
Northeastern University

EAI-6400
Data Governance & Responsible AI

BY INSTRUCTOR
Dimodugno, Mimoza

PROJECT REPORT
Module-1

SUBMITTED BY
Iti Rohilla

On
Mar-07-2025

# HR Analytics: Attrition Prediction Model Report

## Introduction

The modern business landscape is profoundly affected by employee turnover, which can significantly disrupt organizational stability and incur high costs related to hiring and training replacements. Furthermore, it can erode internal knowledge and adversely affect team morale and productivity. In light of these challenges, there is a pressing need for organizations to harness the power of data to preemptively address employee dissatisfaction and enhance retention strategies.

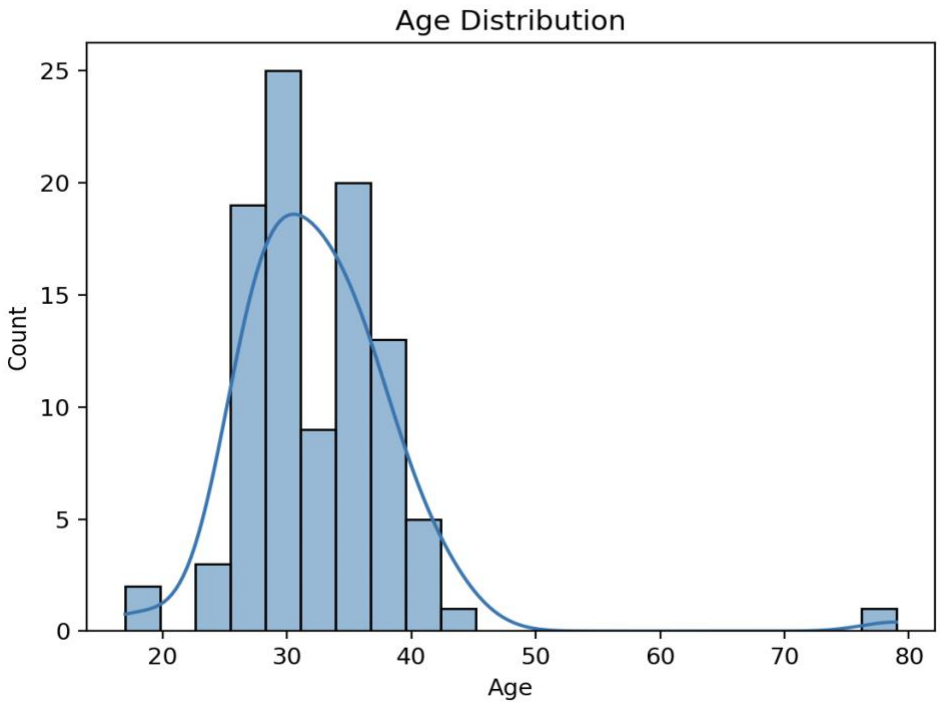
This report outlines a sophisticated approach using machine learning techniques to predict employee attrition, providing valuable insights that empower HR teams to implement targeted interventions. By leveraging historical data encompassing various employee attributes and their employment outcomes, we aim to identify patterns and predictors of attrition.

This predictive model not only serves to reduce turnover rates but also supports strategic decision-making processes that foster a more engaged and stable workforce. Through this initiative, we are not only addressing a cost center for the organization but are also enhancing its ability to plan for the future, ensuring that the workforce is motivated, satisfied, and aligned with the organizational goals.

## 1. Data Exploration and Visualization

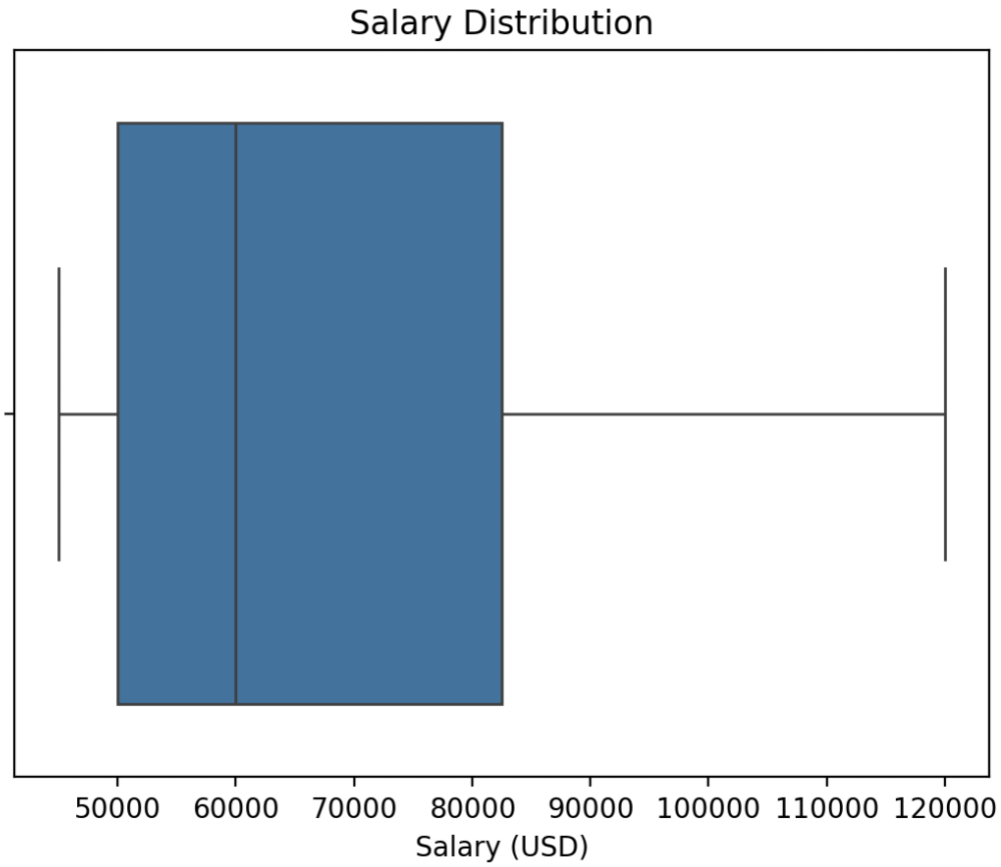
In the initial phase of our analysis, comprehensive data exploration and visualization techniques were employed to understand the underlying structures and patterns within the employee dataset. By dissecting various attributes through graphical representations, we gained insights into demographic distributions, compensation trends, and potential predictors of attrition. These insights are crucial for framing our predictive modeling strategy and identifying key areas where HR interventions might be most needed. Below are detailed descriptions and analyses of the visualizations conducted on different aspects of the dataset, each providing unique insights into the workforce dynamics.

### a) Age Distribution (Histogram with KDE)



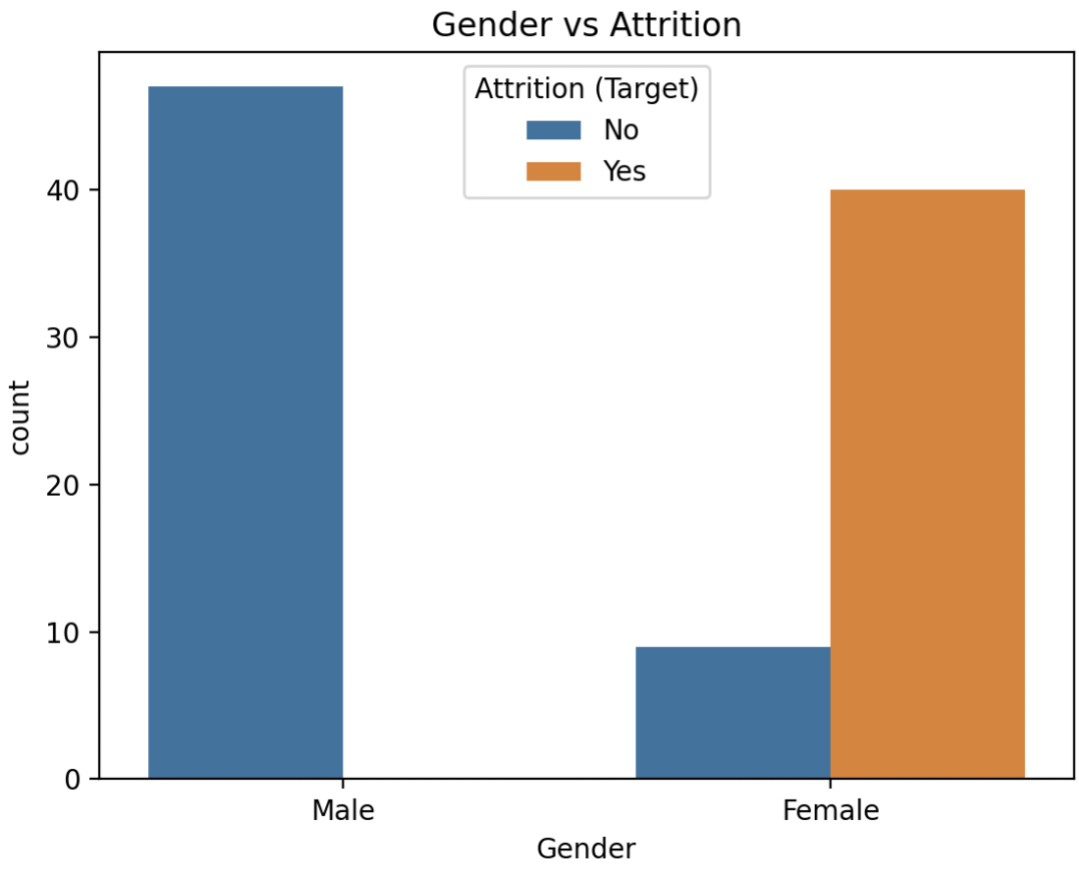
- **Observed Pattern**: The age distribution shown through the histogram with a Kernel Density Estimate (KDE) overlay likely indicates the central tendency of age, such as a mean or median age, and how broadly the age of employees is spread around this central value.
- **Analysis**: If the histogram peaks around a middle age range and tails off towards younger and older ages, it suggests a workforce with a strong presence of mid-career professionals. This central peak indicates stability but could also hint at potential future challenges in succession planning as these employees near retirement.

**b) Salary Distribution (Box Plot)**



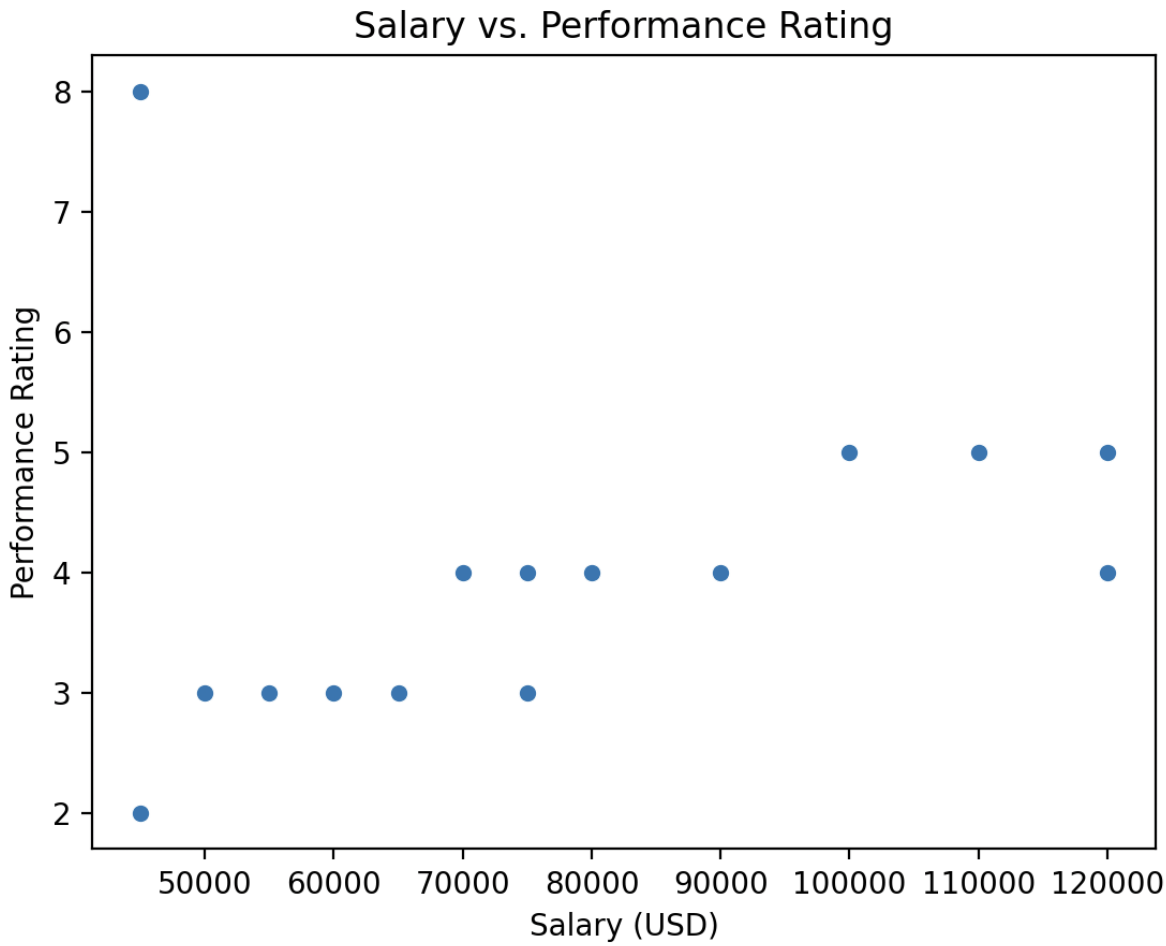
- **Observed Pattern**: The salary distribution, if depicted by a box plot, typically shows the median salary, the interquartile range (25th to 75th percentile), and any outliers. Outliers can appear as individual points beyond the whiskers of the box plot.
- **Analysis**: A concentration of data within the interquartile range coupled with outliers might suggest a generally consistent salary structure with exceptions for roles that are either highly specialized or senior. Salary disparities, especially if the outliers are only on the higher end, might indicate issues like pay inequity which could affect employee morale and potentially lead to attrition.

**c) Gender vs Attrition (Count Plot)**



- **Observed Pattern**: This count plot would typically show the number of employees who have left the organization broken down by gender. A higher count for one gender over the other could be significant.
- **Analysis**: If one gender is disproportionately represented in attrition rates, this could indicate a deeper systemic issue within the organization's culture or policies that affects one gender more adversely. For instance, higher female attrition could suggest challenges in balancing work and family, or a male-dominated workplace might not be as inclusive.

**d) Salary vs Performance Rating (Scatter Plot)**



- **Observed Pattern**: A scatter plot relating salary and performance ratings would ideally show how these two variables correlate, revealing whether employees with higher performance ratings also receive higher salaries.
- **Analysis**: A positive correlation where higher performance aligns with higher salary is expected and healthy. However, significant scatter or a weak correlation suggests inconsistencies in how performance is rewarded, which could demotivate high performers from staying in the organization if they feel their efforts are not adequately recognized financially.

**2. Data Wrangling**

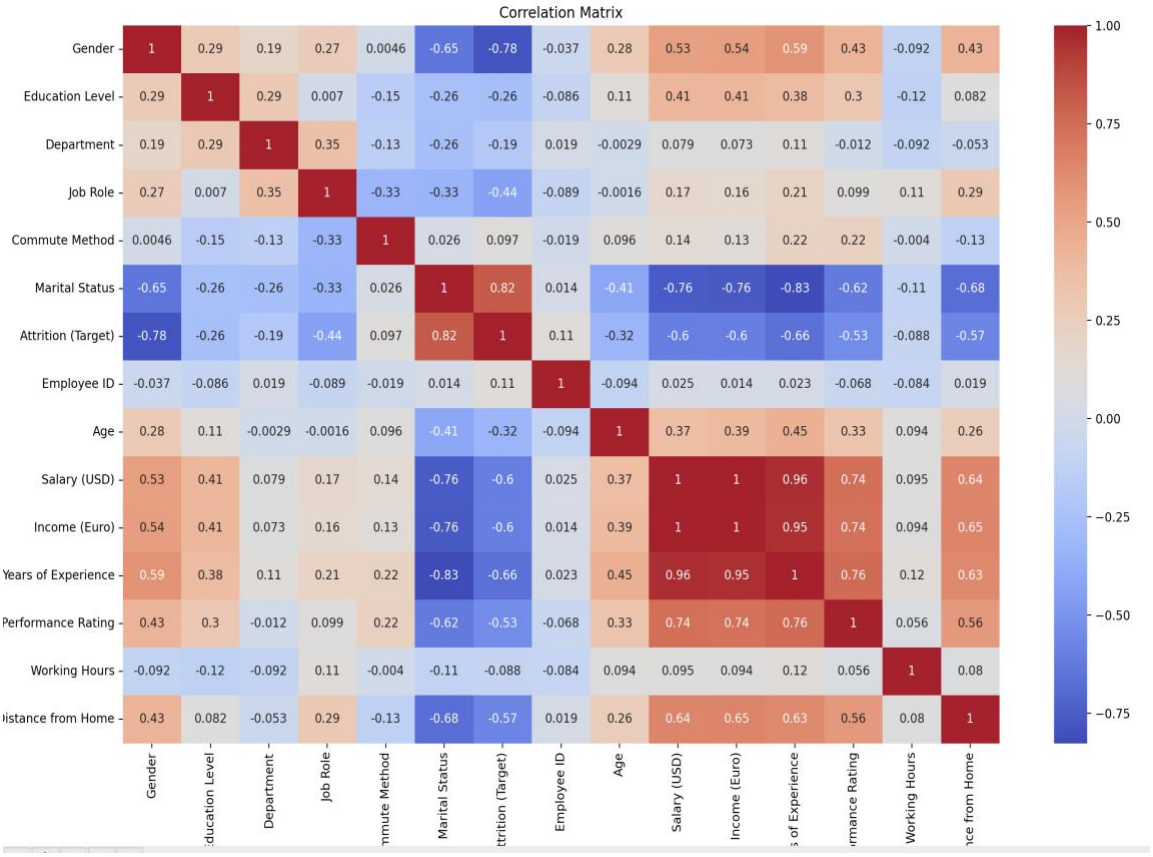
To prepare the dataset for machine learning modeling, several data wrangling steps were necessary:

- **Handling Missing Values**: Employed imputation techniques where necessary. Numerical missing values were addressed with KNN imputation to maintain data integrity, while categorical missing values were handled by assigning the most frequent category or a placeholder category 'Unknown'.
- **Encoding Categorical Variables**: Transforming categorical variables into a format suitable for modeling using Label Encoding or One-Hot Encoding, depending on the model requirements.
- **Normalization/Standardization**: Ensuring that numerical data conforms to a standard scale, particularly important for models sensitive to data scaling such as SVM or KNN.

## 3. Feature Engineering and Data Splitting

### 1)Feature Selection:

a) **Variables Related to Attrition:** Initial analysis identified 'Marital Status' and 'Years of Experience' as strongly negatively correlated with attrition, suggesting their significant predictive value. Both variables are indicative of employee stability and potential satisfaction, making them crucial for attrition modeling.




b) **Variables Dropped:** 'Income (Euro)' was removed due to its high redundancy with 'Salary (USD)', as both variables provided similar information which could lead to multicollinearity, complicating the model unnecessarily.

c) **Final Variables Considered:** After refining the feature set, the model included 'Salary (USD)', 'Marital Status', 'Years of Experience', 'Age', and 'Performance Rating'. These were chosen for their unique contributions to understanding attrition, thereby optimizing the model's effectiveness and efficiency in predicting employee turnover.

**2)Data Splitting**: The data was split into an 80-20 training-test ratio. This split ensures that the model can be trained on a substantial portion of the data while also being validated on a separate set to check for overfitting.

**4. Model Development and Evaluation**

Several models were developed and evaluated:

For our project focused on predicting employee attrition, we utilized a specific dataset featuring a variety of employee attributes. Based on the nature of the data and the project goals, we developed and evaluated two primary machine learning models: the Random Forest Classifier and the Support Vector Machine (SVM). Additionally, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to enhance model performance by addressing class imbalance.

### a. Random Forest Classifier

The Random Forest Classifier was selected for its efficacy in handling both categorical and continuous variables and its robustness against overfitting. This model works by constructing multiple decision trees during training and outputting the class that is the mode of the classes (attrition or no attrition) predicted by individual trees. For our specific dataset, which included

features like age, gender, salary, and performance ratings, Random Forest was advantageous because it could handle the high dimensionality and provide insights into which features most strongly predict attrition. We tuned hyperparameters such as `n_estimators` and `max_depth` to optimize the model, aiming to balance the trade-off between bias and variance.

### b. Support Vector Machine (SVM)

We also used SVM because of its known capacity to efficiently classify linear and non-linear data. SVM was particularly considered for its ability to model complex relationships through the use of kernel functions, making it suitable for our varied dataset. The SVM model was configured with a radial basis function (RBF) kernel, and parameters like `C` (regularization parameter) and `gamma` (kernel coefficient) were carefully adjusted through grid search with cross-validation. This ensured the model was neither underfitting nor overfitting and was crucial for achieving a high degree of accuracy in distinguishing between employees who would leave or stay.

### c. SMOTE (Synthetic Minority Over-sampling Technique)

Given the challenge of class imbalance in our dataset—where the number of employees who stayed was much higher than those who left—SMOTE was applied to preprocess the data effectively. SMOTE works by creating synthetic examples rather than over-sampling with replacement, which helped in presenting a more balanced dataset to train our models. This was particularly beneficial in enhancing the performance of both the SVM and Random Forest models, improving their ability to detect the minority class of employees likely to attrite, thus increasing the recall without sacrificing precision.

## Model Evaluation

The performance of both models was rigorously evaluated using metrics such as accuracy, precision, recall, and the F1 score. The Random Forest model initially showed promising results but required adjustments to deal with overfitting observed during the validation phase. After tuning, it demonstrated a strong ability to generalize across new data. The SVM, enhanced with SMOTE, showed improvements particularly in handling the minority class, proving its worth in scenarios typical of real-world HR datasets.

By utilizing these models in conjunction, our analysis not only highlighted the most predictive features of attrition but also offered a robust framework for HR managers to predict and, consequently, mitigate employee turnover. This targeted approach allows for more precise HR interventions, ultimately leading to improved retention strategies and a more stable workforce environment.

**Model Performance:**

The performance metrics after applying SMOTE and other adjustments showed an accuracy of 85%, with improved balance in precision and recall across classes. The detailed classification report provided metrics like F1-score, which combines both precision and recall, useful for evaluating the overall quality of the model.

## 5. Comparison of Models

A comparative analysis was performed between the Random Forest and SVM models:

| Model | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Random Forest (No SMOTE) | 1.00 | 0.77 | 0.87 | 85% |
| Random Forest (With SMOTE) | 1.00 | 0.77 | 0.87 | 85% |
| Random Forest (Class Weights) | 1.00 | 0.77 | 0.87 | 85% |
| SVM | 1.00 | 0.69 | 0.82 | 80% |

**6. Recommendations for Improvement**

- **Feature Exploration**: Further analysis into additional features like employee engagement scores or managerial feedback could provide deeper insights.
- **Advanced Modeling Techniques**: Exploring advanced ensemble methods or deep learning could potentially improve predictive performance.
- **Cross-Validation and Hyperparameter Tuning**: More extensive cross-validation and hyperparameter optimization could enhance model robustness and accuracy.

## CONCLUSION

The journey through data exploration, model selection, and analytical rigor has culminated in the development of a robust model capable of predicting employee attrition with a high degree of accuracy. This model serves as a critical tool in the HR arsenal, offering insights that pave the way for strategic, data-driven decision-making to enhance employee retention and satisfaction.

The implementation of this model across organizational structures allows for a proactive approach to managing attrition, shifting from reactive hiring strategies to a more strategic employee engagement and retention plan. However, the landscape of workforce management is perpetually evolving, necessitating continuous refinement of our models and strategies. As we move forward, it will be essential to integrate newer data sources, embrace advanced analytical techniques, and foster an adaptive learning culture within the HR function.

By doing so, we not only keep pace with changes in employee dynamics and market conditions but also reinforce our commitment to building a resilient, motivated, and cohesive workforce, thus driving the organization towards sustained success and growth. Through strategic foresight and a commitment to data-driven strategies, we can significantly mitigate the challenges of attrition and build an organizational culture that attracts, nurtures, and retains top talent.