Employee Attrition Prediction using Random Forest

1. Dataset Overview
This dataset contains employee-related information from an HR system, aiming to predict whether an employee is likely to leave the company (Attrition).
It includes demographic data, job role, work environment, and performance-related variables.

Class Imbalance:
Attrition is imbalanced, with significantly more "No" than "Yes" cases.
- In the original data: Most employees stay (`Attrition = No`), only a small portion leave (`Attrition = Yes`).

2. Features in the Dataset:

Age
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
EducationField
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
Gender
HourlyRate
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked
Over18
OverTime
PercentSalaryHike

PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

3. Approach Summary
1. Data Cleaning & Preprocessing
   - Removed constant and ID columns.
   - Label encoded categorical variables.
   - Scaled numerical features with `StandardScaler`.

2. Model
   - Algorithm: Random Forest Classifier
   - First, trained on the original imbalanced dataset.
   - Then, improved using **SMOTE** (Synthetic Minority Oversampling Technique) to balance the classes.

4. Why SMOTE Was Used
When I first used a normal Random Forest on the imbalanced dataset:
- Accuracy** was good (~83%), but
- Recall for "Yes" (leavers) was very low (~10%).

Reason for low recall:
In imbalanced datasets, the model tends to predict the majority class (“No” for attrition) most of the time, because that alone gives high accuracy. As a result, it misses many of the actual “Yes” cases.

To fix this, I used SMOTE:
- SMOTE generates synthetic examples for the minority class ("Yes"), making the dataset balanced.
- This forced the model to learn patterns of employees who actually leave.

Impact of SMOTE:
- Recall jumped from ~10% to 90%.
- Model now catches almost all potential leavers.

## 5. Results

Without SMOTE (Original RF)

Accuracy 83.3%
Precision 41.7%
Recall 10.6%
F1 Score 16.9%
ROC-AUC 0.78

With SMOTE (Improved RF)

Accuracy 92.5%
Precision 94.9%
Recall 90%
F1 Score 92.3%
ROC-AUC 0.97

## 6. Key Insights

- Employees with lower monthly income have higher attrition.
- Certain job roles (e.g., Sales, Laboratory Technicians) see more attrition.
- Overtime and distance from home are potential attrition drivers.
- Balancing the data with SMOTE dramatically improves the ability to detect employees likely to leave.

## 7. Recommendations

- Focus retention strategies on high-risk job roles
- Consider salary adjustments for lower-paid employees.
- Monitor employees with frequent overtime and long commutes.
- Use the SMOTE-balanced Random Forest for ongoing attrition prediction.