Employee Attrition Prediction Report

Project 51 — HR Analytics

---

1. Project Overview

Employee attrition is a critical concern for organizations, as it impacts productivity, morale, and recruitment costs. The objective of this project is to **analyze patterns of employee attrition** and build a predictive model to identify employees at risk of leaving. This allows HR teams to take proactive retention measures.

Dataset Used: IBM HR Analytics Employee Attrition Dataset from Kaggle
Number of Records: 1,470 employees
Features: Age, BusinessTravel, Department, DistanceFromHome, Education, JobRole, MonthlyIncome, OverTime, YearsAtCompany, etc.
Target Variable: Attrition (Yes/No)

---

2. Data Cleaning & Preprocessing

- Converted Attrition to numeric: Yes → 1, No → 0
- Dropped irrelevant columns: EmployeeCount, EmployeeNumber, Over18, StandardHours
- Encoded categorical variables using **one-hot encoding** (BusinessTravel, Department, JobRole, MaritalStatus, OverTime, Gender, EducationField)

---

3. Exploratory Data Analysis (EDA)

Attrition Distribution:

- Majority of employees stayed (No), while a smaller portion left (Yes).
- Attrition rate: ~16% of employees left the company.

Key Observations:

1. **Department:** Attrition is higher in Research & Development compared to Sales and HR.
2. **OverTime:** Employees working overtime are more likely to leave.
3. **Age:** Younger employees tend to have higher attrition.
4. **Monthly Income:** Employees with lower monthly income show higher attrition.

5. **Years at Company:** Employees with fewer years at the company are more likely to leave.

Visualizations created:

- Count plots for Attrition by Department and OverTime
- Boxplots for Age vs Attrition, Monthly Income vs Attrition

---

4. Modeling Approach

Data Split:

- Train set: 80% of data
- Test set: 20% of data
- Stratified split to preserve target proportion

Models Used:

1. **Logistic Regression** – baseline predictive model
2. **Random Forest Classifier** – tree-based ensemble model for better accuracy and feature importance

---

5. Model Evaluation

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.86	0.71	0.53	0.61
Random Forest Classifier	0.88	0.74	0.58	0.65

- **Random Forest performed better** than Logistic Regression.
- Both models show reasonable predictive ability, though recall indicates some employees at risk may be missed.

---

6. Feature Importance (Random Forest)

Top 10 features influencing attrition:

1. OverTime
2. MonthlyIncome

3. Age
4. JobRole
5. YearsAtCompany
6. JobLevel
7. DistanceFromHome
8. EnvironmentSatisfaction
9. TotalWorkingYears
10. RelationshipSatisfaction

Insight: Employees working overtime, earning lower salaries, and younger employees with fewer years at the company are more likely to leave.

---

7. Business Insights & Recommendations

1. **OverTime Management:** Reduce excessive overtime or provide incentives for overtime work.
2. **Compensation Review:** Monitor salary levels and provide fair compensation to reduce attrition.
3. **Employee Engagement:** Focus on younger employees with less experience through mentorship and career development programs.
4. **Job Role Alignment:** Pay attention to high-risk job roles to retain key talent.
5. **Work-Life Balance:** Improve work-life balance programs to retain employees working long hours.

---

8. Conclusion

- Predictive modeling can help identify employees at risk of leaving.
- Random Forest provides both good predictive performance and feature interpretability.
- HR teams can leverage these insights to design proactive retention strategies, reducing attrition costs and improving workforce stability.