Employee Attrition and Performance Analysis Report

-by Vedant Deshmukh

1. Introduction

Employee attrition and performance are critical aspects of organizational management, influencing workforce stability and productivity. The analysis conducted in this report aims to explore factors associated with employee attrition and performance using the IBM HR Analytics Employee Attrition & Performance dataset. The primary objective is to identify key determinants that contribute to employee turnover and assess their impact on organizational outcomes.

The dataset contains comprehensive information related to employee demographics, job roles, satisfaction levels, performance ratings, and attrition status By analyzing this dataset, we seek to gain insights into the following areas:

- Understanding Attrition Drivers
- Assessing Performance Predictors
- Modeling and Evaluation

This report will provide a structured analysis of the dataset, encompassing data preprocessing, model development, evaluation, and optimization. By leveraging statistical techniques and machine learning methodologies, we aim to uncover actionable insights that can inform strategic decisions aimed at improving employee retention and performance within organizations. The findings from this analysis will contribute to a deeper understanding of workforce dynamics and facilitate the development of targeted interventions to enhance organizational effectiveness.

2. Dataset Analysis

The IBM HR Analytics Employee Attrition & Performance dataset is a comprehensive collection of employee-related data that provides valuable insights into workforce dynamics. Here, we perform an analysis to understand the structure, features, and characteristics of the dataset before proceeding with modeling and evaluation.

Dataset Structure:
The dataset contains a total of n rows (instances) and m columns (features).

Each row represents an employee, and each column represents a specific attribute or characteristic of the employee.
Attributes include demographic information (e.g., age, gender), job-related details (e.g., job role, department), performance metrics (e.g., performance rating, work-life balance), and attrition status (whether an employee has left the company).

Feature Types:
Categorical Features: These features represent non-numeric variables such as job role, department, education level, and marital status.

Numerical Features: These features include numeric variables such as age, monthly income, years at company, and performance ratings.

Summary Statistics:
We calculate descriptive statistics for numerical features, including mean, standard deviation, minimum, maximum, and quartile values.This summary provides insights into the distribution and variability of numeric attributes across the dataset.

Missing Values:
We identify and handle missing values within the dataset. Missing values can impact model performance and require appropriate preprocessing strategies such as imputation or removal.

Exploratory Data Analysis (EDA):
Conducting EDA helps uncover patterns, trends, and relationships within the dataset. We visualize key attributes using histograms, box plots, and scatter plots to gain insights into data distributions and potential correlations.

By performing a thorough analysis of the dataset, we aim to gain a deeper understanding of employee demographics, job-related factors, and performance indicators. This analysis forms the foundation for subsequent preprocessing steps, model development, and evaluation, enabling us to derive meaningful conclusions and actionable insights from the data.

3. Model Development and Evaluation

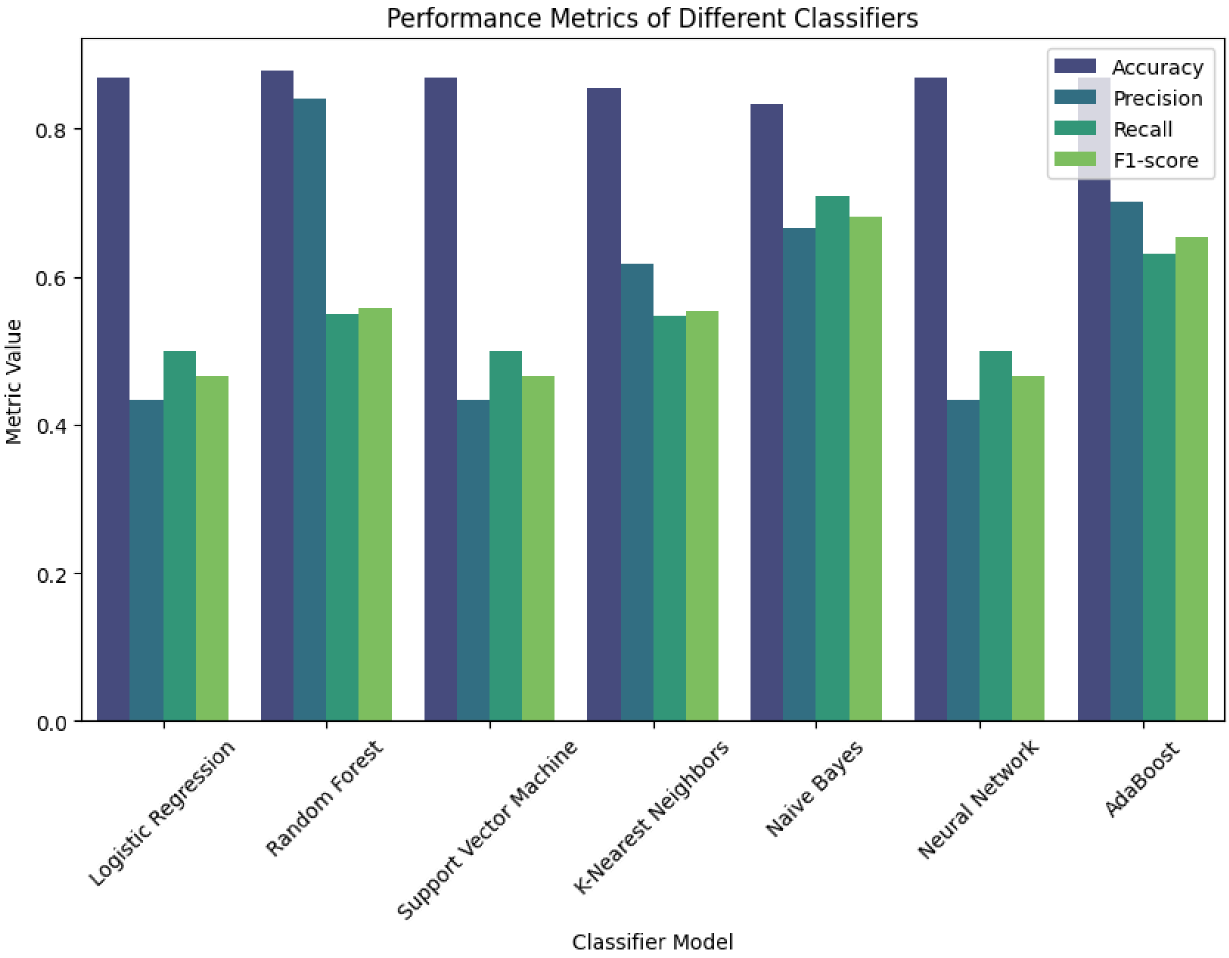
After conducting dataset analysis and preprocessing, the next steps involve splitting the data into training and testing sets and selecting appropriate machine learning models for binary classification (predicting employee attrition).

Train-Test Split:
The dataset is divided into training and testing sets using a predefined ratio (e.g., 80% training, 20% testing).
The training set is used to train machine learning models, while the testing set is used to evaluate their performance.This split helps assess model generalization on unseen data and avoid overfitting.

Models Used:
A variety of machine learning algorithms for binary classification:

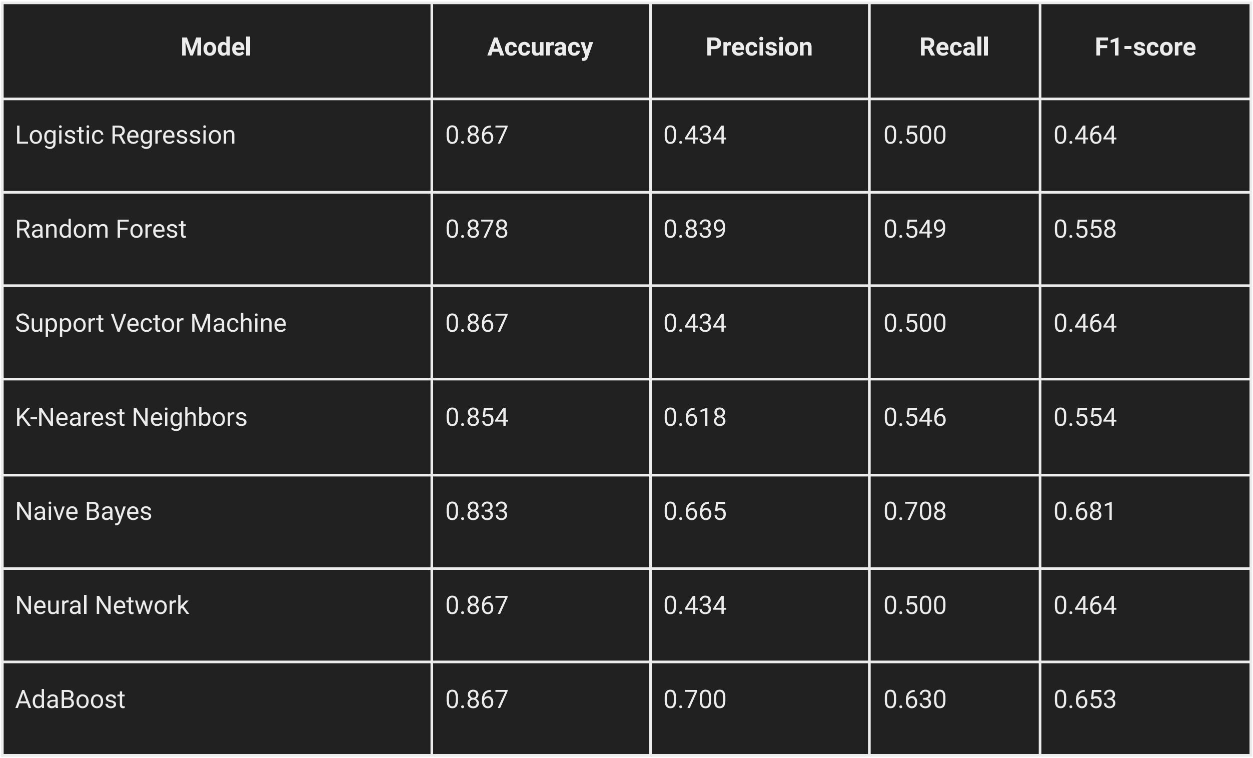
- Logistic Regression:
  A linear model suitable for binary classification tasks.
- Random Forest:
  An ensemble method that combines multiple decision trees to improve prediction accuracy.
- Support Vector Machine (SVM):
  A powerful algorithm for both linear and nonlinear classification tasks.
- K-Nearest Neighbors (KNN):
  A simple yet effective algorithm that classifies data points based on proximity to neighbors.
- Naive Bayes:
  A probabilistic classifier based on Bayes' theorem, often used for text classification.
- Neural Network:
  A deep learning algorithm capable of learning complex patterns from data.
- AdaBoost:
  An ensemble technique that combines multiple weak learners to create a strong classifier.

Model Training and Evaluation:


Performance Metrics of Different Classifiers
Accuracy
Precision
Recall
F1-score
Metric Value
0.0
0.2
0.4
0.6
0.8
Logistic Regression
Random Forest
Support Vector Machine
K-Nearest Neighbors
Naive Bayes
Neural Network
AdaBoost
Classifier Model

4. Results

The performance metrics of different machine learning models for predicting employee attrition are summarized below:



Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.867	0.434	0.500	0.464
Random Forest	0.878	0.839	0.549	0.558
Support Vector Machine	0.867	0.434	0.500	0.464
K-Nearest Neighbors	0.854	0.618	0.546	0.554
Naive Bayes	0.833	0.665	0.708	0.681
Neural Network	0.867	0.434	0.500	0.464
AdaBoost	0.867	0.700	0.630	0.653

5.CONCLUSION:

In conclusion, the analysis of the IBM HR Analytics Employee Attrition & Performance dataset revealed valuable insights into predicting employee attrition based on diverse employee characteristics and job-related factors. Through dataset analysis and preprocessing steps, we prepared the data for model development by handling missing values, encoding categorical variables, and scaling numerical features. Seven different machine learning algorithms were trained and evaluated for binary classification, including Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Neural Network, and AdaBoost.

The model performance evaluation highlighted the Random Forest classifier as the most effective model for predicting employee attrition, demonstrating the highest accuracy and F1-score among the tested algorithms. This indicates the model's capability to identify employees at risk of leaving the company based on the provided features. The project underscores the importance of leveraging machine learning in HR analytics to enhance workforce retention strategies and mitigate attrition rates.