**Employee Attrition Analysis**

This project aims to analyze and predict employee attrition using a combination of data preprocessing, feature engineering, exploratory analysis, and machine learning models. The goal is to identify the key drivers of employee churn and help HR teams take proactive action to improve retention.

**Dataset Overview**

Kaggle Dataset link : <https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study/data>

The dataset includes:

- `general\_data.csv`: Demographics, job role, income, performance.

- `in\_time.csv` & `out\_time.csv`: Employee attendance logs.

- ‘employee\_survey\_data.csv’ : Environment Satisfaction, Job Satisfaction, Work Life Balance

- ‘manager\_survey\_data.csv’ : Job Involvement, Performance Rating

- Label: `Attrition` (Yes/No)

**Objective**

- Identify key behavioral and demographic patterns behind employee attrition.

- Build predictive models to classify whether an employee is likely to leave.

**Data Cleaning & Preparation**

**1. Missing Value Treatment:**

- Replaced or dropped missing values in survey columns.

- Ensured all employees had matching data across files.

**2. Feature Engineering:**

- Derived `avg\_work\_hours` and `absent\_days` from in-time/out-time logs.

- Created binary column `Attrition\_bool` for modeling.

**3. Encoding:**

- Applied label encoding and one-hot encoding on categorical variables.

**4. Multicollinearity:**

- Checked with VIF (Variance Inflation Factor).

- Removed features with high VIF if p-values were insignificant.

**Exploratory Data Analysis (EDA)**

Visual tools like histograms, barplots, and boxplots and Correlation heatmap

were used to uncover patterns.

**Key Insights:**

- Employees with high average work hours showed higher attrition.

- High number of full-day workers left the company possibly due to burn-out. Others left due to very low engagement or absenteeism

- Long gaps since last promotion like at gaps of 3, 10 and 13 years have shown low job satisfaction and were associated with higher attrition.

- Job satisfaction and work-life balance scores were generally lower for those who left.

**Models Used**

|  |  |  |
| --- | --- | --- |
| **Model** | **Purpose** | **Highlights** |
|  |  |  |
| Logistic Regression | Baseline classification | Interpretable coefficients |
| Decision Tree Classifier | Gini-based feature split | Easy to visualize, moderate accuracy |
| Random Forest | Ensemble model | Better accuracy, handles non-linearity |
| XGBoost Classifier | Gradient boosting | Strongest performance |
|  |  |  |

**Model Evaluation**

For **Training** data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | Recall |
| Logistic Regression | 77.1 | 77 | 77.3 | 77.2 | 77 |
| Decision Tree Classifier | 100 | 100 | 100 | 100 | 100 |
| Random Forest | 100 | 100 | 100 | 100 | 100 |
| XGBoost Classifier | 100 | 100 | 100 | 100 | 100 |

For **Testing** Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | Recall |
| Logistic Regression | 75.8 | 60.2 | 77.5 | 34 | 66 |
| Decision Tree Classifier | 96.2 | 91.4 | 97.1 | 84.9 | 91.4 |
| **Random Forest** | **98.8** | **95.4** | **99.4** | **96.9** | **95.4** |
| **XGBoost Classifier** | **98.8** | **95.4** | **99.4** | **96.9** | **95.4** |

**HyperParameter Tuning:**

Hyperparameter tuning was performed using GridSearchCV and RandomSearchCV to tune the models to obtain best results.

**Feature Importance:**

Top predictive features:

**- `avg\_work\_hours`**

**- ‘YearsWithCurrManager’**

**- `JobSatisfaction`**

**- `EnvironmentSatisfaction`**

- **`TotalWorkingYears`**

**Conclusion:**

Random Forest and XGBoost both show good and similar results here with good recall rate of 95.4 as Recall is important here because

**High recall** means model is correctly identifying most of the employees who are *actually going to leave*.

**Low recall** means model is *missing many actual attrition cases* (false negatives).