****

**Assessment Report**

on

**“Employee Attrition Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)(SEC-C)**

By

Group NO: 1

Group members:

Saksham Singh – 202401100400161

Shikha Tomar – 202401100400174

Sujal Kumar – 202401100400193

Utkarsh Pandey- 202401100400203

Vaishnavi Rai – 202401100400205

Yashash Tyagi – 202401100400218

**Under the supervision of**

“MR. ABHISHEK SHUKLA”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

Employee attrition affects organizational productivity, morale, and costs. Anticipating which employees might leave enables HR departments to take proactive steps. This project leverages IBM HR Analytics data to build a machine learning model using classification techniques, focusing on Random Forest, to predict attrition and identify key contributing factors.

**2. Problem Statement**

To build a classification model that predicts whether an employee is likely to leave the company based on features like age, job role, satisfaction level, overtime status, and more, using IBM's HR Analytics dataset.

**3. Objectives**

 Load and preprocess the IBM HR dataset.

 Encode categorical data using LabelEncoder.

 Train a RandomForestClassifier to predict attrition.

 Evaluate performance using metrics like accuracy, confusion matrix, and classification report.

 Visualize the top 15 important features influencing attrition.

**4. Methodology**

**Data Collection:  
IBM HR Analytics dataset (WA\_Fn-UseC\_-HR-Employee-Attrition.csv)**

**Data Preprocessing:**

* **Dropped non-informative columns: EmployeeCount, Over18, StandardHours, EmployeeNumber.**
* **Encoded all categorical features using LabelEncoder.**
* **Defined target column: Attrition.**
* **Split data using train\_test\_split (80% training, 20% testing, stratified).**

**Model Training:  
Used RandomForestClassifier with 100 estimators and random\_state=42.**

**Model Evaluation:**

* **Used accuracy\_score, classification\_report, and confusion\_matrix.**
* **Visualized top 15 feature importances with a Seaborn bar plot.**

**5. Data Preprocessing**

 Label encoding was applied to all object-type (categorical) columns.

 Target (Attrition) and features were separated.

 The data was split into training and testing sets in an 80-20 ratio with stratification for balanced class distribution.

**6. Model Implementation**

 Trained RandomForestClassifier on preprocessed training data.

 Model predictions were generated on the test set.

 Feature importances were extracted and visualized using a bar chart.

**7. Evaluation Metrics**

 **Accuracy Score:**  
Achieved via accuracy\_score(y\_test, y\_pred).

 **Classification Report:**  
Provided precision, recall, and F1-score for each class.

 **Confusion Matrix:**  
Displayed true positives, false positives, etc.

 **Feature Importance Plot:**  
Top 15 most influential features visualized.

**8. Results and Analysis**

* The model showed balanced performance on both classes, identifying key attrition patterns.
* **Top Important Features:** OverTime, JobLevel, MonthlyIncome, Age, DistanceFromHome.
* The confusion matrix helped assess false negatives (important for HR).
* Feature analysis suggested overtime and job level strongly correlate with attrition.

**9. Conclusion**

The Random Forest model predicted employee attrition effectively and revealed important insights. HR teams can use this data to identify employees at risk and implement retention strategies. Future work could explore SMOTE for class imbalance and try other ensemble models like XGBoost.

**10. References**

* IBM HR Analytics Employee Attrition Dataset
* scikit-learn Documentation
* pandas Documentation
* Seaborn Documentation
* Research papers on Employee Churn Prediction



 

 