Table of Contents

[Introduction 2](#_Toc173358647)

[1. Data Preparation and Cleaning 3](#_Toc173358648)

[2. Exploratory Data Analysis (EDA) 4](#_Toc173358649)

[3. Principal Component Analysis (PCA) 5](#_Toc173358650)

[4. Machine Learning 6](#_Toc173358651)

[5. Discussion of Findings 7](#_Toc173358652)

[Conclusion 8](#_Toc173358653)

# Introduction

Understanding and predicting employee attrition is crucial for organizations seeking to manage their workforce effectively. This project delves into various techniques for data preparation, exploratory data analysis (EDA), and machine learning to forecast employee attrition. The objective is to identify key factors influencing attrition and develop accurate predictive models.

# 1. Data Preparation and Cleaning

**1. Data Import and Initial Examination**

To begin, essential libraries for data analysis were imported, and the dataset, comprising 1,470 rows and 35 columns, was loaded into the environment. Initial examinations ensured data integrity by checking variable names and data types.

Missing Values:

The dataset contained 29 missing values. To address this, rows with missing data were removed using the na.omit function, resulting in a cleaner dataset for subsequent analysis.

**2. Data Transformation**

Categorical variables were converted into factors using R's factor function, which facilitated accurate analysis and visualization. Additionally, categorical variables were transformed into numerical codes where necessary to prepare the data for machine learning models.

**3. Exporting Clean Data**

The cleaned dataset was exported to an Excel file using the writexl package, ensuring that the data was available for further use and sharing.

# 2. Exploratory Data Analysis (EDA)

**1. Distribution of Numerical Variables**

Histograms were created to analyze the distribution of numerical variables such as Age, MonthlyIncome, TotalWorkingYears, and YearsAtCompany. These visualizations revealed that most numerical variables were right-skewed, indicating a concentration of values at the lower end of the scale.

**2. Distribution of Categorical Variables**

Using ggplot2, we visualized the distribution of categorical variables. Bar charts and stacked bar charts compared the frequency of categories in relation to attrition. This analysis highlighted patterns and potential predictors of employee turnover.

**3. Attrition vs Numerical Variables**

We employed ggplot2 to create density plots, box plots, and scatter plots to compare numerical variables relative to attrition status. These plots helped identify significant trends or differences in variable distributions between employees who left and those who stayed.

**4. Chi-Square Test**

A Chi-Square test was conducted to determine if there was a significant association between attrition and categorical variables. This statistical test assessed whether observed frequencies deviated significantly from expected frequencies, aiding in the identification of key factors associated with attrition.

# 3. Principal Component Analysis (PCA)

PCA was performed to simplify the dataset and identify key components influencing variance. The following steps were undertaken:

* Standardization: The data was standardized to ensure comparability.
* Covariance Matrix Computation: A covariance matrix was computed to analyze the variance and correlation between variables.
* Matrix Decomposition: The covariance matrix was decomposed into eigenvalues and eigenvectors.

Biplots were used to visualize the principal components, and bar plots displayed the loadings of the first three principal components. PCA was instrumental in reducing dimensionality and identifying the most influential factors affecting employee attrition.

# 4. Machine Learning

**1. Neural Network Models**

Neural network models were trained to predict employee attrition based on various features. Different architectures with varying numbers of hidden layers were evaluated:

* Model 1: One hidden layer with an accuracy of 78.33%.
* Model 2: Two hidden layers with an accuracy of 93.57% (best performance).
* Model 3: Five hidden layers with an accuracy of 83.33%.

The second model, with two hidden layers, demonstrated the highest accuracy, effectively capturing complex patterns and interactions in the data.

**2. Comparison with Other Methods**

The neural network models were compared with other classification methods:

* Support Vector Classifier: Accuracy of 85.22%
* Random Forest Classifier: Accuracy of 84.99%
* Decision Tree Classifier: Accuracy of 77.90%
* Logistic Regression: Accuracy of 87.76%

The neural network model, particularly the one with two hidden layers, outperformed other methods, achieving superior accuracy.

High-Performance Computational Implementation

In addition to R-based models, a Multilayer Perceptron Classifier was implemented in PySpark:

* Model Accuracy: 82.34%

This model utilized VectorAssembler for feature aggregation and StandardScaler for feature scaling. Although the PySpark model performed well, the R neural network model with two hidden layers provided better accuracy.

# 5. Discussion of Findings

The analysis demonstrated that neural network models, especially those with multiple hidden layers, provided the most accurate predictions for employee attrition. PCA played a crucial role in identifying key factors influencing attrition, while EDA and statistical tests offered valuable insights into data distribution and variable relationships.

# Conclusion

Predicting employee attrition is vital for effective organizational planning and retention strategies. This project highlighted that advanced machine learning techniques, particularly neural networks, are highly effective in forecasting attrition. The integration of EDA, PCA, and machine learning provided a comprehensive approach to understanding and predicting employee turnover, offering valuable insights for improving workforce management.

**This report summarizes the methodology and findings of the employee attrition analysis project. The combination of data preparation, exploratory analysis, PCA, and machine learning has proven to be a robust approach to predicting and understanding employee turnover, ultimately supporting more informed decision-making.**