

**Usman Institute of Technology University**

**Department of Computer Science Spring 2025**

CIC-201 – Artificial Intelligence

Project Report

Members:

|  |  |  |
| --- | --- | --- |
| Name | Roll No. | Section |
| Sarmad Khan | 23SP-007-CS | A |
| Talal Aslam | 23SP-008-CS |
| Shaheer Imran | 23SP-031-CS |
| Abdul Mannan Ansari | 23SP-033-CS |

Contents

[Abstract 3](#_Toc200918109)

[Introduction 3](#_Toc200918110)

[Objectives 4](#_Toc200918111)

[Problem Statement 5](#_Toc200918112)

[Methodology 6](#_Toc200918113)

[Data Loading and Cleaning: 6](#_Toc200918114)

[Label Encoding: 6](#_Toc200918115)

[Feature Engineering: 7](#_Toc200918116)

[Visualization: 7](#_Toc200918117)

[Model Training: 8](#_Toc200918118)

[GUI Application: 8](#_Toc200918119)

[Report Generation: 10](#_Toc200918120)

[Results 11](#_Toc200918121)

[Accuracy: 11](#_Toc200918122)

[Visualization Insights: 11](#_Toc200918123)

[GUI Functionality: 11](#_Toc200918124)

[Report Quality: 12](#_Toc200918125)

[Conclusion 13](#_Toc200918126)

[Figure 1 - Dataset used to train the model 4](#_Toc200918127)

[Figure 2 - An example of a stacked histogram generated using seaborn 7](#_Toc200918128)

[Figure 3 - Correlation Heatmap 8](#_Toc200918129)

[Figure 4 - GUI main screen 9](#_Toc200918130)

[Figure 5 - Prediction Result(1) 9](#_Toc200918131)

[Figure 6 - Prediction Result(2) 10](#_Toc200918132)

[Figure 7 - Report generation 10](#_Toc200918133)

[Figure 8 - Report showing model accuracy 11](#_Toc200918134)

[Figure 9 - An example of a visual tool useful for reviews 12](#_Toc200918135)

# Abstract

Employee attrition is one of the most pressing challenges faced by organizations. Retaining talent is important for growing companies as the loss of human capital can make its recovery a lot more challenging. In this project, we leverage machine learning techniques to build a model capable of analyzing employee data and predicts the likelihood of attrition. By working with IBM's open-source HR dataset, we perform an end-to-end analysis that includes preprocessing, data visualization, feature engineering, and predictive modeling. This trained model is then deployed as a user-friendly graphical user interface (GUI) which enables HR professionals to input employee attributes and receive real-time predictions. Additionally, the system automatically generates an in-depth statistical report in PDF format containing visual insights and performance metrics of the model.

This system uses Random Forest Classifier due to its robustness and suitability for handling categorical and numerical data. Advanced feature engineering techniques are used to extract meaningful insights from raw data, such as loyalty index, experience level, and income categories. These features are enhancers to model performance and also offer actionable insights into the factors influencing employee behavior. The application integrates visual components, machine learning, and interactive design, aiming to bridge the gap between AI models and non-technical end-users.

# Introduction

Attrition affects nearly every organization, regardless of size or industry. The departure of talented and valuable employees can disrupt workflow, lower morale, and increase recruitment and training costs. Traditionally, understanding why employees leave required manual HR audits and subjective analysis, which were neither scalable nor data-driven. Today, with access to structured HR data and advances in machine learning, we can predict attrition patterns more accurately.

In this project, we explored the application of artificial intelligence in solving real-world HR problems. The dataset used contains details about employees, such as job role, department, salary, tenure, distance from home, marital status, and more. These variables were used to train a predictive model which could estimate the risk of attrition for any given employee.

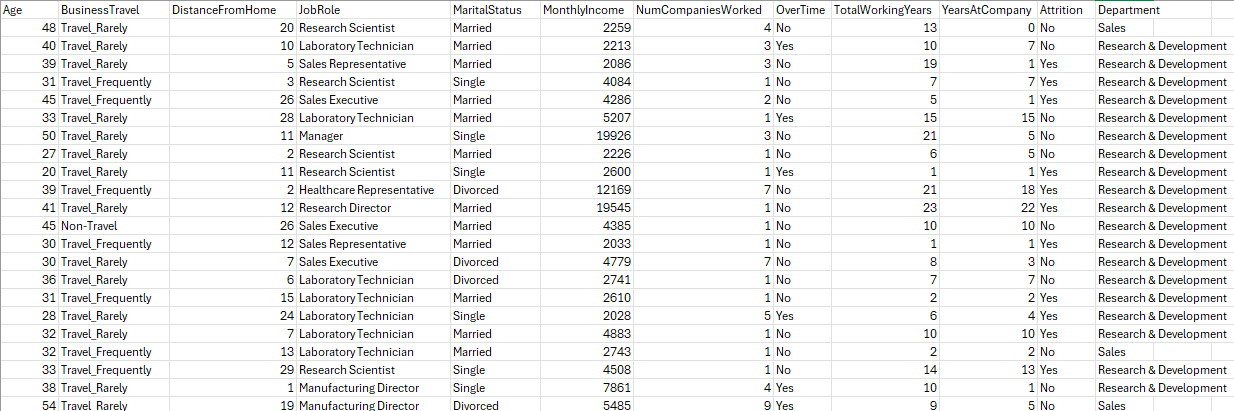


Figure - Dataset used to train the model

By integrating data science techniques and a visual interface, the project provides insights that help organizations take preemptive actions to reduce turnover. The GUI was designed to be accessible for HR personnel without programming skills, making it practical for everyday use.

# Objectives

The primary objective of this project is to provide an AI-powered decision-support system for HR professionals. The following sub-goals were outlined and achieved:

1. Understand and analyze employee behavior and its correlation with attrition.
2. Preprocess real-world HR data by handling missing values and encoding categorical variables.
3. Apply data visualization techniques to highlight key trends and distributions.
4. Engineer high-quality features to improve predictive model performance.
5. Build a machine learning model that accurately classifies employees at risk of attrition.
6. Develop a graphical user interface to make the model accessible to non-technical users.
7. Automate report generation for continuous monitoring and documentation.

# Problem Statement

Employee attrition is a critical challenge that impacts organizational performance and financial stability. The departure of skilled employees results in increased hiring costs, loss of organizational knowledge, disruption of workflows, and a decrease in team morale. Most organizations struggle to identify potential attrition risks early to intervene effectively. Traditional HR systems rely heavily on manual evaluations and qualitative assessments, which are often inconsistent, biased, and reactive in nature.

The growing availability of structured HR data provides a unique opportunity to apply machine learning techniques to predict employee attrition and understand the underlying patterns. However, several problems arise in the practical implementation of such a system:

* HR data is usually complex, consisting of both categorical and numerical variables.
* There are too many features that affect attrition like loyalty, tenure, income stability. These are not directly measurable and need to be derived.
* Models must be interpretable and accessible to non-technical HR staff for real-world application.
* HR teams want visual insights and reports for strategic planning, not just raw predictions.

Our project addresses these gaps by building an automated, intelligent system that can process raw HR data, extract meaningful features, and train a high-performing model to classify employees as ‘likely to stay’ or ‘likely to leave’. This prediction is presented via a user-friendly GUI and is supported by detailed analytical reports that visualize key trends and validate model accuracy.

# Methodology

This project follows a systematic data science workflow, from data acquisition to model deployment.

## Data Loading and Cleaning:

* We use the *load\_data()* function from *AIProjekt.py* to import the CSV file using pandas. This function reads the HR data into a DataFrame.
* Irrelevant or redundant columns such as Age, DistanceFromHome, etc., are removed in the *clean\_data()* function to reduce noise.
* Missing values are filled with either the median or mode, depending on the nature of the column.

## Label Encoding:

* The *encode\_categorical()* function uses *LabelEncoder()* to convert categorical values into numerical representations. This transformation is vital since *scikit-learn* models work with numeric inputs only.

## Feature Engineering:

* *feature\_engineering()* creates new columns derived from existing ones. For example:
  + - *LoyaltyIndex = YearsAtCompany / TotalWorkingYears*
    - *ExperienceLevel* = Categorized *TotalWorkingYears* as Junior, Mid, or Senior
    - *IncomePerYearWorked = (MonthlyIncome \* 12) / TotalWorkingYears*
  + These derived features add domain knowledge into the data, which improves model understanding.

## Visualization:

* + The *plot\_initial\_visualizations()* function generates stacked histograms using Seaborn. For instance, age distribution and income distribution are plotted against attrition.
  + A heatmap is generated to observe correlations among numerical variables.

A graph of different colored bars

AI-generated content may be incorrect.

Figure - An example of a stacked histogram generated using seaborn

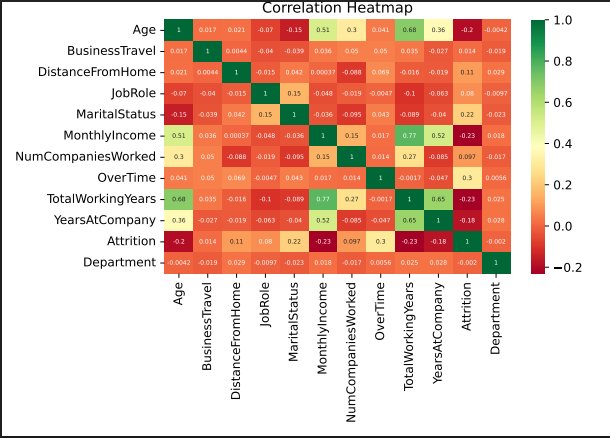


Figure - Correlation Heatmap

## Model Training:

* + The *train\_model()* function splits the data using *train\_test\_split()* and applies a *RandomForestClassifier* with *class\_weight='balanced'* to counter class imbalance.
  + Model performance is evaluated using accuracy score, classification report, and confusion matrix.
  + Artifacts like the model and test predictions are saved using *joblib.dump()* for reuse.

## GUI Application:

* + The *gui()* function in *gui.py* uses *tkinter* and *ttkbootstrap* to create a form-based application.
  + Features are displayed as dropdowns or entry fields, and inputs are converted into model-expected numerical format using dictionaries.
  + Upon clicking “Predict Attrition,” the model's prediction is shown in a message box.
  + The “Generate Report” button calls *generate\_report()* from *reports.py*.

A screenshot of a computer

AI-generated content may be incorrect.

Figure - GUI main screen

A screenshot of a computer

AI-generated content may be incorrect.

Figure - Prediction Result(1)

A screenshot of a computer

AI-generated content may be incorrect.

Figure - Prediction Result(2)

## Report Generation:

* + The *generate\_report()* function in *reports.py* compiles visual plots and performance metrics into a PDF using PdfPages from *matplotlib*.
  + It includes summary statistics, histograms, heatmaps, confusion matrix, classification report, and final accuracy.

A screenshot of a computer

AI-generated content may be incorrect.

Figure - Report generation

# Results

The implemented model achieved strong predictive performance, demonstrating its effectiveness in addressing the problem of employee attrition.

## Accuracy:

The Random Forest Classifier showed high accuracy, indicating its reliability for prediction and its ability to learn complex patterns from the training data without overfitting. This level of accuracy is very important in real-world HR applications where incorrect predictions can lead to poor decision-making.

A white background with black text

AI-generated content may be incorrect.

Figure - Report showing model accuracy

## Visualization Insights:

Employees aged under 30 showed higher attrition rates, which may indicate career exploration phases or job dissatisfaction early in their careers. Similarly, employees who lived far from their workplace or earned lower income levels were more likely to leave. Overtime workload and certain job roles also showed significant correlation with higher attrition, offering insights for possible policy changes or employee support strategies.

## GUI Functionality:

The GUI was designed to accommodate both categorical and numerical inputs, mapped to what the model expects. Predictions are made in real-time with a single button click. The easy to use and neat layout ensures that HR staff can use the system without requiring any technical knowledge or training.

## Report Quality:

The generated report includes all major findings of the model and the data. It provides visual summaries, statistical representations, and evaluation metrics, including classification reports and confusion matrices. These visual tools are highly beneficial for presentations, audits, and future data reviews, enabling HR teams to understand not just what the model predicts but why.

A blue squares with white text

AI-generated content may be incorrect.

Figure - An example of a visual tool useful for reviews

# Conclusion

This AI-powered program provides a complete solution for analyzing and predicting employee attrition. By combining machine learning, feature engineering, and a user-friendly GUI, it offers actionable insights to HR professionals. The project highlights the power of data-driven decision-making in human resource management. The program does not limit itself to an attrition predictor with good accuracy, but it also provides deeper understanding of the factors affecting it through visual tools and automated reports.

In the future, this project can be expanded by:

* Incorporating deep learning models.
* Connecting to live HR databases.
* Adding feedback loops for continuous model improvement.
* Extending the GUI with employee recommendation systems or retention strategies.

Overall, the system stands as a reliable, efficient, and easy-to-use tool to reduce attrition and maintain a stable workforce.