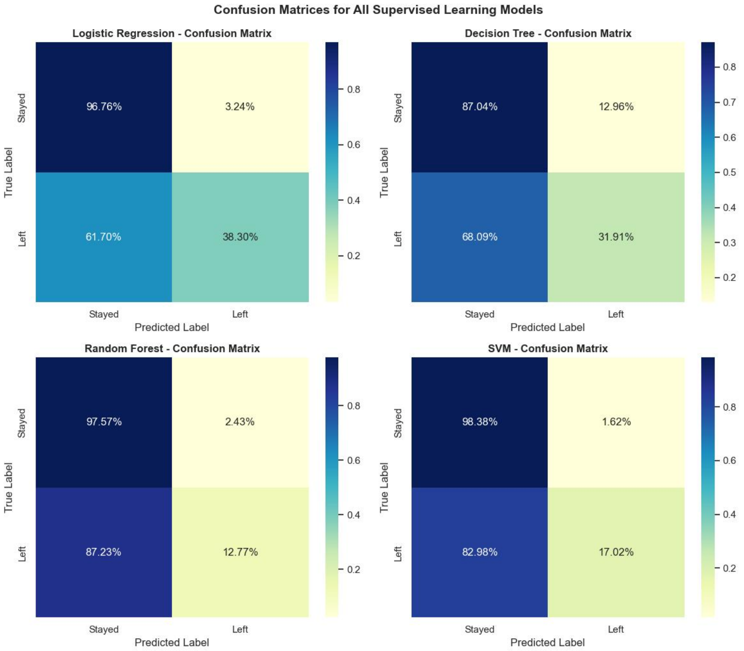

Figure-8: Confusion Matrices of Supervised Learning Models

The **Logistic Regression model** provided the **best balance** between **accuracy and generalization.**

- **It achieved strong performance with a high ROC–AUC score (80.57%), indicating good discriminatory power in distinguishing between employees who stayed and those who left.**
- **The SVM model performed similarly in ROC–AUC but had lower recall, meaning it missed several true attrition cases.**
- **The Random Forest and Decision Tree models underperformed, suggesting overfitting and weaker recall capabilities.**

In conclusion, **Logistic Regression** is selected as the final model for attrition prediction due to its stability, interpretability, and strong predictive balance.

The top predictive features identified were:

- **OverTime**
- **MonthlyIncome**
- **Years At Company**
- **JobSatisfaction**
- **Age**
- **WorkLifeBalance**

These features align closely with the findings from EDA, confirming the reliability of the analysis.

6.6 Summary

The predictive modeling phase successfully identified patterns that distinguish employees at high risk of attrition. While **Logistic Regression** produced the **best overall predictive performance,** the **Random Forest model** was used for **feature interpretation** due to its ability to capture **non-linear relationships and rank variable importance.** This combination provided both **predictive accuracy and actionable insights** — allowing HR teams to understand not only who is at risk of attrition but also which factors contribute most to that risk.

7. Findings and Recommendations

7.1 Key Findings

The predictive analysis of **IBM's HR dataset** revealed several critical factors contributing to employee attrition. **Logistic Regression** emerged as the **most reliable model**, achieving an accuracy of approximately **87% and a ROC–AUC of 0.80**, providing a balanced trade-off between interpretability and predictive strength.

From both EDA and model feature importance analyses, the following variables were identified as the most influential in predicting attrition:

- **"Over Time – Employees working overtime are significantly more likely to leave."**
- **"Monthly Income – Lower-income employees exhibit higher turnover, emphasizing pay disparity as a key factor."**
- **"Job Satisfaction and Environment Satisfaction – Low satisfaction levels consistently correspond to higher attrition rates."**
- **"Years At Company and Total Working Years – Shorter tenure and early-career employees are more prone to leave."**
- **"Work Life Balance – Poor work-life balance strongly correlates with attrition."**

These findings collectively suggest that while compensation and workload are primary drivers of attrition, employee engagement and satisfaction play equally important roles.

7.2 Strategic Recommendations

Based on the analysis, the following actions are recommended to help reduce employee turnover and improve retention rates:

Area	Recommendation	Expected Impact
Workload Management	Monitor and limit overtime hours; implement workload balancing policies.	Reduces burnout and improves retention.
Compensation Review	Review pay structures and ensure fair increments based on experience and performance.	Addresses dissatisfaction among lower-paid employees.
Employee Engagement	Conduct regular surveys on job satisfaction and work-life balance; create recognition programs.	Enhances motivation and loyalty.
Career Growth	Introduce clear promotion pathways and development plans for high-performing employees.	Reduces attrition among skilled staff seeking growth.
Onboarding and Early Support	Strengthen onboarding and mentoring for employees with <3 years tenure.	Reduces early-career attrition and improves role fit.

Table-2: Recommendations proposed based on EDA and Predictive Modelling

7.3 Conclusion

The project successfully applied data analytics and machine learning to predict and understand employee attrition. The combination of **Logistic Regression for prediction** and **Random Forest for feature interpretation** provided actionable insights for HR policy design. Overall, the study demonstrates how data-driven decision-making can effectively support workforce stability and organizational growth.

8. Project Summary and Learning Reflection

8.1 Summary

This project aimed to analyze and predict employee attrition using IBM’s HR Analytics dataset. The study followed a structured data science workflow — beginning with data exploration and visualization, followed by correlation and statistical analysis, and culminating in the application of multiple machine learning models.

Through this analysis, key factors such as Over Time, Monthly Income, WorkL ife Balance, Job Satisfaction, and Years At Company were identified as the primary drivers of employee attrition. Logistic Regression emerged as the most reliable model, **achieving 87.4% accuracy** and a **ROC–AUC of 80.6%**, offering both interpretability and strong generalization.

The findings support the conclusion that workload management, compensation fairness, and job satisfaction play crucial roles in retaining employees, and data-driven insights can effectively guide HR decision-making.

8.2 Personal Reflection

Working on this project enhanced my practical understanding of data science and machine learning workflows. I gained hands-on experience in **data preprocessing, feature selection, model building, and performance evaluation using Python and Scikit-learn.**

Beyond technical execution, the project strengthened my ability to derive **business insights from data,** connecting analytics with organizational outcomes.

This experience also deepened my proficiency in tools like **Jupyter Notebook, and visualization libraries (Matplotlib, Seaborn)** — skills that are essential for real-world analytics and reporting.

8.3 Future Scope

Future improvements could involve testing ensemble models such as **XGBoost or Gradient Boosting,** applying **hyperparameter tuning**, and incorporating **cross-validation** for better generalization. Additionally, integrating **real-time HR data dashboards** in Power BI could provide continuous attrition monitoring for proactive decision-making.