Employee Attrition Prediction Using Random Forest and Power BI

By: **S Lakshmi Priyanka**

1. Introduction

Employee attrition refers to the reduction of a company's workforce due to voluntary or involuntary resignations. It is one of the most critical challenges organizations face today, as high attrition can negatively impact productivity, employee morale, and company performance. With the availability of data, predictive analytics can be employed to foresee attrition trends and help HR departments take proactive measures.

This project aims to predict employee attrition using the **HR_Analytics.csv** dataset. It leverages **Random Forest**, a robust machine learning algorithm, and integrates data visualization through a **Power BI dashboard** to offer insightful and interactive analytics for better decision-making.

2. Abstract

We used a Random Forest classifier to predict whether an employee will leave the company (binary attrition label), training on demographic, job-related and performance variables drawn from the HR_Analytics.csv dataset (1,470 records, 35+ features). To address class imbalance (attrition rate ≈15%), we applied SMOTE oversampling to the minority class. After hyperparameter tuning via 5-fold GridSearchCV, the final model achieved:

- **Accuracy:** 80.5%
- **ROC AUC:** 0.71
- **Precision/Recall for "Will Leave":** 0.38 / 0.30 (f1 ≈ 0.34)

Feature-importance analysis (via SHAP) highlighted that **YearsSinceLastPromotion, OverTime, JobSatisfaction, MonthlyIncome** and **YearsAtCompany** are most predictive. We then built an interactive Power BI dashboard to explore attrition patterns across departments, roles, tenure buckets, income brackets and demographics.

3. Tools Used

- **Data manipulation & EDA:** pandas, numpy, matplotlib, seaborn
- **Preprocessing & modeling:** scikit-learn (LabelEncoder, StandardScaler, SelectKBest, RandomForestClassifier, GridSearchCV), imbalanced-learn (SMOTE)
- **Model interpretation:** SHAP
- **Model persistence:** joblib
- **Dashboard & visualization:** Power BI

4. Steps Involved in Building the Project

a. Data Loading & Initial Exploration

- Loaded HR_Analytics.csv into a pandas DataFrame.
- Dropped non-informative identifiers (EmployeeNumber, Over18, etc.).
- Examined distributions, missing values, class imbalance.

b. Preprocessing

- Encoded all categorical variables with LabelEncoder.
- Scaled numeric features using StandardScaler.
- Optional: univariate feature selection (SelectKBest with ANOVA F-test) to reduce noise.

c. Train/Test Split & Oversampling

- Split into train/test (e.g. 70/30).
- Applied SMOTE on the training set to balance the attrition class.

d. Modeling & Hyperparameter Tuning

- Initialized a RandomForestClassifier(random_state=42).
- Defined grid over n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features.
- Ran 5-fold GridSearchCV optimizing for ROC AUC.
- Re-trained Random Forest with best parameters on the full oversampled training data.

e. Evaluation

- Predicted on the held-out test set.
- Computed accuracy, ROC AUC, confusion matrix, precision/recall/f1.
- Examined the confusion matrix to understand false positives/negatives.

f. Model Interpretation

- Used SHAP to compute feature importances and visualize how each feature affects individual predictions.
- Identified top drivers of attrition.

g. Dashboard Development

- Imported the cleaned data plus model-predicted risk scores into Power BI.
- Built visuals to explore attrition rates by Department, JobRole, YearsSinceLastPromotion, IncomeBracket, Gender, Tenure, etc.
- Enabled slicers for demographics (MaritalStatus, Gender, OverTime) to filter and drill down.

5. Attrition Prevention Suggestions

a. Early Tenure Engagement

- Attrition is high among employees with <2 years of experience. Implement mentorship, career progression plans, and onboarding engagement strategies.

b. Department-Specific Interventions

- Sales and HR departments show higher attrition; consider workload balancing, improved recognition, or better incentives in these units.

c. Promotion Frequency

- Employees who have not been promoted in 0–3 years are at higher attrition risk. Establish a transparent and timely promotion pipeline.

d. Job Satisfaction & Training

- Moderate training hours and job satisfaction scores indicate potential burnout or disengagement. Regular training programs and feedback sessions should be implemented.

e. Income Brackets

- Employees earning in the ₹5k–₹10k and ₹10k–₹15k brackets show higher attrition. Ensure competitive compensation and benefits to retain talent.

6. Conclusion

By combining a robust Random Forest classifier with SHAP-driven explainability and an interactive Power BI dashboard, this project delivers both predictive power and actionable insights. Although the model attains around 80% accuracy, its true value lies in surfacing the **"why"** behind attrition—enabling HR leaders to intervene proactively. Future work could explore alternative algorithms (e.g. XGBoost, neural nets), richer feature engineering (e.g. text analysis of exit interviews), and closer integration into HR workflows for real-time risk management.