Journal of Information Organization

DLINE JOURNALS

Print ISSN : 2278 – 6503
Online ISSN: 2278 – 6511

JIO 2026: 16 (4)
https://doi.org/10.6025/jio/2026/16/4/237-256

Data Quality Assessment and Exploratory Data Analysis of an HR Analytics Dataset

P Paramasivaiah
Department of Studies and Research in Commerce
Tumkur University
TUMKUR- 572 103
India
paramashivaiah@gmail.com

ABSTRACT

In the era of big data, people analytics has become a vital tool for organizations seeking to mitigate the substantial costs and operational risks associated with employee attrition. This study presents a comprehensive data quality assessment and exploratory data analysis (EDA) of an HR analytics dataset comprising 1,473 employee records across 38 variables. The analysis reveals a moderately imbalanced attrition rate of 16.09%, with significant disparities across demographic and occupational lines. Key findings indicate that employee attrition is disproportionately high among Sales Representatives (39.3%), employees working overtime (30.5%), and younger demographics aged 18–25 (35.8%). Furthermore, strong positive correlations were identified between career progression and compensation, notably between JobLevel and MonthlyIncome (r = 0.95). The study also highlights critical data quality challenges, including exact duplicates, conflicting employee identifiers, and isolated missing values in managerial tenure, underscoring the necessity of robust automated preprocessing. By bridging empirical data insights with contemporary literature on machine learning and ethical AI governance, this research emphasizes that effective attrition prediction requires not only advanced algorithms like Logistic Regression and Random Forest but also a commitment to "privacy by design" and organizational justice to ensure ethical, transparent, and actionable HR decision making.

Keywords: Employee Attrition, People Analytics, Exploratory Data Analysis, Predictive Modeling, Data Quality, Ethical AI, Human Resources

Received: 1 April 2026, Revised: 2 June 2026, Accepted: 2 July 2026

Copyright: **DLINE**

1. Introduction

In the era of data science and big data analytics, people analytics helps organizations and their human resources (HR) managers reduce attrition by transforming how they attract and retain talent. Employee attrition presents a critical problem and a significant risk, as it negatively impacts not only organizational productivity but also planning continuity [1]. In today's knowledge-driven industry, employees are among the most critical assets.

Consequently, attrition can severely damage an organization's competitive advantage and incur substantial costs related to the human resources life cycle, lost knowledge, diminished employee morale, and degraded organizational culture [2].

2. Background

2.1 Predictive Analytics and Machine Learning in HR

To address these challenges, researchers have extensively explored predictive models to understand and mitigate attrition. For instance, research measuring people management skills through employee questionnaires regarding their managers revealed a strong negative relationship between effective people management and employee attrition [3].

Various machine learning techniques have been applied to predict resignation decisions. Frye, Boomhower, and Smith utilized Principal Component Analysis, K-Nearest Neighbours, and Random Forest to model statistically significant factors related to resignation. They found that Logistic Regression predicted employee attrition with the highest accuracy among their testing methods, reaching a success rate higher than seventy-four per cent [4]. Similarly, Bindra, Sehgal, and Jain [5] analyzed attrition data across five classes using the Apriori Association Rule Algorithm and the C5.0 Decision Tree Algorithm. Their conclusion was that the algorithm performs better when C5.0 is combined with the association rule algorithm than when C5.0 is used independently. In the IT sector, Raja and Kumar evaluated the circumstances influencing attrition using a sample of 300 workers. Utilizing simple percentage, chi-square, and correlation coefficient methods, they found no correlation between the factors that retain employees and the factors that influence the employee work environment [6]. Beyond attrition, Mishra and Lama proposed a framework comprising data mining and predictive analytics to improve the effectiveness of HR systems and optimize business outcomes [7].

Regression methods remain an integral part of data analysis for explaining the correlation between a response variable and analytical variables, with the logistic regression model being the most frequently used [8]. This model has been widely applied to identify factors determining job satisfaction [9], work-life balance [10] work environment and production loss [11], stress and dissatisfaction [12], and the effects of incentives on employee engagement [13].

2.2 Data Collection and Preprocessing Challenges

The success of these machine learning and predictive models heavily relies on data quality. Carvalho highlights the challenges of raw data collection and preprocessing in real-world environments. Contextual data gathered over several weeks across various offices comes from diverse sources, such as sensors, questionnaires, and apps, and is highly susceptible to interference and noisy conditions. Because data quality directly influences machine learning outcomes, and given the sheer number of instruments used, implementing an automated preprocessing stage is essential to filter noise and improve data quality [14].

2.3 The Integration of AI in Human Resources

Building on predictive analytics, the operational and strategic applications of Artificial Intelligence (AI) in HR have seen intensifying interest over the past five years [15]. AI is now deployed at an impressive pace across multiple stages of the employee lifecycle, including recruitment, performance appraisal, and learning and development.

In recruitment, studies highlight AI's value in aiding candidate evaluation through automated sourcing, screening, and assessments. These systems can parse thousands of resumes, flag ideal candidates, and even assess behavioral cues via video analysis, thereby reducing hiring time and human error [16]. Furthermore, tools such as chatbots, Applicant Tracking Systems (ATS), and algorithmic filters significantly enhance recruiter productivity and alleviate administrative burdens [17].

2.4 Ethical Considerations and AI Governance in HR

(e.g., in MonthlyIncome and TotalWorkingYears) without discarding legitimate extreme values.

• *Phase 2: Exploratory Data Analysis (EDA) & Feature Engineering*
o Conduct univariate, bivariate, and multivariate analyses to identify key drivers of attrition (e.g., overtime status, age demographics, department, and income disparities).

o Engineer new features (e.g., tenure-to-promotion ratio) and encode categorical variables for machine learning compatibility.

• *Phase 3: Predictive Modeling*
o Train and evaluate multiple machine learning algorithms (Logistic Regression, Random Forest, K-Nearest Neighbours, and Decision Trees/C5.0) to predict attrition.

o Implement techniques to handle class imbalance, such as SMOTE (Synthetic Minority Over-sampling Technique) or class-weight adjustments.

• *Phase 4: Ethical AI Governance & Evaluation*
o Apply "privacy-by-design" principles and conduct algorithmic audits to ensure the model does not disproportionately penalize specific demographic groups (e.g., younger employees or specific departments).

o Utilize Explainable AI (XAI) techniques (e.g., SHAP values) to ensure model transparency and foster organizational justice.

3.1 Experimental Setup

• Dataset: The analytical working dataset comprises 1,473 observations and 38 variables (26 numerical, 12 categorical). The target variable is Attrition (Imbalanced: 83.91% Retained, 16.09% Left).

• Environment: Python programming language utilizing libraries such as pandas and numpy for data manipulation, scikit-learn for machine learning modeling, and matplotlib/seaborn for EDA visualization.

• Algorithms to be Tested:
o *Logistic Regression:* For baseline interpretability and identifying linear correlations.

o *Random Forest & Decision Trees (C5.0):* To capture non-linear relationships and feature interactions.

o *K-Nearest Neighbours (KNN) & Principal Component Analysis (PCA):* For dimensionality reduction and pattern recognition.

• Evaluation Metrics: Due to the 16.09% class imbalance, Accuracy will not be the primary metric. The models will be evaluated using Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) to ensure the model effectively identifies the minority attrition class without generating excessive false positives.

• Validation Strategy: Stratified K-Fold Cross-Validation to maintain the class distribution across training and testing folds.

3.2 Proposed System Design / Architecture

To address both the predictive and ethical requirements outlined in the literature, the system is designed as a Four Tier Ethical HR Analytics Architecture.

Despite its benefits, the rapid integration of AI in HR raises significant ethical and governance concerns. Gélinas et al. [18] warn that continuous monitoring, algorithmic decision making, and opaque models can severely diminish employee autonomy.

To counteract this, several scholars advocate for "privacy by design" as a default principle in AI-HRM tools [19]. Proposed governance strategies include conducting algorithmic audits, ensuring employee representation on AI governance boards, and utilizing differential privacy mechanisms to mitigate identification risks. Furthermore, Hernández [20] introduces a multidimensional framework emphasizing transparency, inclusivity, and stakeholder participation, asserting that AI governance must transcend mere compliance to foster organizational justice. Finally, Aizenberg and van den Hoven [21] shift the discourse from empirical evaluation to a normative framework firmly grounded in human rights.

2.5 Research Statement

This research aims to develop a robust, ethically governed predictive analytics framework for employee attrition by leveraging machine learning and people analytics. By transforming raw, noisy HR data into actionable insights, this study seeks to empower organizations to proactively identify attrition risks, optimize talent retention strategies, and reduce the substantial financial and operational costs associated with employee turnover. Furthermore, this research integrates a "privacy by design" and ethical AI governance model to ensure that predictive HR tools enhance organizational decision making without compromising employee autonomy, transparency, or data privacy.

2.6 Research Problem

While employee attrition severely impacts organizational productivity, planning continuity, and competitive advantage, current HR practices often rely on reactive rather than proactive measures. Although machine learning offers powerful predictive capabilities, several critical challenges hinder its effective and ethical deployment in HR:

1. Data Quality and Noise: Real-world HR datasets are highly susceptible to inconsistencies. As identified in the preliminary analysis, datasets contain exact duplicates, conflicting employee identifiers, missing values (e.g., 57 missing records in YearsWithCurrManager), and statistical outliers that can skew model performance.

2. Class Imbalance: Attrition datasets are inherently imbalanced (e.g., 83.91% retained vs. 16.09% attrition in the analyzed dataset). Standard accuracy metrics fail to capture model efficacy in these scenarios, risking the marginalization of the minority (attrition) class.

3. Ethical and Governance Risks: The deployment of AI in HR raises severe ethical concerns. Continuous monitoring, opaque algorithmic decision-making, and biased models can diminish employee autonomy and foster organizational injustice. There is a lack of integrated frameworks that combine high accuracy predictive modeling with strict ethical governance (e.g., privacy by design, algorithmic audits).

3. Methodology

The research will follow a mixed-methods, multi-phase methodology bridging data science and ethical AI governance:

- Phase 1: Data Preprocessing & Quality Assurance

o *Deduplication & Conflict Resolution:* Remove exact duplicate rows and resolve conflicting EmpID records through source verification.

o *Missing Value Imputation:* Apply advanced imputation techniques (e.g., KNN imputation or median substitution) for the 57 missing values in YearsWithCurrManager.

o *Outlier Treatment:* Utilize the Interquartile Range (IQR) method to identify and handle statistical outliers

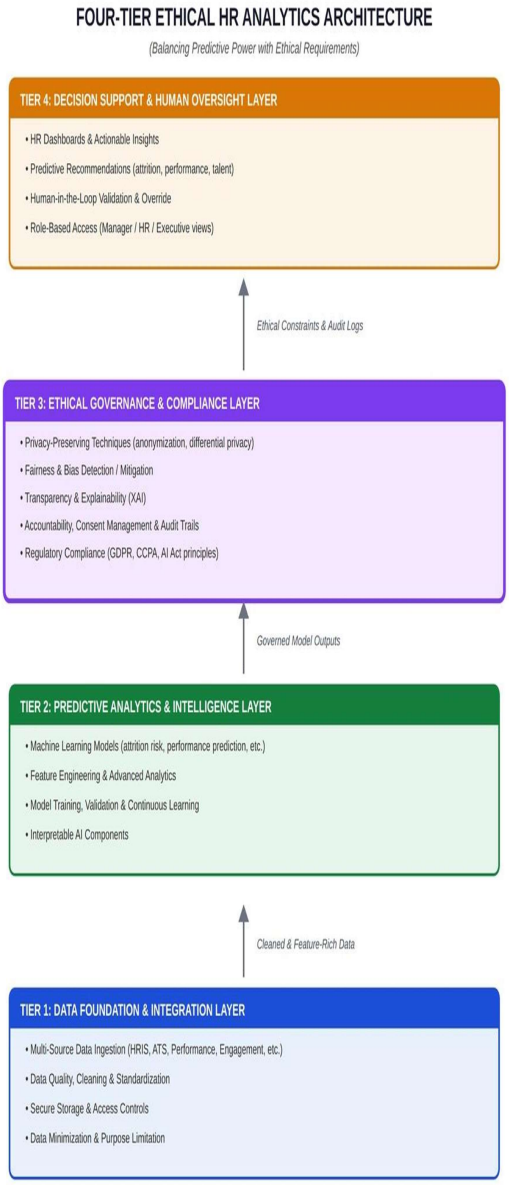

Figure 1. HR Analytics Architecture

Architectural Components Explained:

Tier 1 (Data Layer): Addresses the raw data challenges identified in the study (e.g., handling the 7 exact duplicates, resolving conflicting EmpID records, and managing the 57 missing values in Years With CurrManager).

Tier 2 (Modeling Layer): Acts as the computational core where algorithms like Logistic Regression and Random Forest are trained to predict the 16.09% attrition class, utilizing the strong correlations identified in the EDA (e.g., JobLevel and MonthlyIncome).

Tier 3 (Governance Layer): Directly responds to the ethical concerns raised by scholars like Gélinas, Batool, and Hernández. It ensures the AI does not operate as a "black box," enforces privacy by design, and guarantees that predictive insights do not violate employee autonomy or human rights.

Tier 4 (Application Layer): Translates complex ML outputs into actionable, transparent insights for HR managers, highlighting specific risk areas (e.g., the 39.3% attrition rate in Sales Representatives or the 30.5% risk associated with overtime) while maintaining organizational justice.

3.3 Dataset overview

The source file contains 1,480 employee records and 38 variables. The uploaded table comprises 26 numerical and 12 categorical variables based on their stored data types. Seven exact duplicate rows were identified. Removing only these exact duplicates produced an analytical working table of 1,473 rows. Three employee identifiers remain duplicated with conflicting YearsWithCurrManager values and are retained pending source verification.

3.4 Data quality assessment

Original observations	1480	Rows in uploaded CSV
Variables	38	Columns in uploaded CSV
Numerical variables	26	Based on pandas data types
Categorical variables	12	Based on pandas data types
Exact duplicate rows	7	Exact full-row duplicates
Unique EmpID values	1470	Distinct employee identifiers
EmpID values repeated	10	Distinct IDs appearing more than once
Missing cells	57	All missing cells
Rows after exact duplicate removal	1473	Retains three conflicting repeated-ID pairs for review

Table 1. Overall dataset structure and data-quality assessment

Table 1 summarizes the structural characteristics and initial data-quality status of the HR Analytics dataset. The original dataset comprises 1,480 employee observations and 38 variables, including 26 numerical and 12 categorical variables. This variable structure provides a broad representation of employee demographics, organizational characteristics, compensation, job satisfaction, career progression and employment outcomes.

The initial inspection identified seven exact duplicate records, which were removed to establish an analytical working dataset containing 1,473 observations. However, the employee identifier assessment revealed that 10 EmpID values occurred more than once. Seven repeated records corresponded to exact duplicates, whereas three employee identifiers had records with conflicting values for YearsWithCurrManager. These conflicting records were retained pending source verification rather than being arbitrarily deleted.

The distinction between exact duplicates and repeated identifiers is important because duplicate observations can artificially increase the apparent frequency of particular employee characteristics and influence statistical estimates. The retained conflicting records require further investigation to establish whether they represent data-entry inconsistencies or legitimate differences.

There are 57 missing cells, all in YearsWithCurrManager (57 observations; 3.85%). No missing values were detected in the other variables. Ten EmpID values occur more than once. Seven of these represent exact duplicate rows; the other three repeated identifiers have inconsistent manager tenure values.

BusinessTravel contains both 'Travel_Rarely' (1,038 records after exact row deduplication) and 'TravelRarely' (7 records). These labels appear semantically similar but should be confirmed against the data dictionary before harmonization. EmployeeCount, Over18 and StandardHours are constant in this dataset and have no discriminatory value for modelling. EmpID and EmployeeNumber are identifiers and should not be used as predictors.

The dataset demonstrates an adequate initial structure for exploratory HR analytics, although duplicate identifiers and missing manager-tenure information require explicit treatment before inferential modelling and machine learning applications.

4. Analysis

4.1Descriptive statistics

The mean employee age is 36.92 years (SD = 9.13), mean monthly income is 6,500.23 (SD = 4,706.05), and mean total work experience is 11.28 years (SD = 7.78). Mean tenure at the current company is 7.00 years (SD = 6.12). These statistics are calculated after removal of the seven exact duplicate rows; the three conflicting repeated-ID records remain pending verification.

Variable	count	mean	std	min	25%	50%	75%	max
Age	1473	36.917	9.131	18	30	36	43	60
DailyRate	1473	802.66	403.245	102	465	802	1157	1499
DistanceFromHome	1473	9.197	8.108	1	2	7	14	29
Education	1473	2.911	1.025	1	2	3	4	5
EnvironmentSatisfaction	1473	2.724	1.093	1	2	3	4	4
HourlyRate	1473	65.834	20.35	30	48	66	83	100
JobInvolvement	1473	2.73	0.712	1	2	3	3	4
JobLevel	1473	2.064	1.106	1	1	2	3	5
JobSatisfaction	1473	2.728	1.103	1	2	3	4	4
MonthlyIncome	1473	6500.229	4706.054	1009	2911	4908	8380	19999
MonthlyRate	1473	14321.376	7115.54	2094	8053	14242	20467	26999
NumCompaniesWorked	1473	2.693	2.497	0	1	2	4	9
PercentSalaryHike	1473	15.212	3.657	11	12	14	18	25
PerformanceRating	1473	3.153	0.361	3	3	3	3	4

RelationshipSatisfaction	1473	2.712	1.082	1	2	3	4	4
StockOptionLevel	1473	0.794	0.851	0	0	1	1	3
TotalWorkingYears	1473	11.278	7.776	0	6	10	15	40
TrainingTimesLastYear	1473	2.8	1.289	0	2	3	3	6
WorkLifeBalance	1473	2.762	0.706	1	2	3	3	4
YearsAtCompany	1473	7.005	6.121	0	3	5	9	40
YearsInCurrentRole	1473	4.228	3.621	0	2	3	7	18
YearsSinceLastPromotion	1473	2.184	3.22	0	0	1	3	15
YearsWithCurrManager	1416	4.117	3.559	0	2	3	7	17

Table 2. Statistical Analysis

Table 2 presents the central tendency, dispersion and distributional characteristics of the numerical variables using the mean, standard deviation, minimum, maximum and quartiles.

The average employee age was 36.92 years (SD = 9.13), with a median of 36 years and an observed range of 18–60 years. This indicates that the workforce predominantly comprises employees in the early and middle stages of their careers.

Monthly income exhibited substantial variability, with a mean of 6,500.23 (SD = 4,706.05) and a median of 4,908. The difference between the mean and median suggests a positively skewed income distribution, potentially reflecting substantial compensation differences across organizational levels and occupational roles.

The mean total working experience was 11.28 years (SD = 7.78), whereas average organizational tenure was 7.00 years (SD = 6.12). The median organizational tenure was five years, indicating that the distribution contains employees with substantially different lengths of service.

Career progression variables also exhibited considerable variation. The mean number of years in the current role was 4.23, while the mean number of years since the last promotion was 2.18. These indicators provide a descriptive foundation for examining career mobility and its potential association with employee retention. The average distance from home was 9.20 units (SD = 8.11), while the mean training frequency was 2.80 times during the preceding year.

The descriptive statistics demonstrate substantial heterogeneity in employee demographics, compensation and career progression. In particular, variation in monthly income, working experience and organizational tenure suggests that employee attrition should be investigated using multivariable approaches rather than relying exclusively on individual demographic characteristics.

4.2 Numeric Outliers

Variable	Q1	Q3	IQR	Lower_ Bound	Upper_ Bound	Potential_ Outliers	Percent_of_ Nonmissing
Age	30	43	13	10.5	62.5	0	0
DailyRate	465	1157	692	-573	2195	0	0
DistanceFromHome	2	14	12	-16	32	0	0
Education	2	4	2	-1	7	0	0
EnvironmentSatisfaction	2	4	2	-1	7	0	0
HourlyRate	48	83	35	-4.5	135.5	0	0
JobInvolvement	2	3	1	0.5	4.5	0	0
JobLevel	1	3	2	-2	6	0	0
JobSatisfaction	2	4	2	-1	7	0	0
MonthlyIncome	2911	8380	5469	-5292.5	16583.5	114	7.74
MonthlyRate	8053	20467	12414	-10568	39088	0	0
NumCompaniesWorked	1	4	3	-3.5	8.5	52	3.53
PercentSalaryHike	12	18	6	3	27	0	0
PerformanceRating	3	3	0	3	3	226	15.34
RelationshipSatisfaction	2	4	2	-1	7	0	0
StockOptionLevel	0	1	1	-1.5	2.5	85	5.77
TotalWorkingYears	6	15	9	-7.5	28.5	63	4.28
TrainingTimesLastYear	2	3	1	0.5	4.5	239	16.23
WorkLifeBalance	2	3	1	0.5	4.5	0	0
YearsAtCompany	3	9	6	-6	18	104	7.06
YearsInCurrentRole	2	7	5	-5.5	14.5	21	1.43
YearsSinceLastPromotion	0	3	3	-4.5	7.5	107	7.26
YearsWithCurrManager	2	7	5	-5.5	14.5	13	0.92

Table 3. IQR-based outlier screening

Table 3 presents the identification of potential numerical outliers using the interquartile range (IQR) procedure. The method evaluates observations against lower and upper screening boundaries calculated using the first and third quartiles:

$$IQR = Q_3 _ Q_1$$

$$\text{Lower} = Q_1 _ 1.5\ (IQR)$$

$$\text{Upper} = Q_3 + 1.5\ (IQR)$$

Observations outside these boundaries are classified as potential statistical outliers.

The analysis is particularly relevant to variables such as monthly income, total working years, organizational tenure and distance from home, where substantial variation may occur naturally across employees.

However, potential outliers should not automatically be interpreted as erroneous observations. For example, high monthly income may reflect seniority or specialized occupational responsibilities, while long organizational tenure may represent legitimate employment history.

Scientific interpretation: The IQR procedure provides an initial screening mechanism for identifying unusual observations and assessing the distributional characteristics of numerical variables. Any subsequent exclusion or transformation should be supported by data verification and substantive justification. The screening results should therefore be interpreted as diagnostic evidence rather than confirmation of data errors.

4.4 Attrition distribution

In the analytical working dataset, 1,236 employees (83.91%) are classified as retained and 237 (16.09%) as having left. This outcome imbalance should be considered in predictive modelling; accuracy alone would not adequately characterize model performance.

4.5 Exploratory group comparisons

Observed attrition rates are highest among Sales Representatives (39.3%) and lowest among Research Directors (2.5%). Attrition is 30.5% among employees reporting overtime compared with 10.4% among those not reporting overtime. The age-group attrition rate is highest for employees aged 18–25 (35.8%) and lowest for those aged 36–45 (9.1%). These are unadjusted descriptive associations and should not be interpreted as causal effects.

Department level attrition rates are 20.6% in Sales, 19.0% in Human Resources and 13.8% in Research & Development. Single employees show an observed attrition rate of 25.5%, compared with 12.4% among married employees and 10.1% among divorced employees. These comparisons should be followed by inferential analyses that account for sample size, confounding and multiple comparisons.

Variable_1	Variable_2	Pearson_r	Abs_r
JobLevel	MonthlyIncome	0.95	0.95
JobLevel	TotalWorkingYears	0.782	0.782
PercentSalaryHike	PerformanceRating	0.773	0.773
MonthlyIncome	TotalWorkingYears	0.772	0.772
YearsAtCompany	YearsWithCurrManager	0.765	0.765
YearsAtCompany	YearsInCurrentRole	0.759	0.759
YearsInCurrentRole	YearsWithCurrManager	0.715	0.715

Age	TotalWorkingYears	0.68	0.68
TotalWorkingYears	YearsAtCompany	0.628	0.628
YearsAtCompany	YearsSinceLastPromotion	0.618	0.618
YearsInCurrentRole	YearsSinceLastPromotion	0.548	0.548
JobLevel	YearsAtCompany	0.535	0.535
MonthlyIncome	YearsAtCompany	0.514	0.514
YearsSinceLastPromotion	YearsWithCurrManager	0.513	0.513
Age	JobLevel	0.51	0.51
Age	MonthlyIncome	0.498	0.498
TotalWorkingYears	YearsWithCurrManager	0.46	0.46
TotalWorkingYears	YearsInCurrentRole	0.46	0.46
TotalWorkingYears	YearsSinceLastPromotion	0.405	0.405
JobLevel	YearsInCurrentRole	0.39	0.39

Table 4. Top Correlations

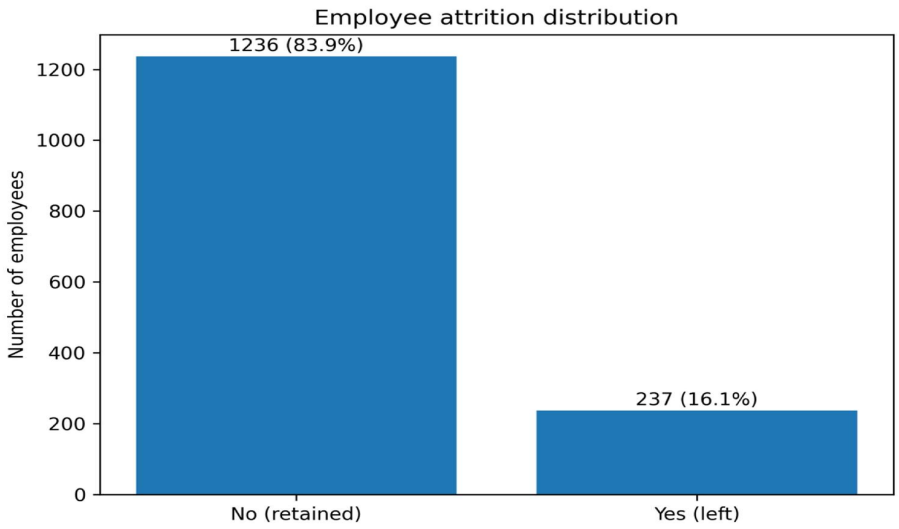

Figure 2. Distribution of employee attrition status, showing the proportion of retained employees and employees who left the organization

The figure shows that 16.1% of the working dataset is labelled as attrition and 83.9% as retained, indicating a moderately imbalanced target.

Figure 2 illustrates the distribution of employee attrition within the analytical dataset. The results indicate that 83.91% of employees remained with the organization, whereas 16.09% were classified as having left. The distribution demonstrates that employee retention represents the dominant outcome, although a meaningful proportion of employees experienced turnover.

The unequal distribution of the two outcome categories is particularly important for subsequent predictive modelling. Models developed using this dataset must account for the comparatively smaller attrition class to avoid systematically favouring employee retention predictions.

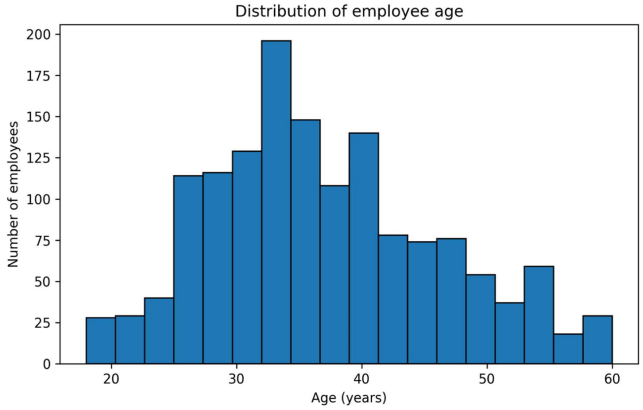

Figure 3. Histogram illustrating the distribution of employee age within the HR Analytics dataset

The histogram summarizes the age profile and allows inspection of concentration, spread and possible tail behavior.

Figure 3 presents the frequency distribution of employee age. The distribution is characterized by a concentration of observations in the working age population, with employee ages ranging from 18 to 60 years. The mean age of 36.92 years and median age of 36 years indicate that the workforce is predominantly composed of early and middle career employees.

The histogram provides an overview of workforce demographic composition and helps identify the concentration and spread of employee ages. This distribution is important because age may be associated with professional experience, organizational tenure, career progression and employee retention.

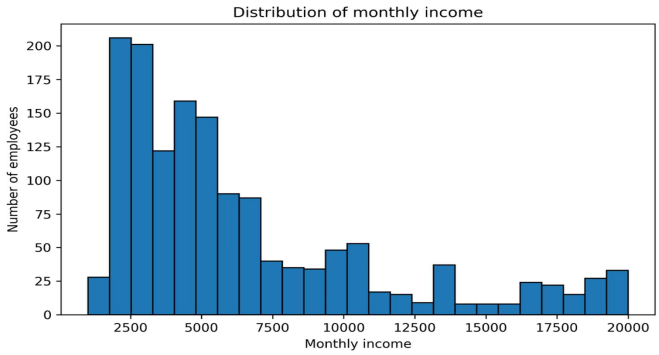

Figure 4. Distribution of monthly employee income illustrating the variability and positive skewness of compensation across the workforce

The histogram displays the spread of employee monthly income. The mean exceeds the median, consistent with a right skewed distribution.

Figure 4 illustrates the distribution of employee monthly income. The mean monthly income was 6,500.23, while the median was 4,908, indicating that the distribution is positively skewed. This pattern suggests that a relatively smaller proportion of employees receive substantially higher compensation than the majority of employees.

The observed income variation may reflect differences in organizational hierarchy, job responsibilities, occupational specialization and professional experience. The income distribution also provides a descriptive foundation for examining compensation inequalities and their potential relationship with employee attrition.

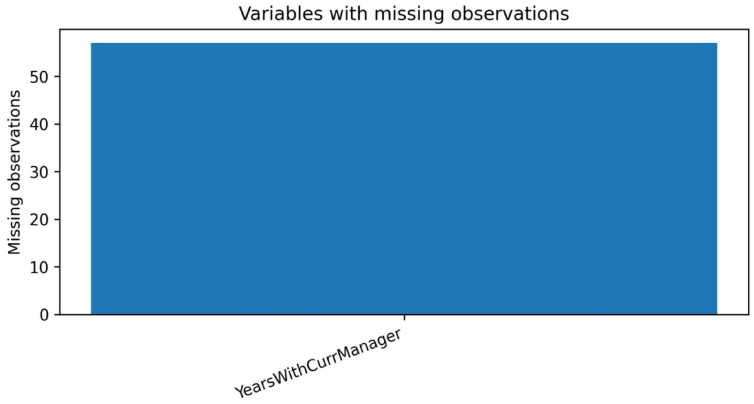

Figure 5. Distribution of missing observations across HR dataset variables, highlighting missing information in YearsWithCurrManager

Missingness is isolated to YearsWithCurrManager, affecting 57 records (3.85% of the working sample).

Figure 5 presents the distribution of missing observations across the dataset variables. The missingness assessment indicates that all 57 missing observations are concentrated in YearsWithCurrManager, accounting for approximately 3.85% of the analytical sample.

The absence of missing observations in other variables indicates relatively high completeness across the remaining dataset dimensions. However, the concentration of missing information in managerial tenure data may influence analyses examining the relationship between supervisory continuity and employee retention. The figure demonstrates that missing value treatment should be variable specific rather than implemented through indiscriminate deletion of incomplete observations.

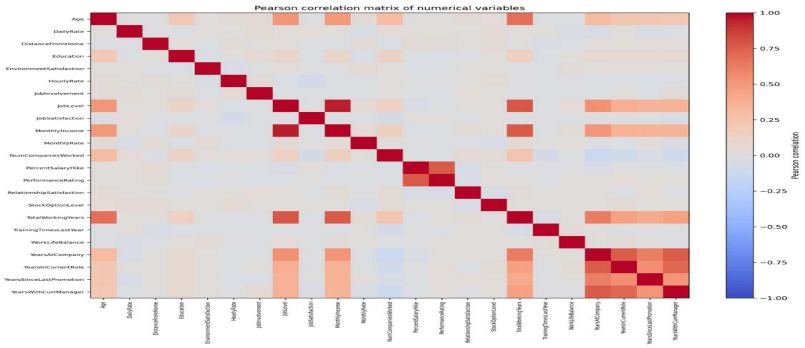

Figure 6. Pearson correlation heatmap illustrating the strength and direction of linear relationships among numerical HR variables

The heatmap visualizes linear associations among numerical variables. Correlations should be interpreted as associations, not causal relationships.

Figure 6 presents the Pearson correlation matrix of numerical variables, providing a comprehensive visualization of the direction and magnitude of linear associations among employee characteristics.

The strongest positive relationship occurs between JobLevel and MonthlyIncome (\(r=0.950\)). Other notable relationships include JobLevel and TotalWorkingYears (\(r=0.782\)), Percent Salary Hike and PerformanceRating (\(r=0.773\)), and MonthlyIncome and TotalWorkingYears (\(r=0.772\)).

The positive associations among organizational tenure, current-role tenure and managerial tenure further demonstrate the interconnected nature of employee career characteristics.

These relationships suggest that employee compensation and career progression are closely related dimensions of workforce structure.

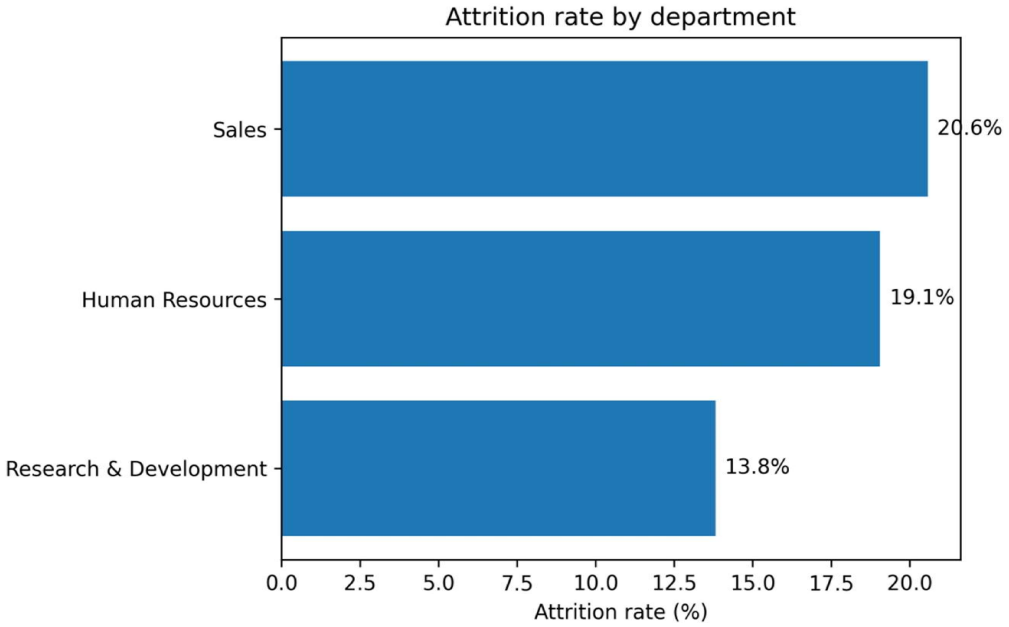

Figure 7. Comparison of employee attrition rates across Sales, Human Resources and Research and Development departments

Sales has the highest observed department attrition rate (20.6%), followed by Human Resources (19.0%) and Research & Development (13.8%).

Figure 7 compares employee attrition rates across organizational departments. Sales exhibited the highest attrition rate (20.58%), followed by Human Resources (19.05%) and Research and Development (13.81%).

The observed differences indicate that employee retention patterns vary across organizational units. The comparatively higher attrition rate in Sales may reflect differences in job demands, performance expectations, working arrangements and compensation structures.

Although Research and Development accounts for the largest number of departing employees, its within-department attrition rate is comparatively lower. This demonstrates the importance of distinguishing employee counts from proportional turnover rates.

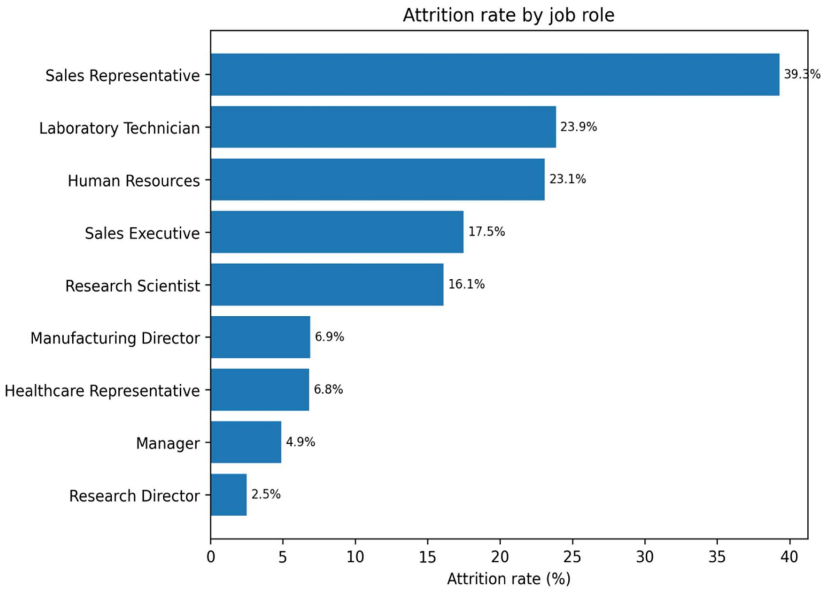

Figure 8. Job-role-wise comparison of employee attrition rates, identifying occupational groups with relatively high and low turnover

Sales Representatives have the highest observed attrition rate (39.3%); Research Directors have the lowest (2.5%). Small role-specific sample sizes should be considered.

Figure 8 illustrates the variation in employee attrition across occupational roles. Sales Representatives demonstrate the highest attrition rate (39.29%), followed by Laboratory Technicians (23.85%) and Human Resources employees (23.08%).

In contrast, Research Directors exhibit the lowest attrition rate (2.50%), followed by Managers (4.90%) and Healthcare Representatives (6.82%).

The substantial variation across job roles indicates that workforce retention is not uniform throughout the organization. Occupational responsibilities, career opportunities, compensation and job related working conditions may represent relevant factors for further investigation.

The comparatively high turnover among Sales Representatives is particularly noteworthy and identifies this occupational group as an important subject for subsequent retention analysis.

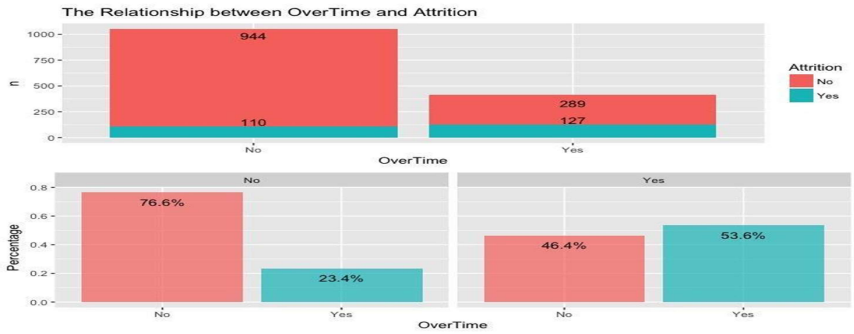

Figure 9. Comparison of employee attrition rates according to overtime status, illustrating higher observed turnover among employees reporting overtime

Employees reporting overtime have a 30.5% attrition rate compared with 10.4% for those not reporting overtime.
Figure 9 compares attrition rates between employees who reported overtime and those who did not.

The results demonstrate that employees reporting overtime experienced an attrition rate of 30.53%, compared with 10.41% among employees without overtime.

The difference indicates a substantial observed association between overtime status and employee turnover. Employees reporting overtime exhibit approximately 2.93 times the attrition rate of employees without overtime.

This finding highlights overtime status as an important organizational characteristic for further investigation. The relationship may be associated with workload demands, work-life balance and employee

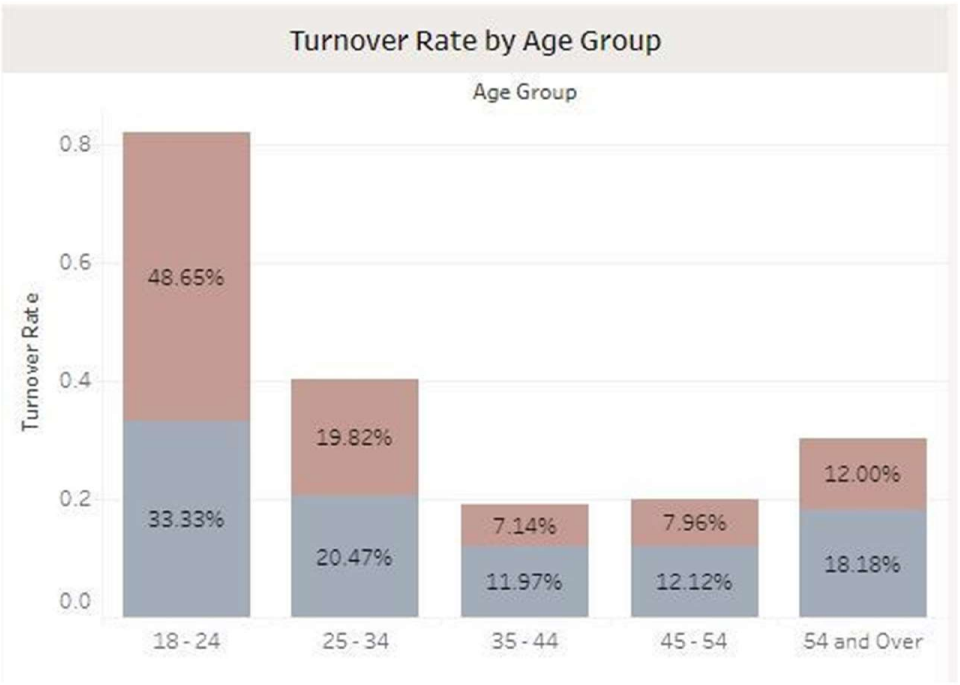

Figure 10. Age-group-wise employee attrition rates illustrating differences in workforce turnover across demographic age categories

The 18–25 group has the highest observed attrition rate (35.8%), while the 36–45 group has the lowest (9.1%).

Figure 10 presents the variation in employee attrition across different age groups. The highest attrition rate is observed among employees aged 18–25 years (35.77%), followed by the 26–35 age group (19.11%).

Employees aged 36–45 years demonstrate the lowest attrition rate (9.15%). The attrition rate subsequently increases to 11.50% among employees aged 46–55 years and 17.02% among those aged 55 years and above.

The distribution demonstrates that employee attrition does not follow a strictly increasing or decreasing relationship with age. Instead, the observed pattern suggests differences in turnover across career stages.

The relatively high attrition among younger employees identifies early-career retention as an important area for organizational investigation.

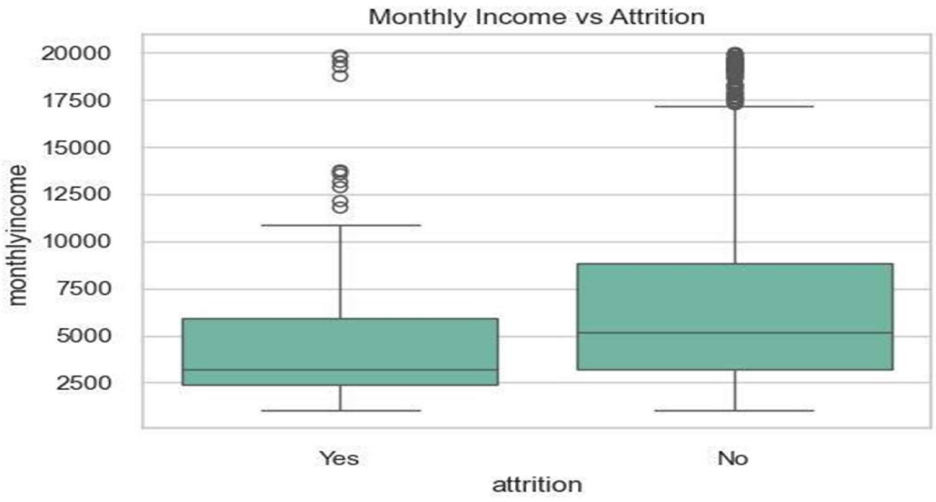

Figure 11. Boxplot comparison of monthly employee income between retained employees and employees who left the organization

The boxplots compare income distributions for retained and departing employees; visual differences require statistical testing and adjustment for job level and role.

Figure 11 compares monthly income distributions between employees who remained with the organization and those who left. The boxplot provides a visual comparison of the median, interquartile range and variability of compensation across the two employment-outcome categories.

Differences in the central distributions may indicate that compensation is associated with employee retention. However, monthly income is closely related to job level, professional experience and occupational role. Consequently, any observed difference between attrition groups may partly reflect differences in workforce composition.

The figure therefore provides a preliminary visualization of compensation related differences that should be examined using inferential statistical methods and multivariable regression.

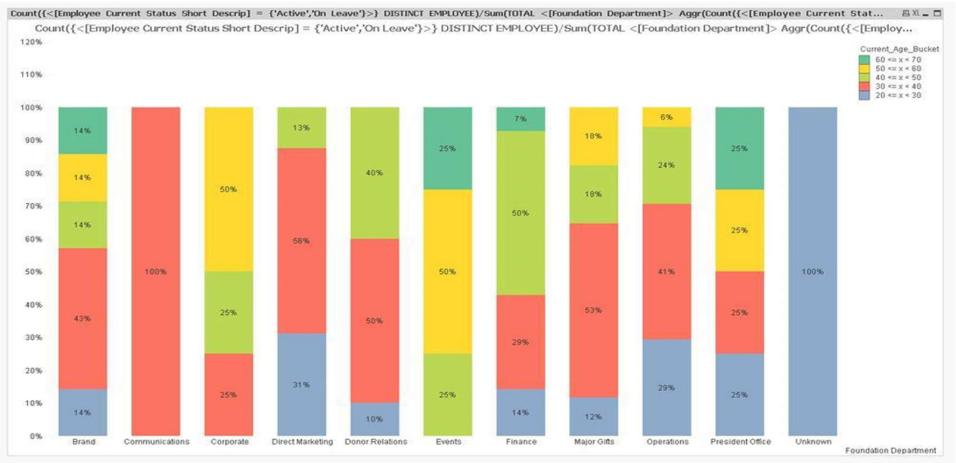

Figure 12. Stacked-bar representation of retained and departing employees across age groups, illustrating the relative composition of employment outcomes within each demographic category

The stacked bars show the proportion retained and leaving within each age group, emphasizing differences in group-specific attrition composition.

Figure 12 illustrates the relative proportions of retained and departing employees within each age group using a stacked-bar representation.

The visualization complements Figure 9 by showing the composition of employment outcomes within each demographic category rather than presenting only the attrition rate.

The proportion of employees leaving is comparatively higher among younger employees, particularly those aged 18–25 years. In contrast, the 36–45 age group demonstrates a greater proportion of retained employees. The stacked-bar visualization makes differences in workforce retention across age categories easier to interpret and highlights the demographic variation observed in the dataset.

5. Discussion of the Results

The exploratory data analysis and data quality assessment provide a foundational understanding of the workforce dynamics and the statistical realities of employee attrition within the organization. The results highlight three critical dimensions: data integrity, demographic and occupational risk factors, and the structural drivers of retention.

5.1 Data Quality and Preprocessing Imperatives Aligning with Carvalho's observations on the challenges of real world data collection, the initial dataset assessment revealed inherent noise and structural inconsistencies. The identification of seven exact duplicate rows, three conflicting employee identifiers, and 57 missing values exclusively concentrated in YearsWithCurrManager (3.85% of the sample) demonstrates that raw HR data is rarely ready for immediate inferential modeling. The decision to retain conflicting identifiers pending source verification, rather than applying indiscriminate deletion, reflects a necessary commitment to data integrity. These findings confirm that an automated, variable-specific preprocessing pipeline is essential to filter noise, handle missing managerial tenure data, and manage statistical outliers (e.g., in MonthlyIncome and TotalWorkingYears) without discarding legitimate extreme values.

5.2 Demographic and Occupational Vulnerabilities The EDA uncovered stark disparities in attrition rates across different employee cohorts, validating the need for targeted retention strategies. The overall attrition rate of 16.09% masks severe localized risks. Sales Representatives exhibit a critical attrition rate of 39.3%, and the Sales department as a whole leads organizational turnover at 20.6%. Furthermore, employees reporting overtime face an attrition rate nearly three times higher than their non-overtime counterparts (30.5% vs. 10.4%). Demographically, the youngest cohort (18–25 years) shows the highest turnover (35.8%), suggesting that early career retention and onboarding processes require urgent strategic intervention. These unadjusted descriptive associations align with literature emphasizing that work environment, job demands, and people management skills are strong predictors of resignation.

5.3 Compensation, Career Progression, and Model Implications The analysis of numerical variables revealed a positively skewed income distribution, reflecting hierarchical compensation structures. The Pearson correlation matrix highlighted a near perfect linear relationship between JobLevel and MonthlyIncome (r = 0.95), alongside strong correlations involving TotalWorkingYears and YearsAtCompany. This indicates that compensation and career mobility are deeply intertwined dimensions of workforce structure. For predictive modeling, these multicollinearities suggest that dimensionality reduction or careful feature selection will be required to prevent model overfitting.

Furthermore, the severe class imbalance in the target variable (83.91% retained vs. 16.09% attrition) dictates that standard accuracy metrics will be misleading. Future machine learning applications must employ techniques like SMOTE and prioritize evaluation metrics such as the F1-Score and ROC-AUC to ensure the minority attrition class is accurately identified.

5.4 Ethical and Governance Implications Finally, the granular insights derived from this EDA such as flagging specific departments or age groups as “high risk” must be handled with ethical caution. As warned by Gélinas et al. and Batool et al., algorithmic decision making can diminish employee autonomy if deployed opaquely. Therefore, the integration of these predictive insights into HR dashboards must be governed by a “privacy by design” framework. Predictive models should not be used to penalize or surveil high risk cohorts (e.g., young sales representatives working overtime) but rather to inform empathetic, systemic interventions regarding workload management, compensation equity, and career development, thereby fostering organizational justice.

6. Conclusion

This study successfully mapped the structural, demographic, and occupational landscape of employee attrition through rigorous data quality assessment and exploratory data analysis. The findings confirm that employee turnover is not a uniform organizational issue but a highly localized phenomenon driven by specific occupational hazards (e.g., Sales roles, overtime), demographic life-stages (e.g., early career employees), and compensation disparities.

However, the transition from descriptive insights to predictive action requires overcoming significant data quality hurdles and methodological challenges, such as class imbalance and multicollinearity. More importantly, this research underscores that the deployment of AI and machine learning in HR cannot be purely a mathematical exercise. To truly optimize business outcomes and retain critical talent, organizations must build predictive frameworks that are intrinsically aligned with ethical AI governance. By combining robust, noise-filtered data pipelines with transparent, “privacy by design” algorithms, HR leaders can shift from reactive turnover management to proactive, equitable, and human centric talent retention strategies.

References

[1] Yahia, N. B., Hlel, J., Colomo-Palacios, R. (2021). From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access, 9*, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559.

[2] Setiawan, I., Suprihanto, S., Nugraha, A. C., Hutahaean, J. (2020, April). HR analytics: Employee attrition analysis using logistic regression. In *IOP Conference Series: Materials Science and Engineering* (Vol. 830, No. 3, p. 032001). IOP Publishing.

[3] Hoffman, M., Tadelis, S. (2018). *People management skills, employee attrition, and manager rewards: An empirical analysis*. National Bureau of Economic Research.

[4] Frye, A., Boomhower, C., Smith, M., Vitovsky, L., Fabricant, S. (2018). Employee attrition: What makes an employee quit? *MU Data Science Review, 1*(1).

[5] Bindra, H., Sehgal, K., Jain, R. (2019). Optimisation of C5.0 using association rules and prediction of employee attrition. In *International Conference on Innovative Computing and Communications*. Lecture Notes in Networks and Systems. Singapore.

[6] Raja, D. V. A. J., Kumar, R. A. R. (2016). A study to reduce employee attrition in IT industries. *International Journal of Marketing and Human Resource Management (IJMHRM), 7*(1), 1–14.

[7] Mishra, S. N., Lama, D. R. (2016). A decision making model for human resource management in organizations using data mining and predictive analytics. *International Journal of Computer Science and Information Security, 14*(5), 217.

[8] Hosmer, D. W., Jr., Lemeshow, S., Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley Sons.

[9] Villar-Rubio, E., Delgado-Alaminos, J., Barrilao-González, P. (2015). Job satisfaction among Spanish tax administration employees: A logistic regression analysis. *Journal of Labor Research, 36*(2), 210–223.

[10] Karhula, K., Puttonen, S., Ropponen, A., Koskinen, A., Ojajärvi, A., Kivimäki, M., & Härmä, M. (2017). Objective working hour characteristics and work–life conflict among hospital employees in the Finnish public sector study. *Chronobiology International, 34*(7), 876–885.

[11] Lohela-Karlsson, M., Hagberg, J., Bergström, G. (2015). Production loss among employees perceiving work environment problems. *International Archives of Occupational and Environmental Health, 88*(6), 769–777.

[12] Halkos, G., & Bousinakis, D. (2017). The effect of stress and dissatisfaction on employees during crisis. *Economic Analysis and Policy, 55*, 25–34.

[13] Merriman, K. K., Sen, S., Felo, A. J., Litzky, B. E. (2016). Employees and sustainability: The role of incentives. *Journal of Managerial Psychology, 31*(4), 820–836.

[14] Carvalho, C., Moreira, R. S., Torres, J. M. (2021). Data Quality Visual Analysis (DQVA): A tool to process and pinspot raw data irregularities. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (p. 1036–1045). IEEE. https://doi.org/10.1109/CCWC51732.2021.9375961.

[15] Iancu, C., Oprea, S. V. (2025). AI and human resources in a literature-driven investigation into emerging trends. *IEEE Access, 13*, 81897–81916. https://doi.org/10.1109/ACCESS.2025.3568338.

[16] Malin, C. D., Fleiß, J., Seeber, I., Kubicek, B., Kupfer, C., Thalmann, S. (2024). The application of AI in digital HRM: An experiment on human decision-making in personnel selection. *Business Process Management Journal, 30*(8), 284–312. https://doi.org/10.1108/BPMJ-11-2023-0884.

[17] Madanchian, M. (2024). From recruitment to retention: AI tools for human resource decision-making. *Applied Sciences, 14*(24), 11750. https://doi.org/10.3390/APP142411750.

[18] Gélinas, D., Sadreddin, A., Vahidov, R. (2022). Artificial intelligence in human resources management: A review and research agenda. *Pacific Asia Journal of the Association for Information Systems, 14*(6), 1–42. https://doi.org/10.17705/1pais.14601.

[19] Batool, A., Zowghi, D., Bano, M. (2025). AI governance: A systematic literature review. *AI Ethics, 5*, 3265–3279. https://doi.org/10.1007/s43681-024-00653.

[20] Hernández, E. G. (2024). *Towards an ethical and inclusive implementation of artificial intelligence in organizations: A multidimensional framework*. arXiv. https://arxiv.org/pdf/2405.01697.

[21] Aizenberg, E., van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*. https://doi.org/10.1177/2053951720949566.