# Employee Attrition Prediction

## Objective

The goal of this project is to build predictive models that determine whether an employee is likely to leave the organization. By identifying key factors contributing to attrition, companies can proactively improve employee retention and reduce turnover costs.

## Dataset Used

- **Source**: Kaggle - WA_Fn-UseC_-HR-Employee-Attrition
- **Size**: 1470 records
- **Target Variable**: Attrition (Yes/No)

**Features Categories:**

- **Demographics**: Age, Gender, MaritalStatus
- **Job-Related**: Department, JobRole, JobLevel, JobSatisfaction, YearsAtCompany
- **Compensation**: MonthlyIncome, DailyRate, PercentSalaryHike
- **Work-Life**: WorkLifeBalance, OverTime, BusinessTravel
- **Tenure Metrics**: YearsSinceLastPromotion, YearsWithCurrManager

## Model Chosen

We implemented and compared the following models:

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors (KNN)
- Gradient Boosting
- Support Vector Classifier (SVC)

Each model was evaluated using standard classification metrics.

## Performance Metrics

We evaluated model performance using:

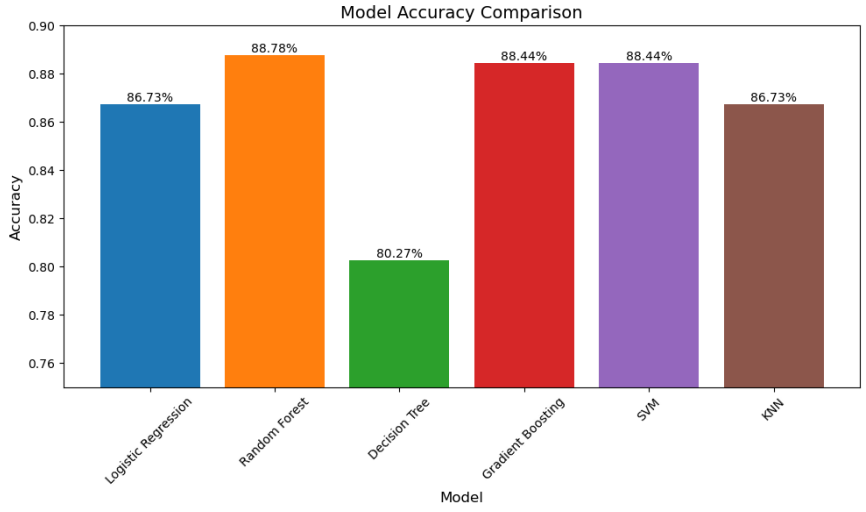
- **Accuracy**

- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**

| Model | Accuracy | Precision (Class 1) | Recall (Class 1) | F1-Score (Class 1) | Precision (Class 0) | Recall (Class 0) | F1-Score (Class 0) |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.8673 | 0.50 | 0.23 | 0.32 | 0.89 | 0.96 | 0.93 |
| Random Forest | 0.8878 | 1.00 | 0.15 | 0.27 | 0.89 | 1.00 | 0.94 |
| Decision Tree | 0.8027 | 0.31 | 0.38 | 0.34 | 0.90 | 0.87 | 0.88 |
| Gradient Boosting | 0.8844 | 0.63 | 0.31 | 0.41 | 0.90 | 0.97 | 0.94 |
| SVM | 0.8844 | 0.86 | 0.15 | 0.26 | 0.89 | 1.00 | 0.94 |
| KNN | 0.8673 | 0.50 | 0.10 | 0.17 | 0.88 | 0.98 | 0.93 |

- Class 1: Employees who left the company (minority class, more critical to predict).
- Class 0: Employees who stayed (majority class).
- Best Accuracy: Random Forest and Gradient Boosting are top performers.
- Best Recall (Class 1): Decision Tree does slightly better, but overall recall for class 1 is low across all models—suggests class imbalance issues.

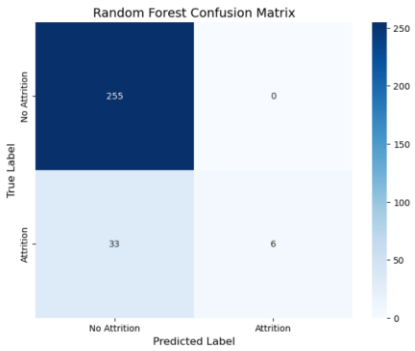
## Model Comparison Summary



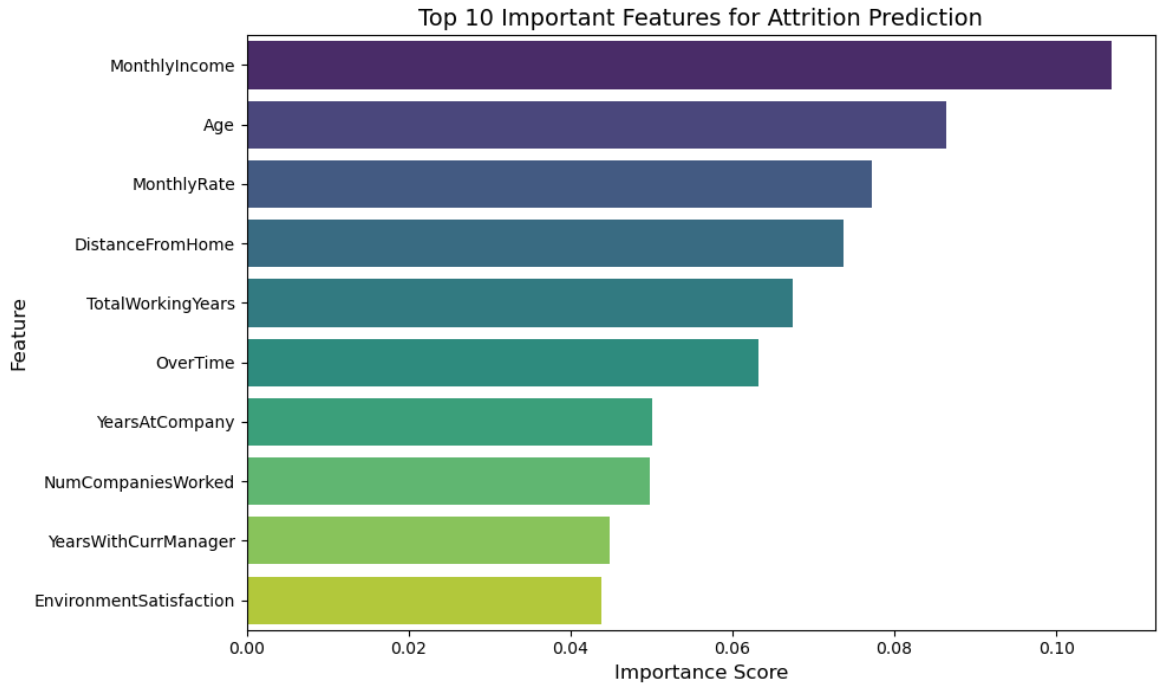
## Best Model Confusion Matrix

After this performace matrix we can conclude that best model is random forest with accuracy 88.78%.

### Confusion matrix of random forest:




## Feature Importance




---

## Challenges & Learnings

### Challenges

- **Class Imbalance**: The target variable (Attrition) was imbalanced, making it challenging for some models to detect the minority class.
- **Model Selection**: Choosing the best-performing model required careful cross-validation and metric evaluation.

- **Hyperparameter Tuning**: Required experimentation to optimize each model's performance.

**Learnings**

- **Tree-based models** like Random Forest and Gradient Boosting performed better in identifying attrition patterns.
- **Feature scaling** and **encoding** significantly affect model performance.
- Visualizations helped in understanding the underlying data distribution and model behavior.
- Using multiple models provides insights into which algorithm best fits the data characteristics.