Name of Student:	James Bobby Kiawu
Submitted To:	Ratinder Rajpal, Lecturer.
Willis College:	Introduction to Artificial Intelligence
Intro AI	ML Final Project Report
Date:	November 16, 2025

1. Introduction

- Why predicting employee attrition matters
- Purpose and scope of the project
- Machine learning approach applied

2. Dataset Overview

- Dataset source and size
- List of key features (demographics, job-related, satisfaction metrics)
- Target variable description
- Note on class imbalance

3. Methodology

- Data preprocessing steps
- Handling missing values
- One-Hot Encoding of categorical variables
- Scaling numerical features
- Train–test split strategy
- Tools and libraries used
- Exploratory Data Analysis (EDA)
- Summary of observed trends and insights

4. Model Development and Evaluation

- Models used: Logistic Regression, Random Forest, SVM
- Rationale for choosing each model
- Evaluation metrics
- Accuracy
- Precision, Recall, F1-score
- Confusion Matrix
- ROC Curve & AUC score
- Key results and best-performing model (Random Forest)
- Feature importance interpretation

5. Model Optimization

- Hyperparameter tuning with Grid Search
- Regularization testing for Logistic Regression
- Feature selection approaches
- Impact of optimization on model accuracy and performance

6. Model Deployment

- Saving and loading the final model (joblib)
- Gradio interface implementation in Google Colab
- Walkthrough of user input → prediction flow

- Real-world applicability and use-case scenario

---

7. Conclusion

- Summary of findings
- Final model selection
- Practical implications for HR retention strategies
- Limitations and future improvement opportunities
-

---

8. References

- Dataset source
- Libraries & documentation
- Supplementary resources

FINAL MACHINE LEARNING PROJECT REPORT

Employee Attrition Prediction Using Machine Learning

Executive Summary

This project uses machine learning to predict employee attrition using the IBM HR Employee Attrition dataset. The objective was to build a model capable of identifying employees likely to leave an organization, allowing leaders to implement proactive retention strategies. Three models were evaluated—Logistic Regression, Random Forest, and SVM—with the Random Forest Classifier delivering the strongest performance based on accuracy, precision/recall, and ROC-AUC. The model identified key attrition drivers such as Overtime, Monthly Income, Job Satisfaction, Environment Satisfaction, and Work Life Balance.

After hyperparameter tuning using Grid Search, the final model was deployed through a Gradio interface in Google Colab. This allowed users to input employee details and receive real-time predictions. The project demonstrates a complete end-to-end ML workflow including preprocessing, EDA, modeling, optimization, and deployment.

1. Introduction

Employee attrition is a recurring challenge that impacts organizational performance and culture. Predicting attrition can help HR teams intervene early and reduce costly turnover. This project applies supervised classification models to forecast which employees are likely to leave, using demographic, job-related, and satisfaction-related variables.

2. Dataset Overview

Dataset: IBM HR Analytics Employee Attrition Dataset[1]
Size: 1,470 records and 35+ features
Target: Attrition (Yes/No)

The dataset includes variables such as age, salary, job role, overtime, commute distance, satisfaction metrics, tenure, performance ratings, and more. Attrition is imbalanced, with fewer "Yes" cases.

3. Methodology

Data Preprocessing

- Checked for missing values (none found).
- Applied One-Hot Encoding to categorical attributes.
- Standardized numerical columns using StandardScaler.
- Split data into 80% train / 20% test (stratified).

Exploratory Data Analysis

Key patterns:

- Employees working frequent overtime showed significantly higher attrition rates.

- Lower job satisfaction and environment satisfaction were strongly linked to leaving.
- Lower income and longer commute distances modestly increased attrition risk.

4. Model Development and Evaluation

Three models were trained:

Logistic Regression

- Served as baseline
- Moderate performance
- Struggled predicting minority class (attrition)

Support Vector Machine (SVM)

- Performed better than Logistic Regression
- Good for high-dimensional data
- Slightly less accurate than Random Forest

Random Forest Classifier (Best Model)

- Highest accuracy and balanced precision/recall
- Strong ROC-AUC (0.80)
- Lower false negative rate (important for identifying resignations)
- Provided clear feature importance insights

Top Predictive Features:

- Overtime
- Monthly Income
- Job Satisfaction
- Environment Satisfaction
- Work Life Balance

5. Hyperparameter Tuning

Grid Search optimized Random Forest parameters such as:

- n_estimators
- max_depth
- min_samples_split

- min_samples_leaf

This improved performance and model stability.

6. Deployment

The final Random Forest model was saved using joblib and deployed in Google Colab using Gradio. The interface allows users to input employee characteristics (age, income, overtime, commute distance, satisfaction score) and receive instant predictions ("Likely to Leave" or "Likely to Stay"). Deployment demonstrates how ML models can integrate into HR decision workflows.

7. Conclusion

This project successfully built and deployed a machine learning model that predicts employee attrition with high accuracy. Random Forest proved to be the strongest model due to its ability to handle complex patterns, its robustness to noise, and its interpretability through feature importance.

The model highlights actionable insights for employers:

- Reduce excessive overtime
- Improve satisfaction metrics
- Support work-life balance
- Address income-related concerns

While the project produced strong results, future improvements may include testing additional ensemble models (XGBoost, LightGBM), applying SMOTE for class imbalance, integrating external factors (engagement surveys, performance trends), or deploying the model into a full web application.

Overall, this work demonstrates the full end-to-end machine learning pipeline—from data prep to deployment—and delivers a practical tool that organizations can leverage to improve retention and workforce planning.

References:

- Scikit-learn Documentation
- Pandas Documentation
- Matplotlib & Seaborn Libraries
- Kaggle: IBM HR Analytics Employee Attrition Dataset
- Gradio Documentation

1. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data