# Employee Attrition Analysis and Prediction Using Data Analytics Techniques

# List of TABLES

# List of Figures

# TABLE OF CONTENTS

# 1. Introduction

## 1.1 Problem Statement

- Employee attrition—also known as employee turnover—is a persistent issue that affects organizations across industries. While some degree of attrition is natural and even healthy, **high or unanticipated attrition rates** create substantial challenges for HR departments.
- When an experienced employee leaves, the organization not only loses a skill set but also faces the **burden of recruitment, onboarding, and training** a replacement. These processes are time-consuming and resource-intensive. More importantly, it leads to a **loss of institutional knowledge**, reduces team morale, and often disrupts ongoing projects or client relationships.
- In today's knowledge economy, where **employee experience and satisfaction** are as important as salary, understanding why employees choose to leave has become more complex. Traditional HR methods like surveys and exit interviews are reactive and often biased. Therefore, **proactively identifying the signs and factors that drive attrition is no longer a luxury but a necessity** for forward-thinking HR teams.
- This project takes a modern approach to this issue by applying **machine learning and data analytics** to uncover patterns in employee behavior and engagement that might signal a risk of attrition.

## 1.2 Project Summary

- The goal of this project is to build a system that can **predict whether an employee is likely to leave the organization** using historical HR data. The dataset contains a variety of employee attributes including **age, salary, education field, job satisfaction, years at company, relationship with manager, overtime hours**, and more.
- Instead of relying on gut feelings or post-exit analysis, the system learns from past data to detect early warning signs of disengagement. These might include frequent overtime, low satisfaction scores, or being stuck in the same role for too long without promotion.
- The predictive system doesn't stop at just giving a "yes or no" output. It is enhanced with **explainable AI (SHAP values)**, so HR leaders can understand **why** a certain prediction was made—enabling them to take precise and targeted actions.
- Additionally, the project integrates **Power BI dashboards** to convert complex data into intuitive visualizations. This ensures that even non-technical HR stakeholders can interact with the data and insights easily.

## 1.3 Aim & Objectives

### 1.3.1 To develop a data-driven predictive model for employee attrition.

The primary goal is to move beyond guesswork and gut-feeling decisions. By training models like **Logistic Regression and Decision Trees**, the system learns from historical employee data and makes statistically-grounded predictions about who might leave the company next. This model will allow HR to **act before an employee submits their resignation letter**.

### 1.3. 2. To visualize insights using dashboards for HR decision-making.

Visual storytelling is key in data analytics. Through **Power BI**, complex results such as attrition by department, salary bands, and job role impact are transformed into easy-to-understand visuals. This empowers managers and HR professionals to explore patterns interactively and draw **actionable conclusions** from real data.

### 1.3 3. To interpret the model's decisions using explainable AI techniques.

A model is only useful if its decisions are transparent. That's why this project uses **SHAP (SHapley Additive Explanations)**, a method rooted in game theory, to explain each prediction. It tells you which features—like 'OverTime', 'WorkLifeBalance', or 'YearsAtCompany'—played the biggest role in influencing a prediction. This builds **trust** in the model and supports ethical AI practices.

### 1.3. 4. To recommend strategies to reduce attrition based on analysis.

Data is only valuable when it leads to action. After identifying key risk factors for attrition, the project concludes by offering **HR-specific strategies**—like enhancing employee engagement in high-turnover departments, reviewing compensation policies, and developing career growth frameworks. These recommendations are tailored to the insights gained from both **machine learning models and data visualizations**, making them evidence-based and practical.

# 2. System Analysis

## 2.1 Motivation

Employee retention plays a critical role in the success and stability of any organization. Retaining skilled and experienced employees helps reduce training costs, fosters team cohesion, and maintains productivity. However, many companies still rely on **reactive HR strategies**—responding to resignations instead of preventing them.

This project is motivated by the idea that **attrition can be predicted and prevented** if organizations use data-driven approaches. With the rise of big data and machine learning, it's now possible to uncover hidden patterns in employee behavior that precede resignation.

Rather than relying solely on **exit interviews or satisfaction surveys**, this project empowers HR managers with tools that **forecast attrition risk before it's too late**. The goal is to **transition HR from a support role to a strategic decision-maker**, using predictive analytics to build a more engaged, stable, and productive workforce.

## 2.2 Literature Survey (Expanded)

The topic of employee attrition has been widely researched in both academic and industrial settings. Several machine learning techniques have been applied, including:

### 2.2.1. Decision Tree Models

- Simple to understand and interpret.
- Breaks down decisions into a tree of rules based on employee features.
- However, prone to **overfitting**, and performance drops when data is imbalanced.

### 2.2.2. Logistic Regression

- Highly interpretable and suitable for binary classification like attrition (Yes/No).
- Produces probability-based predictions and is easy to implement.
- However, it assumes linear relationships and may not capture complex interactions.

### 2.2.3. Random Forests

- Ensemble method combining multiple decision trees for better accuracy.
- Handles non-linear relationships well.
- Lack of transparency, i.e., **“black box”** nature makes it difficult to explain decisions.

### 2.2.4. Support Vector Machines (SVM)

- Good for small to medium-sized datasets.
- Can handle non-linear boundaries.
- Limited interpretability and sensitivity to feature scaling.

### 2.2.5. SHAP with Logistic Regression (SHAP-LR)

- Combines the interpretability of LR with the feature attribution capability of SHAP.
- Explains individual predictions and reveals **global feature importance**.
- Best suited when explainability and fairness are priorities.

## 2.6 Research Gap

Despite a large volume of work on attrition prediction, several critical gaps remain:

### 2.6.1. Interpretability is Overlooked

Most high-performing models like Random Forest or XGBoost provide accurate results but offer **little to no insight into *why* an employee is at risk of leaving**. In HR, this is a major limitation because actionable interventions require a clear understanding of root causes.

### 2.6.2. Class Imbalance is Ignored

Attrition datasets are often **imbalanced**, meaning far more employees stay than leave. Many models fail to address this issue, leading to biased predictions—accurately predicting those who stay while **missing the ones who are actually at risk**.

### 2.6.3. Lack of Business-Readable Insights

Many tools do not bridge the gap between data scientists and HR professionals. The end result is that **HR teams don’t trust or use the predictions**, because they cannot relate them to their workflows or policies.

### 2.6.3.1 RESULT TABLE OF MODEL ACCURACY AND SCALABILITY

| Model | Accuracy | Scalability | Remarks |
|---|---|---|---|
| **Decision Tree** | Moderate | High | Easy to implement and interpret; works well on small to medium datasets. |
| **Logistic Regression** | Moderate | Very High | Ideal for quick, interpretable models; handles large datasets efficiently. |
| **Random Forest** | High | High | Accurate and robust; scalable with parallel processing; less interpretable. |
| **Support Vector Machine (SVM)** | Moderate | Moderate-Low | Good accuracy on smaller datasets; struggles with very large datasets. |
| **SHAP + Logistic Regression** | Moderate) | Moderate | Adds model explanation but adds some processing time due to SHAP complexity. |

**TABLE 2.6.3.1 MODEL ACCURACY AND SCALABILITY**

# 3. Design, Methodology & Implementation Strategy

## 3.1 Overview

The design of this project follows a structured pipeline: starting from data gathering, moving to preprocessing, model training, evaluation, interpretation, and finally reporting the insights through dashboards. This modular design ensures flexibility, transparency, and adaptability to real-world HR systems.

## 3.2 Hardware and Software Requirements

### 3.2.1 Hardware Requirements:

| **Component** | **Specification** |
|---|---|
| Processor | Intel i5/i7 or equivalent (4+ cores) |
| RAM | Minimum 8 GB (16 GB recommended) |
| Storage | 250 GB SSD or HDD (for dataset and tools) |
| GPU (Optional) | For faster SHAP visualizations (NVIDIA CUDA support) |

**TABLE 3.2.1 Hardware Requirements**

### 3.2.2 Software Requirements:

| **Software/Tool** | **Purpose** |
|---|---|
| Python 3.10+ | Primary language for scripting |
| Jupyter Notebook | Development and visual coding interface |
| Anaconda / pip | Environment and package manager |
| Power BI Desktop | Data dashboard and report visualization |
| scikit-learn | Model training & evaluation |
| SHAP | Model interpretability |
| Matplotlib & Seaborn | Data visualization |
| Pandas & NumPy | Data wrangling and numeric operations |

**TABLE 3.2.2 Software Requirements**

## 3.3 Design Methodology

The project adopts a **hybrid waterfall and iterative model**, where foundational steps like data preprocessing and model selection are sequential, but model tuning and dashboarding are done iteratively. This ensures early testing, better accuracy, and quick turnaround in stakeholder feedback.

### 3.3.1 Key Phases of Design:

1. **Problem Formulation**
   → Understand business pain points in HR attrition.
   → Frame as a binary classification problem.

2. **Data Understanding**
   → Analyze structure, missing values, balance, feature types.
   → Explore potential bias or outliers (e.g., outliers in salary or tenure).

3. **Feature Engineering & Selection**
   → Drop irrelevant columns (e.g., EmployeeCount, StandardHours).
   → Encode categorical variables (OneHot, LabelEncoder).
   → Normalize numerical features for model stability.

4. **Model Design**
   → Start with **Logistic Regression** for transparency.
   → Add **Decision Tree** for rule-based decisions.
   → Use **class_weight='balanced'** to manage imbalance.
   → Evaluate via Accuracy, Precision, Recall, F1-Score.

5. **Interpretability Layer**
   → Integrate **SHAP** to make model decisions explainable.
   → Produce global and local explanations using summary and waterfall plots.

6. **Visualization Layer**
   → Export data to Power BI.
   → Build dashboards for HR (Attrition by Department, Role, Salary Band).

## 3.4 Implementation Strategy

A robust and structured workflow was followed to ensure clarity, traceability, and iterative improvements. Below is a modular breakdown of implementation:

### 3.4.1 System Flowchart of  HR ANALYTICS:

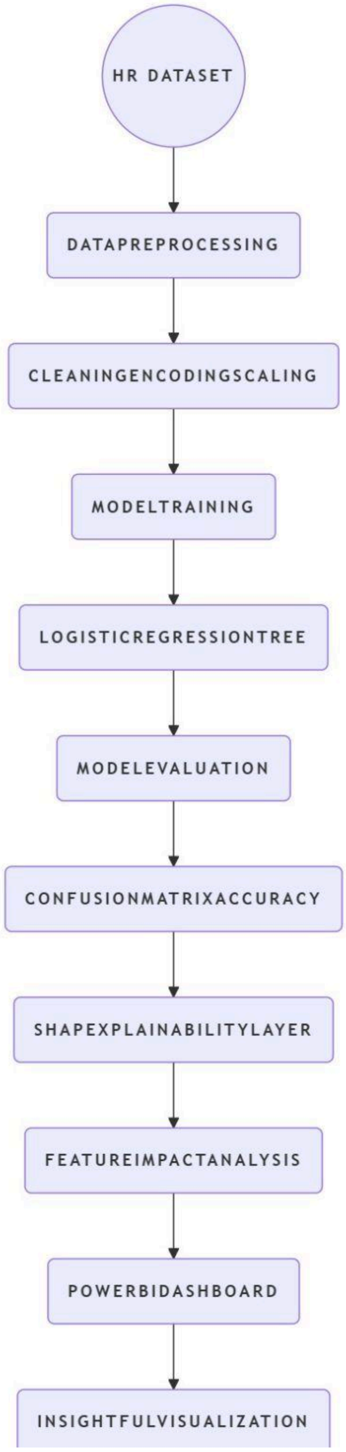


**Fig 3.4.1 HR ANALYTICS FLOWCHART DIAGRAM**

## 3.5 Timeline Chart

| DAYS | Tasks |
|---|---|
| DAY 1 | Literature Review, Dataset Exploration |
| DAY 2 | Data Preprocessing, Feature Engineering |
| DAY 3 | Logistic Regression Model (initial) |
| DAY 4 | Decision Tree Model + Evaluation |
| DAY 5 | SHAP Integration for Interpretability |
| DAY 6 | Dashboard Prototyping in Power BI |
| DAY 7 | Attrition Insight Analysis + Prevention Strategies |
| DAY 8 | Report Writing, Final Testing, and Presentation |

**TABLE 3.5 Timeline Chart**

## 3.6 Modular Specifications

| Module | Function |
|---|---|
| | |
| preprocessing.py | Cleans and encodes HR data, handles nulls, scales values |
| train_model.py | Trains and evaluates LR & Decision Tree models |
| shap_analysis.py | Generates SHAP summary plots and feature importance visualizations |
| visual_dashboard.pbix | Interactive Power BI Dashboard file for HR decision makers |
| attrition_suggester.py | Generates insights and suggestions based on SHAP and department-level data |

**TABLE 3.6 Modular Specification**

## 3.7 Summary of Design & Implementation

- The system is designed around **modularity**, ensuring smooth data-to-dashboard flow.
- The choice of **Python + SHAP + Power BI** balances transparency, accuracy, and usability.
- Clear hardware/software specs help ensure reproducibility.
- Timeline-based development ensures consistent progress and feedback loops.

# 4. Implementation

## 4.1 System Flow: Step-by-Step Process

1. **Load Data from CSV**
   The project begins by importing the dataset (HR_Analytics.csv) using pandas.read_csv(). This data consists of multiple features such as employee demographics, job role, income, work experience, satisfaction scores, and attrition status.

2. **Conduct Exploratory Data Analysis (EDA)**
   The data is explored visually and statistically to identify patterns and anomalies. Charts such as histograms, barplots, heatmaps, and pie charts are used to observe trends like attrition rate by department, job role, or salary band.

3. **Encode and Scale Features**
   Categorical variables are transformed using Label Encoding and One-Hot Encoding, while numerical variables are standardized using StandardScaler from Scikit-learn. This ensures the model treats all features fairly.

4. **Train/Test Models (Logistic Regression & Decision Tree)**

   i. Logistic Regression: Trained for its interpretability and ability to handle binary classification.

   ii. Decision Tree: Used for its intuitive decision-making paths and visualization capabilities.
   Both models are evaluated using accuracy, confusion matrix, and classification reports.

### 4.1.1 LOGISTIC REGRESSION:

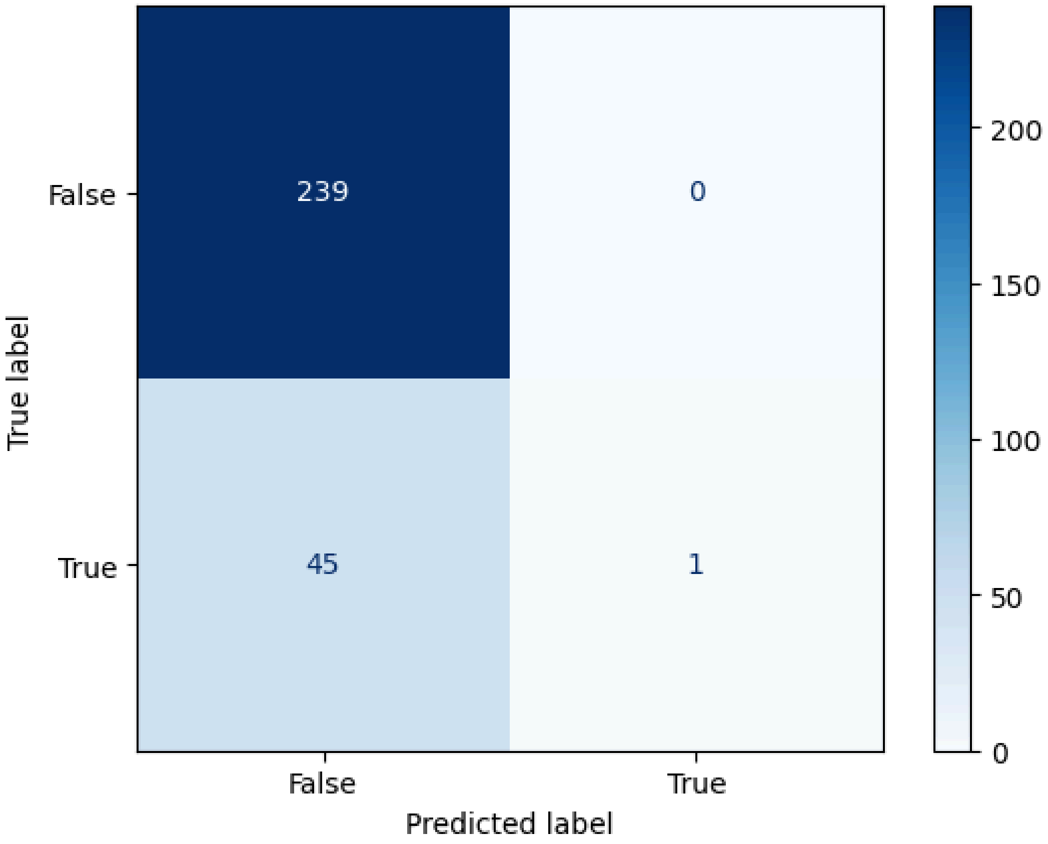


**Fig 4.1.1 Logistic Regression**

### 4.1.2 DECISION TREE :

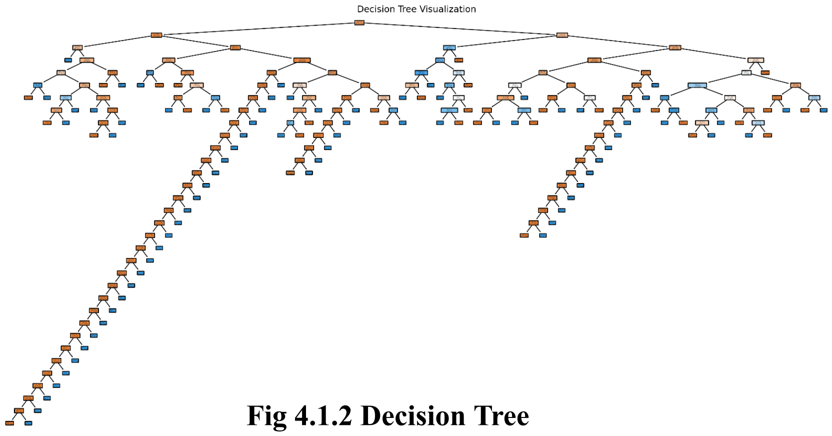


**Fig 4.1.2 Decision Tree**

5. **Interpret Predictions Using SHAP**
   SHAP (SHapley Additive exPlanations) is integrated to explain individual predictions. This helps HR understand why the model predicts an employee might leave, showing feature contributions for each case.

## 4.1.3 Shap result Prediction

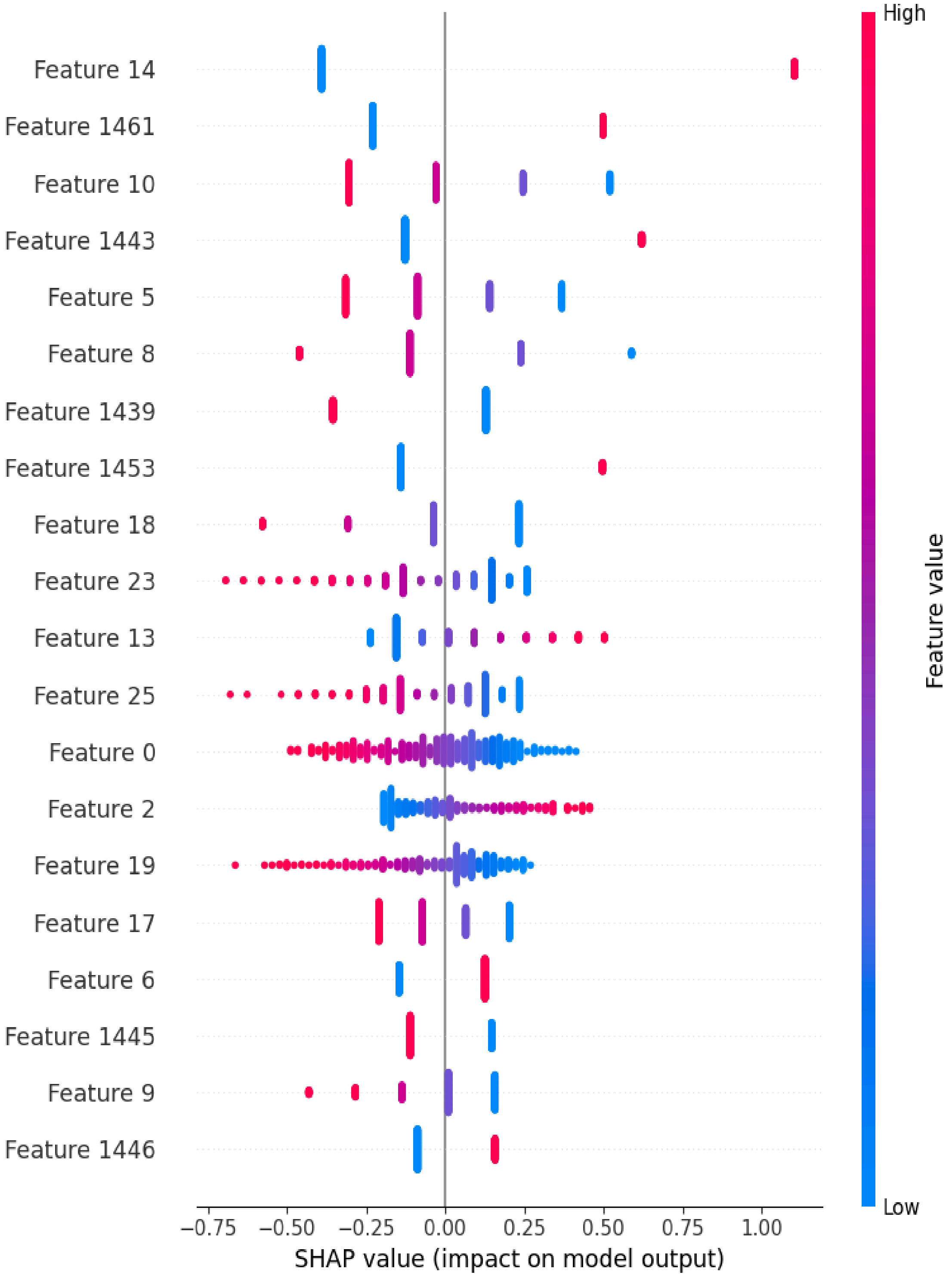
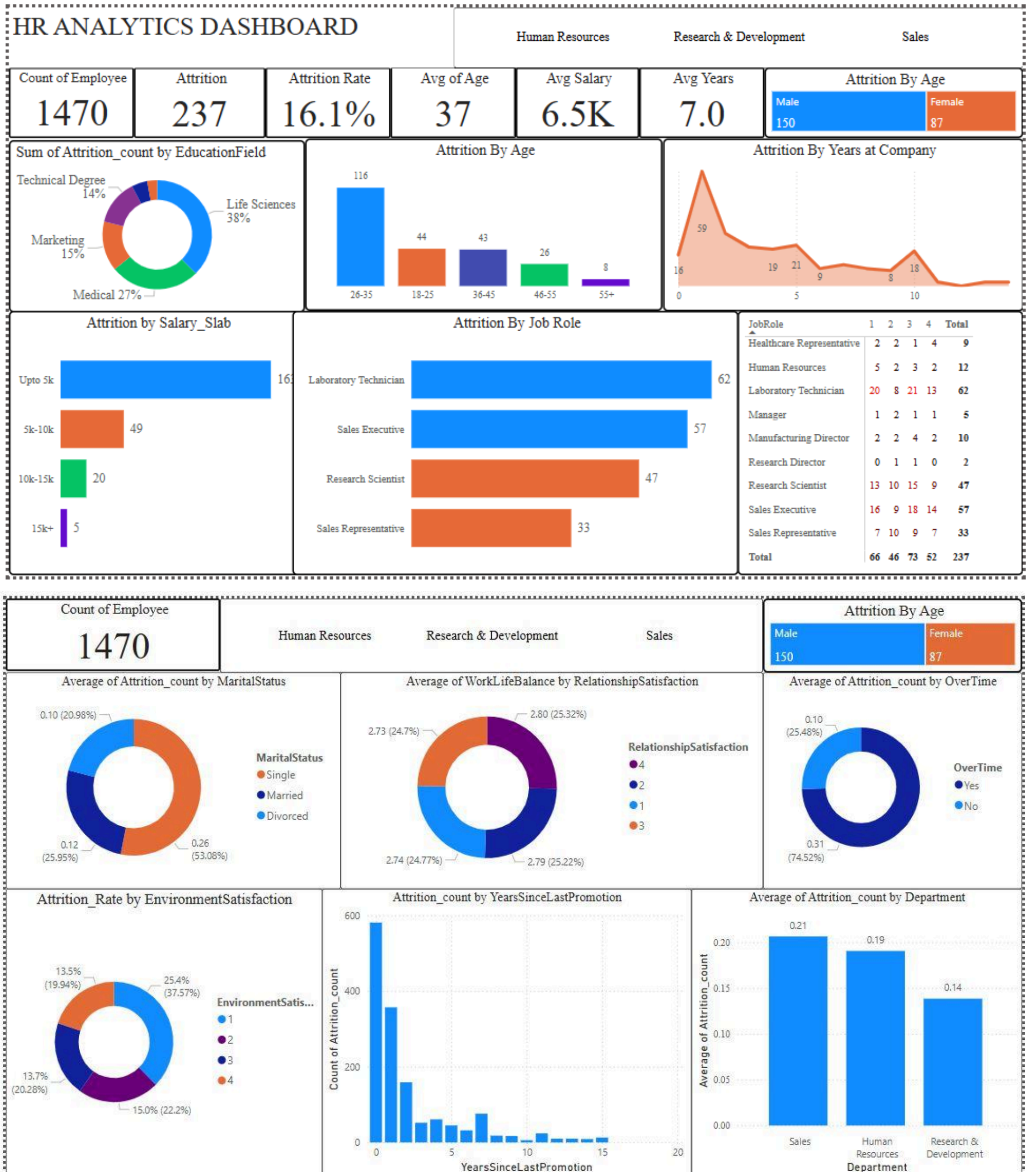


**Fig 4.1.3 Shap Result Prediction**

6. **Visualize Results in Power BI**
   Key findings are exported and visualized in Power BI dashboards, allowing HR personnel to interact with attrition data, explore high-risk segments, and filter results based on roles, age, departments, etc.

7. **Recommend HR Strategies Based on Findings**
   Based on the model insights and SHAP values, concrete strategies are proposed to reduce attrition — such as adjusting overtime policies, improving work-life balance, enhancing career progression, and improving environmental satisfaction.

## 4.1.4 HR ANALYTICS BOARD USING POWER BI:



| JobRole | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| Healthcare Representative | 2 | 2 | 1 | 4 | 9 |
| Human Resources | 5 | 2 | 3 | 2 | 12 |
| Laboratory Technician | 20 | 8 | 21 | 13 | 62 |
| Manager | 1 | 2 | 1 | 1 | 5 |
| Manufacturing Director | 2 | 2 | 4 | 2 | 10 |
| Research Director | 0 | 1 | 1 | 0 | 2 |
| Research Scientist | 13 | 10 | 15 | 9 | 47 |
| Sales Executive | 16 | 9 | 18 | 14 | 57 |
| Sales Representative | 7 | 10 | 9 | 7 | 33 |
| Total | 66 | 46 | 73 | 52 | 237 |

**Fig 4.1.4, 4.1.5 HR_ANALYTICS Dashboard**

## 4.2 Program/Module Specification

### 1) Data Preprocessing

This module is responsible for preparing the raw HR dataset for analysis. It begins by handling missing values, removing irrelevant or constant columns, and ensuring data consistency. Categorical variables are encoded using Label Encoding and One-Hot Encoding, and numerical features are scaled using StandardScaler. These steps standardize the dataset for optimal model performance and reduce bias caused by unprocessed features.

### 2) Exploratory Data Analysis (EDA)

EDA is used to uncover key trends, correlations, and patterns in employee data that influence attrition. Through bar plots, histograms, and heatmaps, this module visualizes attrition rates across different departments, salary ranges, job roles, and satisfaction levels. These insights help identify which segments of the workforce are at higher risk and guide model feature selection.

### 3) Model Training

This module focuses on implementing and training machine learning models—specifically Logistic Regression and Decision Tree classifiers. These models are selected for their interpretability and classification ability. The training process involves splitting the data into train-test sets, fitting models, and evaluating performance using accuracy, confusion matrix, precision, recall, and F1-score metrics.

### 4) SHAP Explainability

To ensure model transparency, SHAP (SHapley Additive exPlanations) is integrated in this module. SHAP values explain the impact of each feature on individual predictions, helping interpret why the model classified an employee as "likely to leave." It provides both global and local interpretability, giving HR teams actionable understanding rather than just black-box predictions.

### 5) Dashboard Visualization

Insights generated from EDA and model outputs are converted into interactive visual dashboards using Power BI. This module presents key metrics such as attrition by department, overtime trends, and satisfaction breakdowns. The dashboards are dynamic, allowing HR users to filter views and explore the data visually to support evidence-based decision-making.

### 6) Report Generation & Recommendations

The final module compiles findings from all previous steps into a comprehensive report. It includes model evaluations, SHAP plots, and data summaries. Based on these insights, strategic recommendations are made for HR to reduce attrition, such as improving work-life balance, restructuring promotions, or targeting high-risk departments with tailored policies.

## 4.3 Project Implementation Timeline:

| Day | Task | Deliverables |
|---|---|---|
| **Day 1** | **Data Collection & Cleaning** | Load dataset, handle missing values, remove duplicates, and drop irrelevant columns. |
| **Day 2** | **Feature Engineering** | Encode categorical variables, scale numeric features, and finalize the input matrix. |
| **Day 3** | **Exploratory Data Analysis (EDA)** | Visualizations (bar charts, heatmaps), correlation matrix, initial attrition insights. |
| **Day 4** | **Model Training** | Train Logistic Regression & Decision Tree models, perform train/test split, and fine-tune parameters. |
| **Day 5** | **Model Evaluation** | Generate accuracy score, confusion matrix, classification report for both models. |
| **Day 6** | **SHAP Analysis** | SHAP summary plot, individual prediction explanation, feature importance chart. |
| **Day 7** | **Power BI Dashboard Creation** | Create and link visuals (attrition by job role, salary slab, satisfaction), filter setup. |
| **Day 8** | **Documentation & Final Report** | Prepare a full report (introduction, system design, results), conclusions, and HR suggestions. |

**TABLE 4.3.1 Project Implementation Timeline**

# 5. Conclusion

- This project has successfully demonstrated the value of applying data science and machine learning to a pressing HR challenge: predicting and understanding employee attrition. By using logistic regression and decision tree classifiers, we built predictive models capable of identifying which employees are most likely to leave the organization.
- What sets this work apart is not just its predictive power, but also its interpretability. The integration of SHAP (SHapley Additive exPlanations) allowed us to move beyond simple predictions and understand the *why* behind them. This form of explainable AI is critical for HR teams who need to make fair, informed, and transparent decisions. For example, we could clearly see how features such as overtime status, salary level, tenure, and job satisfaction directly influenced an employee's risk of leaving.
- Exploratory Data Analysis (EDA) also uncovered strong patterns—particularly the high attrition rates among employees who worked frequent overtime, earned low salaries, or reported low satisfaction in their job environment or relationships. These findings were essential for crafting data-backed HR strategies.
- The visual layer of this project—built using Power BI—transformed complex findings into interactive dashboards. This empowers HR personnel and managers to explore attrition trends, drill down by department or job role, and make data-informed decisions on the go.
- Ultimately, this project brings together three powerful components: predictive analytics, explainable AI, and intuitive business dashboards. Together, they form a comprehensive decision-support tool to address employee attrition proactively, reduce turnover, and enhance organizational resilience.

# 5.1 Suggestions for HR Action

1. **Implement Early Warning Systems:**Use predictive scores to flag employees at risk and start retention conversations early.

2. **Optimize Workload Distribution:** Monitor and reduce excessive overtime, especially in lower salary brackets.

3. **Improve Engagement for Long-Tenured Employees:** Offer lateral moves, mentorship programs, or leadership paths to avoid stagnation.

4. **Revise Compensation Strategies:**Benchmark salaries against industry standards and offer performance-based raises in critical roles.

5. **Promote a Healthy Work Culture:**Use relationship satisfaction surveys and emotional intelligence training for managers.

6. **Monitor KPIs via Dashboards:** Continue using Power BI for live HR insights on attrition, employee satisfaction, and department-level risk.

## 6. REFERENCES

Scikit-learn: Machine Learning in Python
https://jmlr.org/papers/v12/pedregosa11a.html

A Unified Approach to Interpreting Model Predictions (SHAP)
https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Kaggle Dataset: HR Analytics - Employee Attrition
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Python Official Documentation
https://docs.python.org/3/

Scikit-learn Documentation
https://scikit-learn.org/stable/

SHAP GitHub Repository
https://github.com/slundberg/shap

Matplotlib Documentation
https://matplotlib.org/stable/index.html

Seaborn Documentation
https://seaborn.pydata.org/

NumPy Documentation
https://numpy.org/doc/

Power BI Documentation
https://learn.microsoft.com/en-us/power-bi/

McKinsey Report: The Business Value of Retention in Tech Workforces (2022)
https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights