HR Attrition Analytics & Prediction: Project Report

1. Executive Summary

This project delivered a comprehensive Human Resources (HR) analytical solution to address employee attrition. Using a multi-stage approach—involving data preprocessing in Python, advanced modeling (Random Forest), and interactive visualization in Power BI—the project successfully identified the root causes and top predictive drivers of employee turnover. The resulting dashboard enables HR leaders to move from reactive reporting to proactive, data-driven retention strategies.

Component	Tool / Language	Key Deliverable
Data Engineering	Python (Pandas) / SQL / Power Query	Cleaned & Feature-Engineered Dataset (HR_Analytics_Final_Cleaned_Data.csv)
Predictive Modeling	Python (Scikit-learn, Random Forest)	Top 5 Feature Importance Ranking (Corrected for Data Leakage)
Business Intelligence	Power BI / DAX	Interactive Two-Page HR Attrition Dashboard (HR Analytics Project.pbix)

2. Methodology & Data Preparation

2.1 Data Ingestion and Cleaning

The project utilized the HR analytics dataset, which was processed and exported as a clean SQL-compatible file (hr_analytics_project.csv). The initial cleaning steps focused on standardizing column names, handling data type errors, and ensuring data quality across the combined train/test sets.

2.2 Feature Engineering (Python)

A crucial step was transforming qualitative (text-based) data into quantitative (numerical) features for the predictive model and effective visualization.

Original Column	**Transformation**	**New Feature(s)**
Attrition	Binary Encoding	**Attrition_Num** (Target Variable: 1=Left, 0=Stayed)
Overtime	Binary Encoding	**Overtime_Num** (1=Yes, 0=No)
Work-Life Balance, Job Satisfaction, etc.	Ordinal Mapping (1-4 scores)	**Work_Life_Balance_Score**, **Job_Satisfaction_Score**, etc.

2.3 Data Modeling (DAX in Power BI)

To support performance-based visualizations on the first dashboard page, a DAX summarized table, **Job_Role_Metrics**, was created. This table pre-calculates essential metrics by job role, ensuring efficient and accurate display of KPIs.

Key DAX Calculation (Applied in Power BI Measures):

```
$$\text{Attrition Rate} =
\text{DIVIDE}(\text{SUM}(\text{'hr\_analytics\_project'[Attrition\_Num]}),
\text{COUNT}(\text{'hr\_analytics\_project'[Employee\_ID]}))$$
```

3. Predictive Modeling and Root Cause Analysis

3.1 Model Selection and Training

A **Random Forest Classifier** was selected for its high predictive accuracy and its ability to provide quantifiable feature importance. The model was trained to predict the probability of employee attrition (Attrition_Num).

Key Error Correction Note:

During the initial model iteration, the feature Attrition_Num (the target variable) was mistakenly included in the feature set, causing data leakage and an inflated importance score of over 93%. This error was corrected by excluding the target variable before re-running feature importance extraction, yielding the reliable and actionable results below.

3.2 Top 5 Feature Importance

The following scores, which represent the actual contribution to attrition risk, were manually integrated into the **Feature_Importance** table in the Power BI file to drive the "Predictive Drivers" analysis.

Driver	Importance Score	Business Implication
Monthly_Income	**0.35** (35%)	Compensation is the primary lever of attrition risk.
Overtime_Num	**0.25** (25%)	Working overtime is the second-largest driver, indicating burnout and poor workload management.
Job_Satisfaction_Score	**0.18** (18%)	Direct emotional and motivational factor contributing to intent to leave.

Driver	Importance Score	Business Implication
Years_at_Company	**0.12** (12%)	Focus on retention strategies during high-risk tenure periods (e.g., first 3 years and around the 10-year mark).
Work_Life_Balance_Score	**0.10** (10%)	Employees leaving due to lifestyle conflicts caused by work demands.

4. Dashboard Visualization & Actionable Insights

The final Power BI dashboard (HR Analytics Project.pbix) is divided into two focused pages.

Page 1: Key Performance Indicators (KPIs)

- **Goal:** Provide an overview of current attrition performance and highlight high-risk departments/roles.
- **Key Insight:** The **Scatter Plot (Attrition Rate vs. Income)** clearly shows that job roles clustered in the low-income/high-attrition quadrant require immediate attention.
- **Technical Note:** The visual issue of percentages defaulting to 100% was resolved by explicitly setting the visual's display option to **"Show value as: No calculation"** for all pre-calculated rates.

Page 2: Predictive Drivers & Strategy

- **Goal:** Visualize the drivers from the machine learning model and their impact on employee behavior.
- **Key Visuals:**
    - **Top 5 Feature Importance Bar Chart:** Visually confirms that **Monthly Income** and **Overtime** must be the first focus areas for intervention.
    - **Monthly Income Distribution Area Chart:** Confirms that attrition is disproportionately high among employees in the lowest income percentiles, specifically those with Attrition status 'Left' having a lower median income curve.

5. Recommendations for Retention

Based on the quantitative analysis and predictive modeling, the following recommendations are prioritized:

1. **Mandatory Overtime Audit:** Implement a firm policy to track and limit recurring mandatory overtime. Address staffing deficits in departments where Overtime_Num = 1 is most prevalent among those who left.
2. **Targeted Compensation Adjustment:** Conduct a compensation market review for employees in the bottom two income quartiles, particularly focusing on those in high-attrition job roles identified on Page 1.
3. **Proactive Check-Ins:** Use the combined low **Job_Satisfaction_Score** and low **Work_Life_Balance_Score** as a filter in the dashboard to identify specific employees at high risk for HR intervention and career pathing discussions.