Detailed Report: Predicting Employee Attrition with Machine Learning

1. Introduction

Employee attrition is a critical issue for organizations, leading to increased recruitment costs, loss of institutional knowledge, and reduced productivity. This project leverages machine learning (ML) to predict employee attrition risk using the IBM HR Analytics dataset. The goal is to identify key factors influencing turnover and develop a predictive model to help HR teams take proactive retention measures.

Key Objectives:

Perform Exploratory Data Analysis (EDA) to understand attrition patterns.

Preprocess data and engineer relevant features.

Train and evaluate multiple ML models.

Address class imbalance using techniques like SMOTE and PCA.

Deploy the best-performing model using Hugging Face for real-time predictions.

2. Project Overview

Problem Statement

High employee attrition negatively impacts business operations. Traditional reactive approaches (exit interviews, post-attrition analysis) are insufficient. A predictive model can help identify at-risk employees early.

Solution

Build an ML model to predict attrition likelihood.

Provide insights into key factors driving turnover.

Enable HR to implement targeted retention strategies.

Dataset

IBM HR Analytics Employee Attrition & Performance Dataset

Contains 1,470 employee records with 35 features (demographics, job-related factors, satisfaction levels).

Target Variable: Attrition (Yes/No).

Focus Areas

Exploratory Data Analysis (EDA) – Identify trends and correlations.

Feature Engineering – Select and transform key predictors.

Model Development – Compare Logistic Regression, XGBoost, Decision Tree, Random Forest, and SVM.

Handling Class Imbalance – Apply SMOTE and PCA.

Model Deployment – Host on Hugging Face for real-time predictions.

3. Exploratory Data Analysis (EDA)

Key Insights from the Dashboard (Built with Python Dash)

Attrition Rate: 16% of employees leave.

Gender Distribution: 60% male, 40% female.

Overtime Impact: Employees working overtime are more likely to leave.

Department Analysis:

Research & Development (R&D) has the most employees but the lowest attrition rate.

Sales and HR show higher attrition.

Other Findings:

Employees with lower job satisfaction are more likely to leave.

Higher monthly income correlates with lower attrition.

4. Data Preprocessing & Feature Engineering

Steps Taken:

Data Cleaning

Handle missing values (if any).

Remove outliers.

Feature Selection

Use correlation analysis and feature importance to select key predictors.

Examples: Overtime, MonthlyIncome, JobSatisfaction, YearsAtCompany.

Class Imbalance Handling

Attrition (Yes) = 16% (minority class).

Techniques applied:

No Balancing (Baseline)

SMOTE (Oversampling)

PCA + SMOTE (Dimensionality Reduction + Oversampling)

Train-Test Split

80% training, 20% testing for model validation.

5. Machine Learning Models & Performance

Five models were trained and evaluated:

Logistic Regression (Baseline)

XGBoost (Gradient Boosting)

Decision Tree (Interpretable)

Random Forest (Ensemble Method)

Support Vector Classifier (SVC)

Evaluation Metrics:

AUC-ROC (Area Under the Curve)

F1-Score (Balances Precision & Recall)

Precision (Minimize False Positives)

Recall (Minimize False Negatives)

Results Summary

Model	AUC	F1-Score	Precision	Recall
Logistic Regression (Imbalanced)	0.89	0.45	0.91	0.68
XGBoost (Oversampling)	0.85	0.39	0.88	0.58
Decision Tree (Imbalanced)	0.78	0.32	0.90	0.27
Random Forest (Oversampling + PCA)	0.82	0.42	0.89	0.44
SVC (Imbalanced)	0.87	0.40	0.90	0.50

Key Observations:

Logistic Regression performed well but suffered from low recall (missed true attrition cases).

XGBoost improved AUC but had lower precision.

Random Forest + PCA + SMOTE provided a balanced performance.

Ensemble Model (Stacking Classifier) was the best overall:

AUC: 0.80

F1-Score: 0.52

Precision: 0.93

Recall: 0.43

6. Deployment with Hugging Face

Why Hugging Face?

Simplifies model hosting and API integration.

Ensures scalability and security.

Steps Taken:

Model Upload – Trained ensemble model exported and hosted.

API Integration – REST API allows real-time predictions from web/mobile apps.

Security Compliance – Data encryption and secure endpoints.

Web App Features:

Input employee details (job role, satisfaction, overtime, etc.).

Output: Attrition Risk Score (High/Medium/Low).

7. Conclusion & Future Work

Key Takeaways:

✅ Ensemble Model (Stacking) performed best with balanced metrics.

✅ Overtime, Job Satisfaction, and Income are critical predictors.

✅ SMOTE + PCA helped mitigate class imbalance.

Future Improvements:

Incorporate more employee behavioral data (e.g., performance reviews).

Experiment with Deep Learning models (Neural Networks).

Expand deployment to HR analytics dashboards.

8. Team Members

Yousef Khaled Shawkyy

Fares Essam Mostafa

Ali Fathy Abdelghany

Amr Sabry Awad

Abdelrahman Mohamed Abdelrazek