TABLE OF CONTENTS

LIST OF ABBREVIATIONS

No.	Abbreviation	Explanation
1	ANN	Artificial Neural Network
2	BRF	Balanced Random Forest
3	IQR	Interquartile Range
4	MLP	Multi-layer Perceptron
5	Q-Q	Quartile - Quartile
6	SMOTE	Synthetic Minority Over-sampling Technique
7	SVC	Support Vector Classifier
8	SVM	Support Vector Machine
9	WLB	Work-life balance

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1. Rationale of the Research

Employee performance analysis is a critical component of organizational management as it enables companies to assess employee contributions systematically and identify areas for improvement. This process is essential for optimizing workforce productivity and aligning individual objectives with strategic goals, thereby fostering a culture of continuous improvement (Synergita, 2025). Moreover, performance evaluations have been shown to positively impact employee morale and engagement by acknowledging achievements and addressing deficiencies, which can lead to increased job satisfaction and motivation (Indeed, 2025). Empirical evidence suggests that organizations prioritizing performance management tend to outperform their peers financially (McKinsey). However, to ensure its transparency and accuracy, there is a need for a critical approach and an effective support system backed by suitable data collection methods (Muslih, 2022).

2. Research Aims and Objectives

This research aims to contribute to the field of employee performance management by pursuing three primary objectives. Firstly, it seeks to identify the key drivers of employee performance, focusing on both quantitative and qualitative factors that significantly influence individual and organizational outcomes. Secondly, the study aims to develop a predictive model for employee performance using advanced machine learning techniques, which will enable organizations to forecast performance trends and make informed decisions about talent development and resource allocation. Lastly, the research aims to provide actionable recommendations for improving employee performance based on the insights derived from the predictive model, thereby supporting organizations in designing targeted interventions and strategies to enhance workforce productivity and engagement.

3. Research Scope

The subject of this research is employee performance analysis, with a focus on identifying key performance drivers and developing predictive models. The scope of the study encompasses both theoretical and practical aspects of performance management, including data analysis techniques and machine learning models. The research will explore various organizational contexts to ensure that the findings are applicable across different industries and settings.

4. Research Questions

This research is guided by three primary questions:

- What are the key drivers of employee performance in contemporary organizational settings?
- How can machine learning models be effectively used to predict employee performance based on historical data?
- What recommendations can be derived from data analysis to improve employee performance and enhance organizational productivity?

5. Research Method

The research will employ a mixed-methods approach, combining both qualitative and quantitative techniques. Particularly, a dataset of employee performance metrics will be analyzed using machine learning algorithms to develop a predictive model. The model's accuracy will be evaluated, and insights from the analysis will be used to formulate recommendations for improving employee performance.

6. Contribution of the Research

This research contributes to the existing body of knowledge on employee performance by identifying key drivers, developing a predictive model, and providing actionable recommendations. It builds upon previous studies that highlight the importance of factors such as workplace environment, employee satisfaction, management standards, and training in influencing employee performance (Frontiers, 2022; Dergipark, 2022). The predictive model developed in this study enhances the ability of organizations to forecast performance trends, allowing for proactive talent management and strategic decision-making. Furthermore, by emphasizing data-driven insights, this research supports the creation of personalized development programs and targeted interventions, which can improve job satisfaction and retention (Alwardi, 2023). Overall, this study contributes to the development of more effective employee performance management strategies, aligning with broader organizational goals and fostering a high-performing organizational culture (Mesiya, 2022).

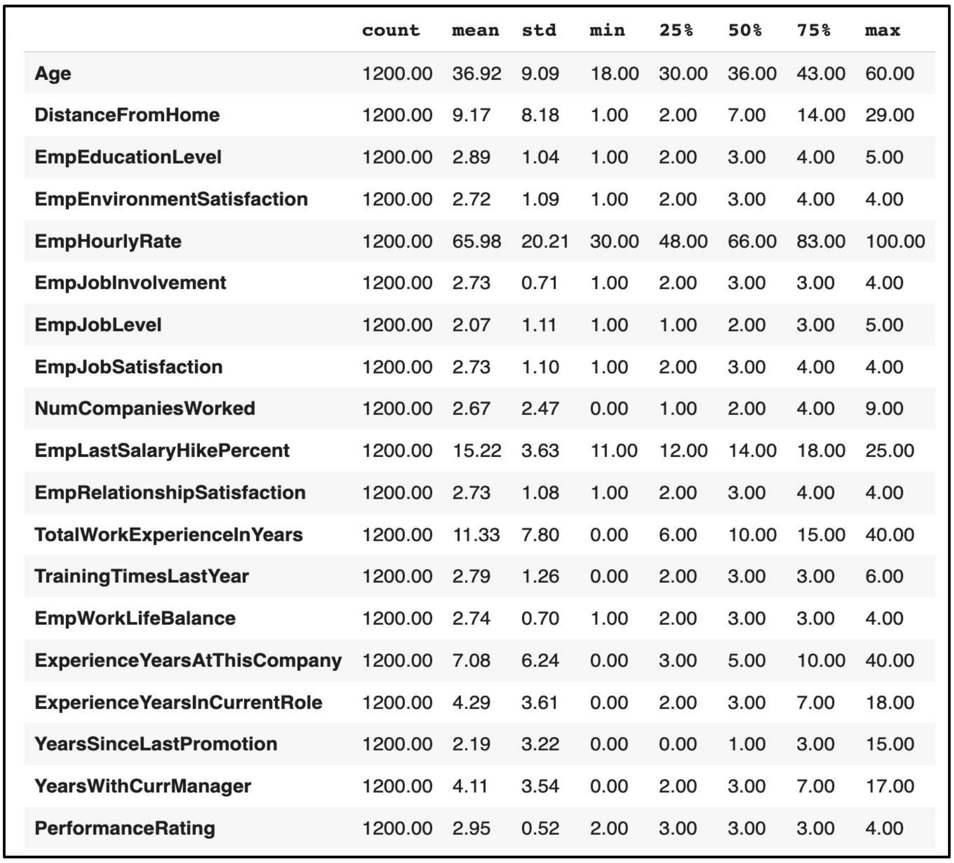
CHAPTER 1: DATA OVERVIEW AND DESCRIPTIVE STATISTICS

1.1. Overview of the Dataset

This paper makes use of an employee performance dataset consisting of 1,200 observations and 28 features. The dataset encompasses both numerical and categorical variables capturing various aspects of employee performance, workplace attributes, and individual characteristics. The primary objective of this analysis is to leverage these features to predict employee performance ratings, which serve as the target variable.

Among 19 quantitative variables, 11 are numerical (e.g., age, hourly rate, years of experience) and 8 are ordinal (e.g., job satisfaction, work-life balance). Additionally, the dataset contains 8 qualitative features, which represent categorical attributes, such as department, marital status, job role, and among others. The "EmpNumber", an alphanumeric identifier, is excluded as it does not contribute to employee performance prediction.

1.2. Descriptive Statistics



	count	mean	std	min	25%	50%	75%	max
Age	1200.00	36.92	9.09	18.00	30.00	36.00	43.00	60.00
DistanceFromHome	1200.00	9.17	8.18	1.00	2.00	7.00	14.00	29.00
EmpEducationLevel	1200.00	2.89	1.04	1.00	2.00	3.00	4.00	5.00
EmpEnvironmentSatisfaction	1200.00	2.72	1.09	1.00	2.00	3.00	4.00	4.00
EmpHourlyRate	1200.00	65.98	20.21	30.00	48.00	66.00	83.00	100.00
EmpJobInvolvement	1200.00	2.73	0.71	1.00	2.00	3.00	3.00	4.00
EmpJobLevel	1200.00	2.07	1.11	1.00	1.00	2.00	3.00	5.00
EmpJobSatisfaction	1200.00	2.73	1.10	1.00	2.00	3.00	4.00	4.00
NumCompaniesWorked	1200.00	2.67	2.47	0.00	1.00	2.00	4.00	9.00
EmpLastSalaryHikePercent	1200.00	15.22	3.63	11.00	12.00	14.00	18.00	25.00
EmpRelationshipSatisfaction	1200.00	2.73	1.08	1.00	2.00	3.00	4.00	4.00
TotalWorkExperienceInYears	1200.00	11.33	7.80	0.00	6.00	10.00	15.00	40.00
TrainingTimesLastYear	1200.00	2.79	1.26	0.00	2.00	3.00	3.00	6.00
EmpWorkLifeBalance	1200.00	2.74	0.70	1.00	2.00	3.00	3.00	4.00
ExperienceYearsAtThisCompany	1200.00	7.08	6.24	0.00	3.00	5.00	10.00	40.00
ExperienceYearsInCurrentRole	1200.00	4.29	3.61	0.00	2.00	3.00	7.00	18.00
YearsSinceLastPromotion	1200.00	2.19	3.22	0.00	0.00	1.00	3.00	15.00
YearsWithCurrManager	1200.00	4.11	3.54	0.00	2.00	3.00	7.00	17.00
PerformanceRating	1200.00	2.95	0.52	2.00	3.00	3.00	3.00	4.00

Figure 1.1: Summary Statistics of the Dataset

Source: Authors' data own processing (2025)

The average age of employees is 37 (36.92) years old (SD = 9.09), ranging from 18 to 60, which shows that the company employs both early-career and experienced workers. Commute distances vary significantly (Avg = 9.17 km, SD = 8.18), with commutes ranging from 1 to 29 km. This indicates that while some employees live closer to the workplace, other employees have much longer commutes.

This dataset contains metrics related to engagement and job satisfaction as well. Overall, the employee involvement score lies at 2.73, which is moderate given the range of from 1 to 4. The average level of job satisfaction is noted at 2.73 as well, similar to that of the relationship satisfaction among employees. Due to low standard deviations (around 1.1), it is clear that most employees are within a similar level of fulfillment when it comes to these parameters. The environment satisfaction lying at 2.72, indicates that employees have a neutral to positive sentiment towards the comfort and employees within their workplace.

Patterns of career advancement are quite different. The average hourly pay is $65.98 (SD = 20.21), ranging from $30 to $100 per hour. Salary hikes average 15.22%, with increments from 11% to 25%. Employees have an average of industry experience of 11.33 years but only 7.08 years with the current company. Role change occurs around every 4.29 years, but some workers go up to 15 years without a promotion, suggesting a lack of career growth for some.

On the 2 to 4 scale of performance ratings, the mean is 2.95 with SD 0.52 suggesting that most employees have comparable evaluations which is quite stable. Finally, training participation is moderate because attendance at training sessions per employee averages 2.79 per year with a range of 0 to 6 times last year, which may limit growth and development in a career.

The visualization of data in the analysis highlights some key aspects of the employees, their level of satisfaction, the compensation, and the career progression. Connecting these aspects will further enrich our understanding of what drives job satisfaction, retention, and performance.

CHAPTER 2: EXPLORATORY DATA ANALYSIS

2.1. Univariate Analysis

2.1.1. Categorical Variables

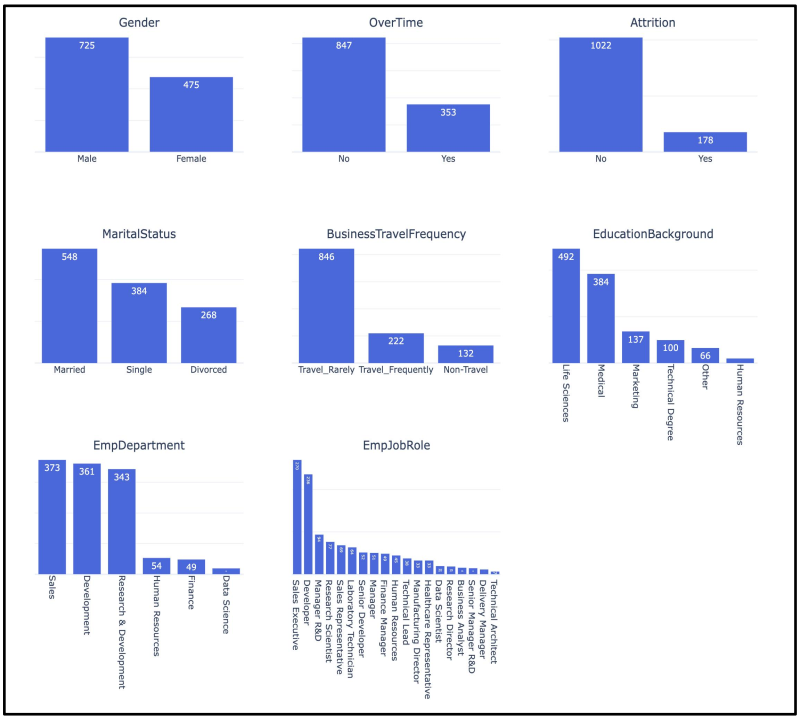


Figure 2.1: Distribution of Categorical Variables

Source: Authors' data own processing (2025)

Based on the gender distribution, it is apparent that there are more male than female employees which may have some effects on workplace relations and diversity policies.

The majority of employees not working overtime indicates that for most people, there is some level of work-life balance. Nonetheless, a significant portion of employees do work overtime, which may have several implications for their job satisfaction and productivity.

Employees' attrition rates imply that most of them are staying with the company, although some have left. The causes of employee turnover or retention due to job discontent, advancement possibilities, or a heavy workload could provide useful guidance on retention plans.

The most prominent married group is followed by single and divorced groups. This may have consequences associated with work-life balance as well as policies and benefits catered for different life stages.

The majority of employees travel infrequently while a small proportion travel frequently. Employees who travel frequently may have higher demands of work requiring them to be more engaged or may be more stressed.

Analysis of academic backgrounds show that a large number of employees have qualifications from Life Sciences and Medical fields while less employees have qualifications in Marketing, Technical areas, or Human Resources. This means that employees, to a large extent, specialize in scientific and technical areas.

In terms of the distribution among the departments, greater employment work in Sales, Development and Research and Development and less employment work in Human Resource and Finance department. It seems that this distribution corresponds to the company priorities and its main directions of work.

Lastly, job roles vary greatly, with a concentration in executive and technical jobs, compared to more specialized positions with less employees. This pattern of employment suggests that the organization emphasizes research, sales, and operation.

These categorized observations help analyze the workforce structure, potential organizational problems, and identify possibilities for improving employee motivation and retention.

2.1.2. Numerical Variables

According to the 'Performance Rating' chart, most employees are rated 3. Ratings of 2 and 4 are outliers, while none of the employees are rated 1 or 5, which indicates a narrowed evaluation scale.

'Employee Environment Satisfaction' is mostly rated 3 or 4, suggesting general satisfaction. The box plot suggests a balanced distribution with no major outliers.

'Employee Last Salary Hike Percent' range between 11% and 15%, with a gradual decline in higher hikes, creating left-skewed distribution with some outliers at the upper end.

The 'Employee Work-Life Balance' is moderate (average = 3), showing that most employees are satisfied with their work life balance. The box plot indicates a concentrated distribution with no significant outliers.

'Experience Years in Current Role' right-skewed, showing that most employees have 5 years or less. This figure decreases with greater experience. There are several outliers as some employees have remained in the same role longer than 15 years.

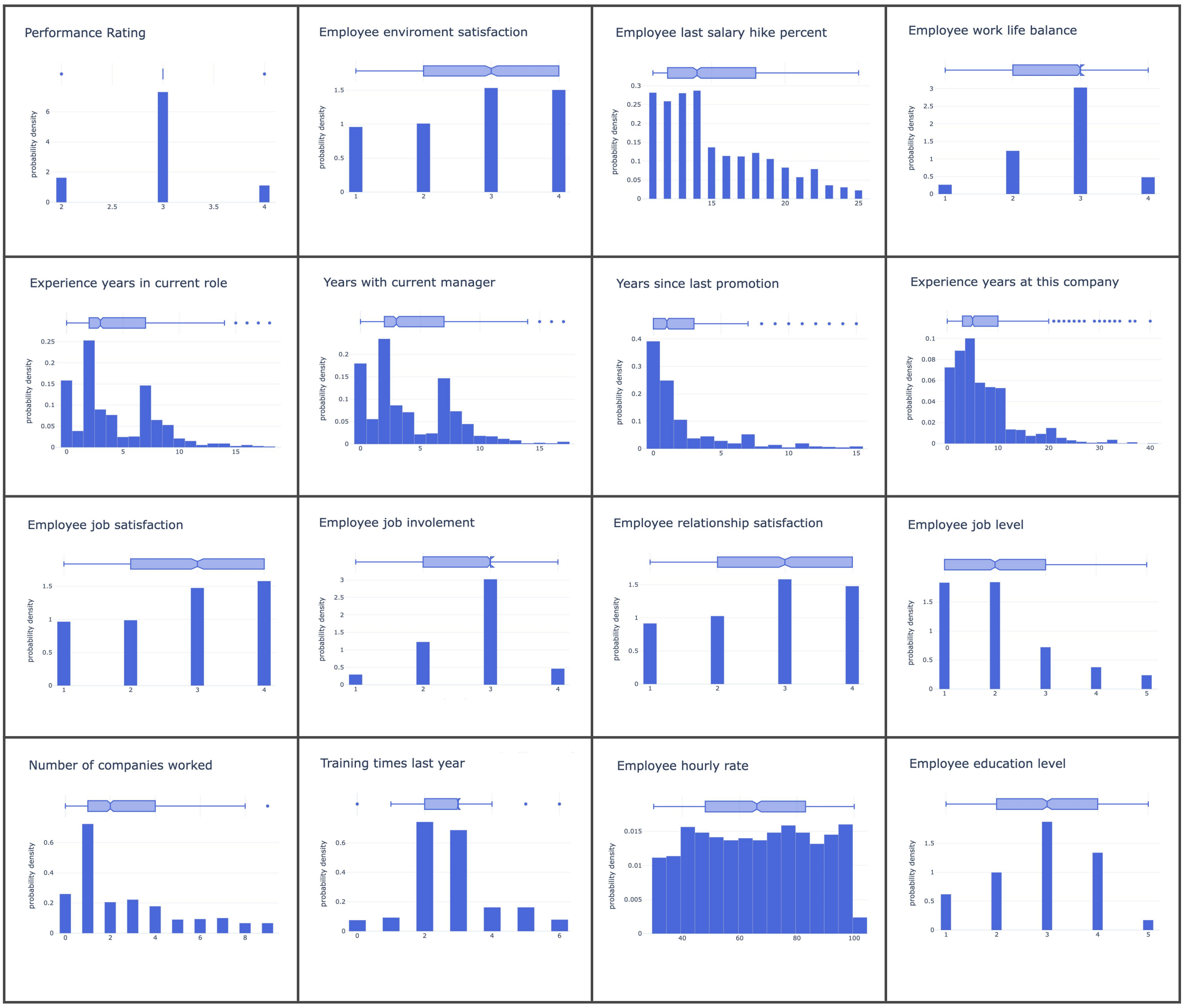


Figure 2.2: Distribution and Boxplots of Selected Numerical Variables

Source: Authors' data own processing (2025)

'Years with Current Manager' chart shows frequent managerial changes, with a very small number of employees having the same manager for over 15 years. There are some outliers as many employees have remained under the same leadership for an unusually long time.

The 'Years Since Last Promotion' chart demonstrates that most employees were promoted within the last 2 years. The likelihood of promotion is higher in initial years. The distribution is right-skewed, with many outliers beyond ten years, highlighting career stagnation for some.

'Experience Years at The Company' follows a highly right-skewed pattern, with most employees having less than 10 years, with a sharp decline indicating fewer long-tenured employees. The box plot highlights several outliers with some employees having up to 40 years of experience.

'Employee Job Satisfaction' remains fairly constant, with a slight preference in higher levels (3 and 4), with no significant outliers.

'Employee Job Involvement' is mostly moderate (many rated as 3), followed by rating 2. The box plot confirms the spread of responses is stable as no extreme outliers exist.

'Employee Relationship Satisfaction' presents a fairly even distribution, with satisfaction ratings of 3 and 4 being more common, without significant outliers.

'Employee Job Level' is right-skewed, with most workers in the junior levels (level 1 & 2) and fewer in senior rank (level 3 & 4). No extreme outliers are apparent.

The 'Number of Companies Worked' chart illustrates that many employees have worked at a single company, with the declining trend as previous company count rises. The box plot features one outlier with extensive job-hopping.

The 'Training Times Last Year' chart shows that most had either 2 or 3 training sessions, with very few having 0 and some having 5 or 6 as outliers.

The 'Employee Hourly Rate' chart demonstrates a uniform distribution with no extreme values which means that income levels are spread across the scale.

'Employee Education Level' is evenly distributed, with levels 3 and 4 being the most common, while fewer have levels 1 or 5, with no clear outliers.

2.2 Bivariate and Multivariate analysis

2.2.1. Overview of Employee Performance Rating

Table 2.1: Distribution of Employee Performance Ratings

PerformanceRating	Count
3	874
2	194
4	132

Source: Authors' data own processing (2025)

An analysis reveals a distribution of performance rating using the .value_counts() function, with Rating 3 being the most common (874 employees), making up about 72.8% of the dataset. Meanwhile, the figure for Rating 2 and Rating 4 is much lower, accounting only 194 and 132 employees, respectively. The significant gap observed from the given dataset reveals that there are further potential insights.

Next, the authors categorized employees based on two basic demographic factors, which are gender and age. First, by determining the minimum and maximum ages, our group segmented the dataset into five age groups as the provided bar chart, using *pd.cut()* function, allowing for a structured comparison that covers different age categories. Subsequently, we grouped the given dataset based on gender and performance rating and used *unstack()* for counting occurrences. Similar approach was also applied to the age group. Missing values (NaN) were replaced with 0 by using.fillna(0) to ensure comprehensiveness. Finally, the group of authors visualized the result by a pie chart and bar chart, putting them together for easier comparison.

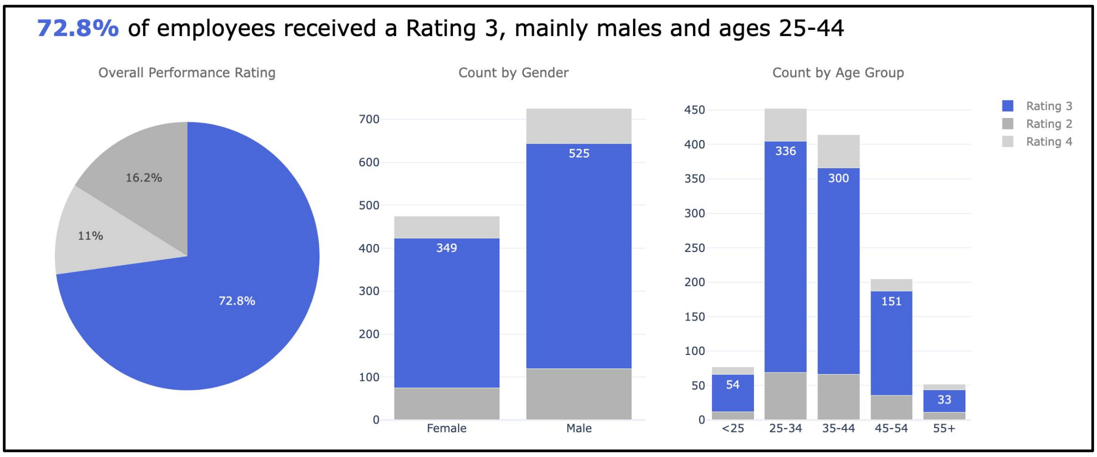


Figure 2.3: Overview of Employee Performance Ratings

Source: Authors' data own processing (2025)

Overall, the pie chart reveals that the distribution of employee performance ratings shows a strong concentration at Rating 3, accounting for 72.8% of the dataset, suggesting deeper insights into the efficiency of the organization’s evaluation system. When the system rates nearly three quarters of employees the same rank, it means that the organization fails to adequately distinguish between different levels. However, the tendency to give employees mid-range levels to avoid conflicts by some managers may weaken motivation for higher performers to contribute their passion to the work because of being undervalued.

Breaking down the result by gender, due to the difference in number of employees by sex, we found that the number of male and female employees receiving Rating 3 witnesses a huge discrepancy, with the former comprising 525 people and the latter comprising only 349 people. Furthermore, as illustrated in the last chart, the figure for employees aged 25-34 receiving Rating 3 surpasses the other age groups. It can be explained in several ways. During such an age range (25-44), the employees have accumulated sufficient experience to perform steadily. In contrast, the younger age group (<25) shows a lower performance, possibly due to being in the adaptation stage. Similar trend is also true for older employees (55+), which is possibly attributed to the adaptation to technological changes, physical and cognitive demands or different evaluation criteria.

2.2.2. Employee Performance by Department & Job Involvement

To better understand and uncover insights from employee performance data, the group of authors first calculated the total of performance ratings for each department. To maintain the chart's clarity and simplicity, we categorized all other departments under "Others", visualizing the final results on the pie chart below. Taking one step further, the authors also examined the average performance rating of employees within departments in a gender-based approach with a view to conducting a more comprehensive analysis.

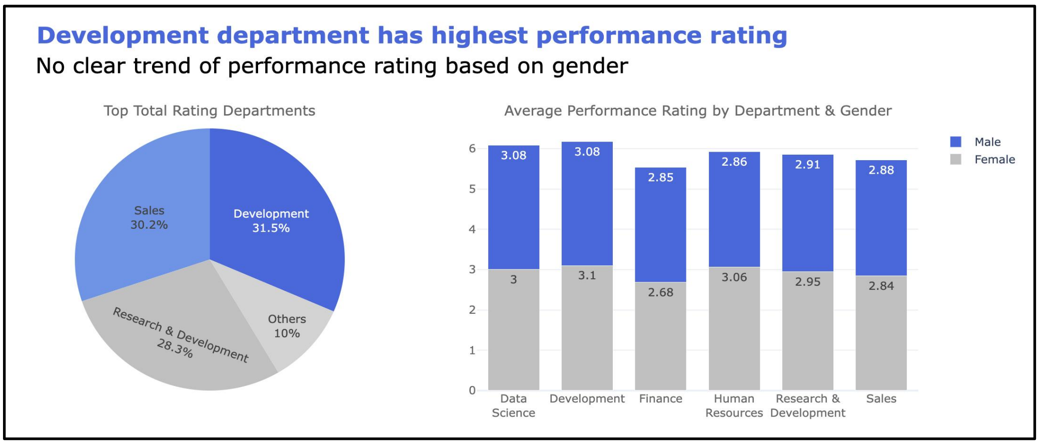


Figure 2.4: Department-Wise Employee Performance

Source: Authors' data own processing (2025)

From the given pie chart, it is evident that the 'Development' department holds the highest performance rating, accounting for 31.5% of the company's overall productivity. Following closely behind are the Sales (30.2%) and Research & Development (28.3%) department. Collectively, these three divisions take up nearly 90% of the total employee performance ratings, highlighting their strategic importance within the organization and their critical role in driving the company's success. The dominance of these three departments can be attributed to the nature of work, which is closely tied to revenue generation, product innovation, relationship building and sustainability-driven effort (the core pillars of any business), and to more performance-based incentives may be present.

The provided stacked bar chart illustrates the average gender-based performance rating across various departments. However, no significant gap between the female and male performance rating is evident, notwithstanding minor variations. The Development department, despite having the highest total performance rating, maintains a relatively balanced between genders. This finding suggests that gender does not play a decisive role in evaluating the performance rating within the divisions. Any minor disparity observed can be due to other variables such as work experience, department-specific demands rather than gender-based discrepancy. From the chart we can clearly observe that the largest gap is in the Finance division, in which the performance rating of both female and male is lower than the others, at 2.68 and 2.85 respectively, resulting in a disparity of 0.17. Despite small fluctuations, having the lowest performance rating indicates that further investigation should be conducted to analyze the reason behind this lower rate to take timely actions.

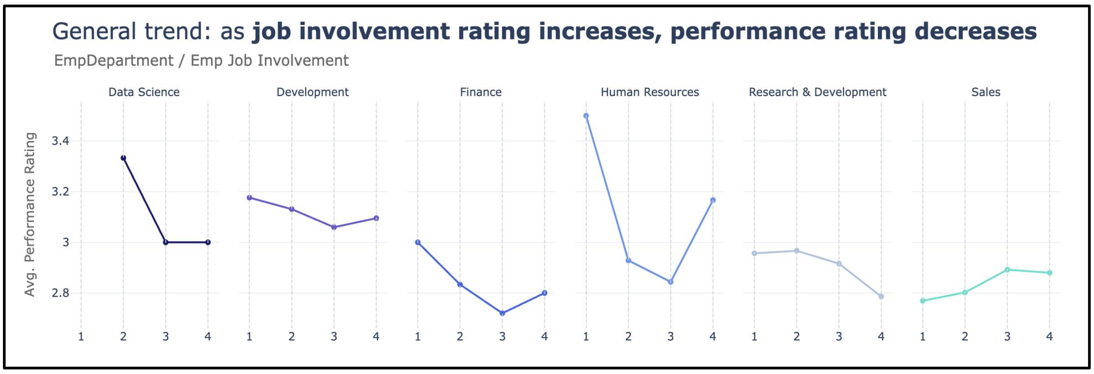


Figure 2.5: Employee Performance Across Departments, by Job Involvement

Source: Authors' data own processing (2025)

Job involvement is considered to be an important intrinsic factor contributing to employee performance. It can be defined as the degree to which employees “emotionally and cognitively” invest in their work (Hngoi et al., 2024) with a perception that it occupies a major position in their life interest (Dubin, 1956). Similar view was also supported by Lodahl & Kejner (1965), Signh & Gupta (2015) and Salessi & Omar (2019).

In the provided chart, an obviously contradictory trend can be observed. Generally, as the job involvement rating within departments increases, performance rating experiences a downward trend, highlighting that higher engagement levels do not necessarily result in higher productivity. Greenhalgh and Rosenblatt (1984) indicated that the heavy workload potentially causes anxiety for employees and drives them to distraction, which negatively impacts individual performance (called "productivity paradox"). However, there is limited research on the relationship between job involvement and employee performance, or if yes, they produced divergent results, suggesting that the relationship between these two variables may vary depending on contextual factors.

The sales department stands out as the only department in the chart witnessing an upward trend in employee performance since the job involvement increases. The driving force behind this increase can be attributed to a widely held perception. It is usually perceived among people working in the Sales department that higher involvement would directly lead to higher sales, more client acquisition and more deal closures. Another reason can be due to the nature of work in this department, which follows a commission-based structure that serves as a strong motivator. Generally, a higher level of involvement could result in more financial rewards, making the positive correlation between involvement and performance become more apparent and significant.

2.2.3. Employee Performance & Workplace Factors: Job Level, Work-Life Balance & Job Satisfaction

After assessing how job engagement influences the performance within an organization, the authors also examine other factors regarding the workplace. These factors include job level, work-life balance, and job satisfaction. These correlations are adequately demonstrated by the writers in three different above bar charts. Overall, it is evident that job level has a negative correlation with average performance rating while work-life balance and environment satisfaction witness a gradual and significant increase, suggesting that these workplace-based determinants have various degrees of impacts on their individual performance.

To be more specific, observing from the first bar chart that there is a slight decrease in the average rating as the job level increases. In other words, moving up to a higher job level does not directly translate into higher performance. In contrast, employees experienced a reduction in their productivity. Holding a higher job role means that they may deal with extreme job demands or role overload, which could lead to emotional exhaustion.

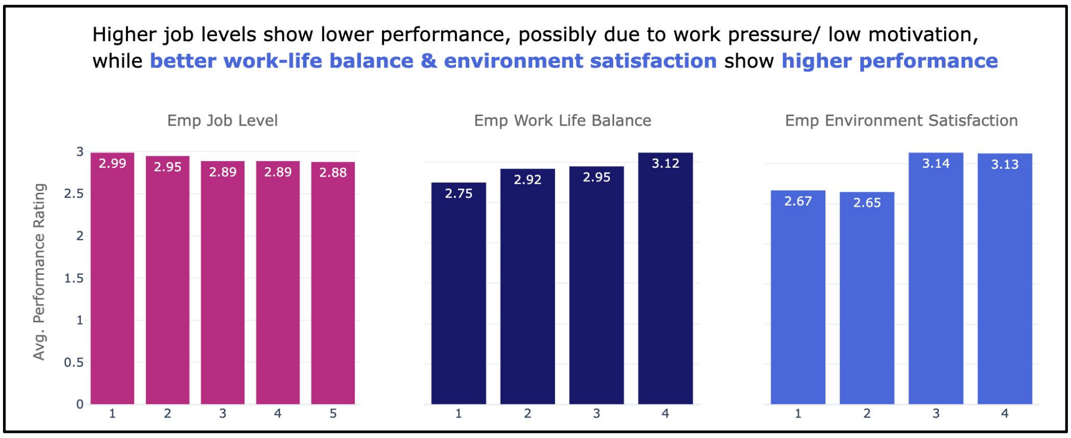


Figure 2.6: Employee Performance Affected by Workplace Factors

Source: Authors' data own processing (2025)

Past researchers also proved that role overload has a strong impact on employee's mental health, such as Janssen et al., (1999), which could potentially result in lower performance rating at work. The general trend of job roles on performance across departments is clearly illustrated as below (Figure 2.7).

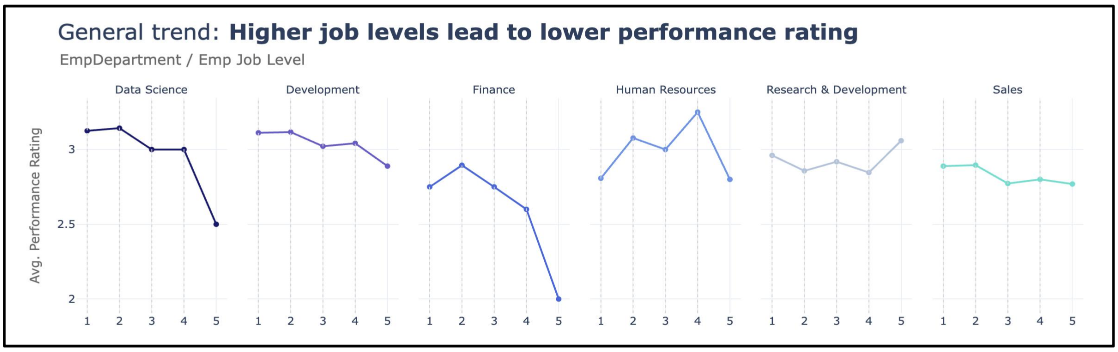


Figure 2.7: Employee Performance Across Departments, by Job Levels

Source: Authors' data own processing (2025)

The opposite trend is true for the two other factors, which are work life balance (WLB) and environmental satisfaction. The two bar charts suggests that there is a positive correlation between two mentioned factors and employee performance. Such a relationship has attracted considerable attention from researchers in the past.

To be specific, according to Scholarios and Marks (2004), WLB plays an important role in forming employee attitudes toward an organization, such as organizational attachment, job satisfaction. In another research, Kanwar et al., (2009) uncovered that there is also a closely positive relationship between job satisfaction and WLB. Perry-Smith et al., (2000) suggested that effort of an organization in adopting more work-family policies is associated with a higher performance. Moreover, family is of great importance in terms of uplifting performance and boosting positive energy at work (Baral, 2009). From the data, we also found a similar trend with these mentioned results, that is employees with higher WLB have a tendency to perform better at their workplace. This finding suggests that organizations should implement flexible work schedules and mental health support programs to obtain more firm-level achievements.

2.2.4. Employee Performance & Career Growth, Workload: Promotions, Salary Hikes & Overtime

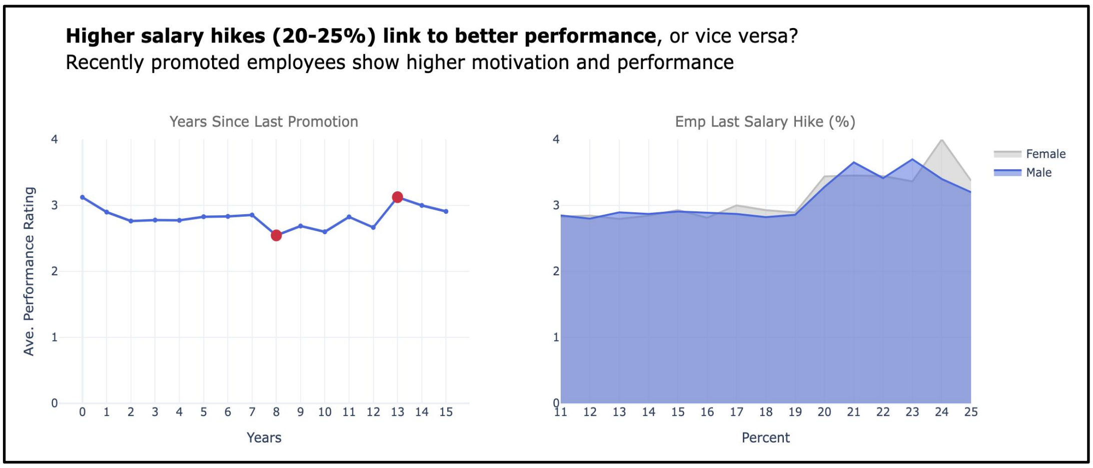


Figure 2.8: Employee Performance Affected by Promotions & Salary Hikes

Source: Authors' data own processing (2025)

Regarding the left line chart on the left, a general downward trend indicates that employees tend to experience a decline in performance over time when promotions are delayed.

According to Appendix 3, the lowest performance rating of 2.545, occurring at 8 years since the last promotion, may suggest a period of burnout, dissatisfaction, or disengagement due to an extended lack of career advancement. In contrast, the highest performance rating of 3.125 is observed at 13 years, indicating that long-term employees develop expertise which sustains their performance and might push harder in hopes of a late-stage career boost.

Employees who were recently promoted exhibit relatively higher ratings, yielding around 3.0. However, there is a notable drop between 0 and 2 years post-promotion, implying that the initial motivation fades relatively quickly. From year 2 onwards, performance ratings remain steady but at a lower level than immediately after a promotion.

Regarding the right stacked area chart and Appendix 4, for salary hikes below 20%, performance ratings remain relatively stable at approximately 3.0 for both genders. This suggests that salary increases within this range are likely determined by factors such as tenure, or job role adjustments, rather than individual performance.

However, a significant shift occurs once the salary hike surpasses 20%. Between 20% and 23%, male employees exhibit slightly higher average performance ratings than their female counterparts. Notably, at 24%, female employees reach their peak performance rating of 4.00, whereas male employees peak slightly lower at 3.40. This could indicate that women receiving this salary hike level presumably demonstrate exceptional performance to be rewarded similarly to men.

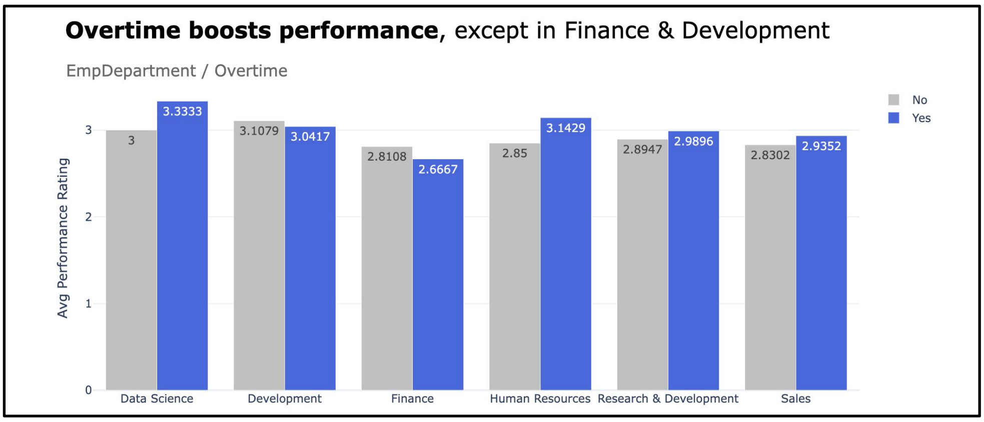


Figure 2.9: Employee Performance Across Departments, by Overtime

Source: Authors' data own processing (2025)

Beyond 24%, performance ratings decline for both genders, dropping to 3.375 for females and 3.20 for males. This implies that, beyond a certain salary hike threshold, the increasing pressure to maintain high performance may outweigh the benefits of the raise, leading to diminished productivity.

Overall, while employees in Data Science, Human Resources, Research & Development, and Sales who work overtime generally exhibit higher performance ratings than those who do not, employees in the Development and Finance departments with overtime are associated with slightly lower performance ratings compared to their counterparts working standard hours.

Notably, Appendix 5 shows that Data Science demonstrates the largest positive gap of 0.333, suggesting that employees benefit from overtime due that are conducive to focused and extended work periods. Danesh and Nourdad (2017) determine that roles requiring creative problem-solving can be enhanced through dedicated time and techniques, which helps explain why Data Science employees achieve a higher average performance rating of 3.333 during overtime.

In contrast, the Finance department illustrates the largest negative gap of -0.144, indicating that employees in financial roles who put in extra hours may risk burnout due to repetitive tasks and intense nature of their responsibilities. Sheng et al. (2019) also found that while moderate time pressure can enhance engagement, excessive time pressure can impair performance, particularly in high-pressure environments where there is a shortage of psychological capital and diminished sleep quality. This helps explain why Finance employees working overtime exhibit the lowest performance rating of 2.666.

2.3. Correlation Analysis

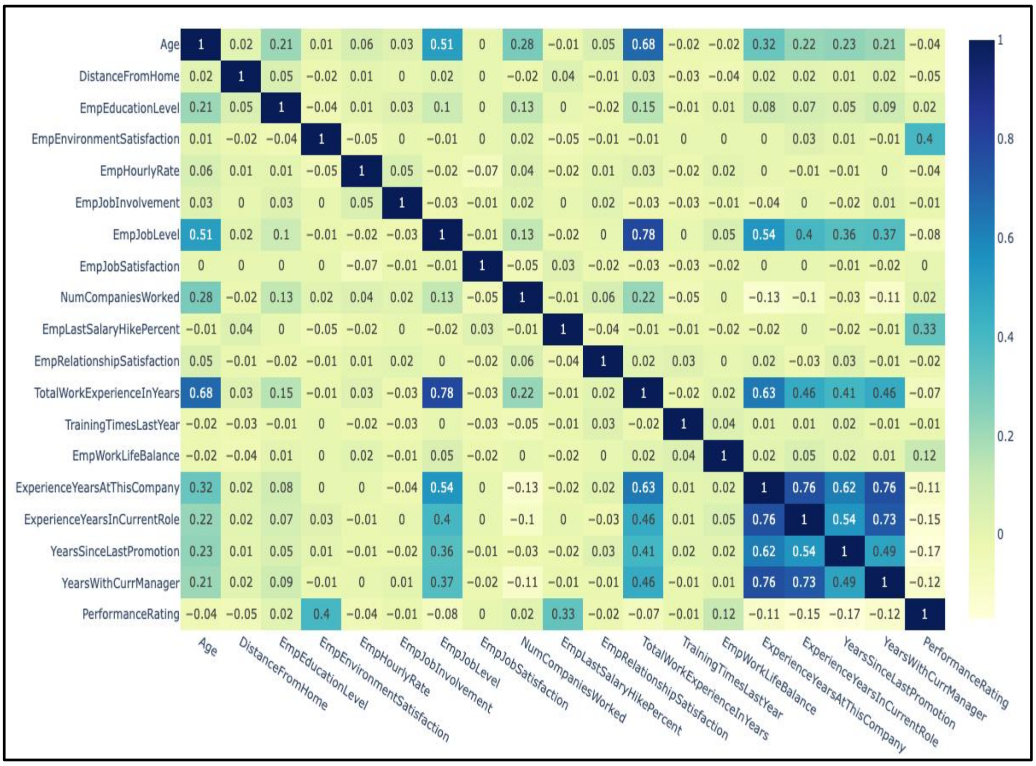


Figure 2.10: Correlation Heatmap (Numerical Variables)

Source: Authors' data own processing (2025)

The correlation matrix computed using the "dataframe's corr()" function reveals several key relationships.

On the positive side, there are strong correlations such as:

EmpJobLevel and TotalWorkExperienceInYears (0.78): Employees with greater total work experience tend to hold higher job levels, which supports the idea that accumulated experience drives timely promotions and career advancement.

ExperienceYearsAtThisCompany and ExperienceYearsInCurrentRole (0.76)/ YearsWithCurrManager (0.76): Employees who have been with the company longer tend to remain in their current roles and have longer working relationships with their current managers, suggesting a stable workforce and low managerial turnover, although it might also indicate limited career mobility within the organization.

On the negative side, 'PerformanceRating' shows slight declines in relation to several factors:

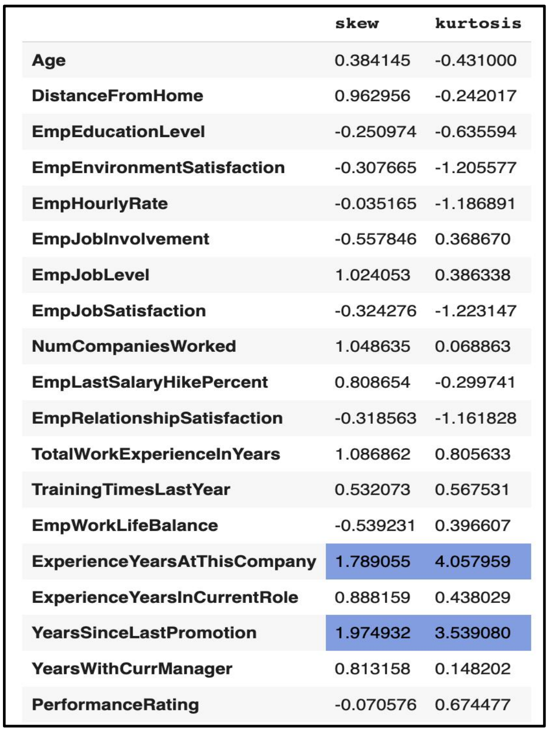
PerformanceRating and YearsSinceLastPromotion (-0.17): As mentioned above, performance ratings decrease as time since last promotion increases. This suggests that employees who have not been promoted for a long period may experience reduced motivation or engagement, potentially leading to performance stagnation or decline.

PerformanceRating and ExperienceYearsInCurrentRole (-0.15): Employees who remain in the same role for longer periods tend to have marginally lower performance ratings, which may indicate that prolonged tenure without advancement can lead to complacency.

Notably, there is no correlation between 'EmpJobSatisfaction' and 'PerformanceRating', suggesting that job satisfaction may be more strongly influenced by factors such as work-life balance, benefits, or team dynamics rather than performance outcomes alone.

2.4. Check Skewness & Kurtosis

Inferred from the table, other variables remain relatively stable with insignificant skewness. Only 'ExperienceYearsAtThisCompany' (4.057959) and 'YearsSinceLastPromotion' (3.539080) exhibit strong positive skewness, suggesting that a large portion of employees have comparatively short tenures or were promoted recently. The skew in 'ExperienceYearsAtThisCompany' may indicate high labor turnover, while the skew in 'YearsSinceLastPromotion' could illustrate limited career advancement opportunities for some employees. Consequently, these two variables are processed using a log transformation in Chapter 3 to reduce skewness and restore a normal distribution within the allowable range.



	skew	kurtosis
Age	0.384145	-0.431000
DistanceFromHome	0.962956	-0.242017
EmpEducationLevel	-0.250974	-0.635594
EmpEnvironmentSatisfaction	-0.307665	-1.205577
EmpHourlyRate	-0.035165	-1.186891
EmpJobInvolvement	-0.557846	0.368670
EmpJobLevel	1.024053	0.386338
EmpJobSatisfaction	-0.324276	-1.223147
NumCompaniesWorked	1.048635	0.068863
EmpLastSalaryHikePercent	0.808654	-0.299741
EmpRelationshipSatisfaction	-0.318563	-1.161828
TotalWorkExperienceInYears	1.086862	0.805633
TrainingTimesLastYear	0.532073	0.567531
EmpWorkLifeBalance	-0.539231	0.396607
ExperienceYearsAtThisCompany	1.789055	4.057959
ExperienceYearsInCurrentRole	0.888159	0.438029
YearsSinceLastPromotion	1.974932	3.539080
YearsWithCurrManager	0.813158	0.148202
PerformanceRating	-0.070576	0.674477

Figure 2.11: Skewness and Kurtosis for Numerical Variables

Source: Authors' data own processing (2025)

CHAPTER 3: DATA PRE-PROCESSING

Before entering into the modelling stage, data preprocessing is a significant step to convert data into meaningful insights. It includes getting the data ready for analysis while guaranteeing its accuracy and suitability for the intended use.

3.1. Missing Value Checking

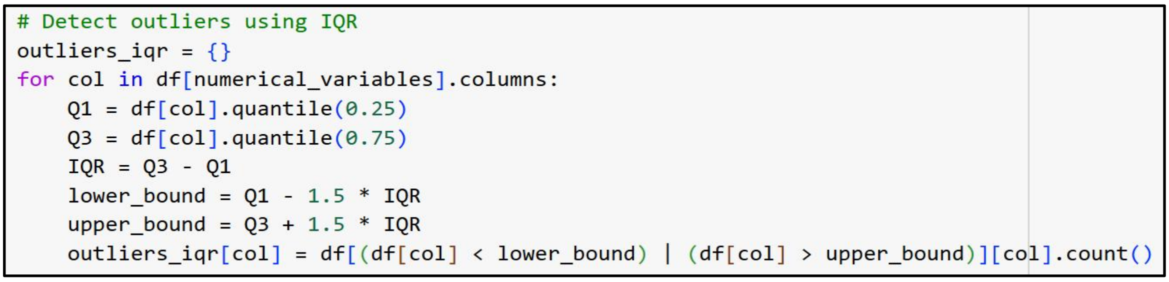
Firstly, missing values happen when no data is stored for a particular variable in an observation (i.e., data entry errors, system malfunctions, or respondents not answering), affecting the model accuracy and insights. Therefore, checking missing values is necessary in order to verify whether the dataset is complete and reliable for analysis, thereby enhancing the model performance and statistical validity.

Using the code “df.isnull().sum()”, the authors get the count of null values in the column. No missing values are found, claiming all the data valid for next steps.

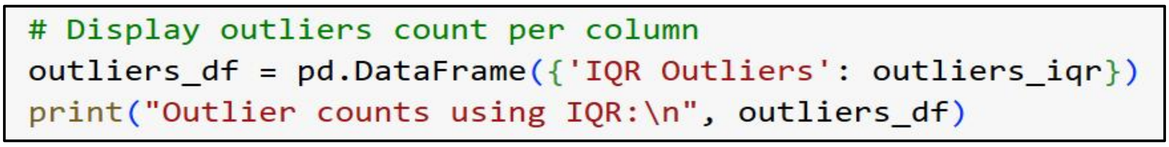
3.2. Outliers Handling

After identifying missing values, the authors begin to detect and handle outliers. Outliers are the extreme data that are significantly different from others, which distort and affect the model prediction.

Here, the authors use the Interquartile Range (IQR) method to identify outliers in the dataset. By setting up a "fence" outside of Q1 and Q3, any values falling beyond 1.5 times the interquartile range (Q1-Q3) are considered outliers.



```python
# Detect outliers using IQR
outliers_iqr = {}
for col in df[numerical_variables].columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers_iqr[col] = df[(df[col] < lower_bound) | (df[col] > upper_bound)][col].count()
```



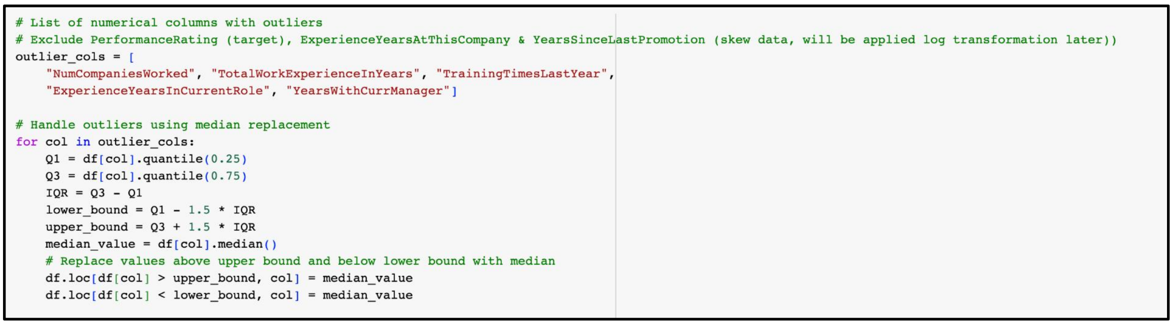
```python
# Display outliers count per column
outliers_df = pd.DataFrame({'IQR Outliers': outliers_iqr})
print("Outlier counts using IQR:\n", outliers_df)
```

Table 3.1: The List of Outliers in the dataset

Variables	IQR Outliers
NumCompaniesWorked	39
TotalWorkExperienceInYears	51
TrainingTimesLastYear	188
ExperienceYearsAtThisCompany	56
ExperienceYearsInCurrentRole	16
YearsSinceLastPromotion	88
YearsWithCurrentManager	11
PerformanceRating	326

Source: Authors' data own processing (2025)

As such, the authors tackle this situation by the following codes:



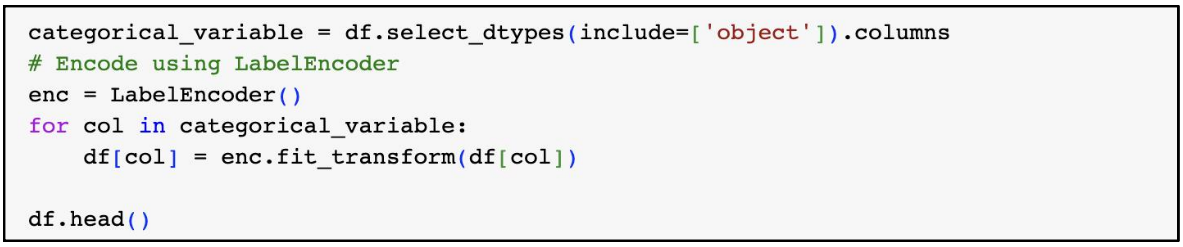
```python
# List of numerical columns with outliers
# Exclude PerformanceRating (target), ExperienceYearsAtThisCompany & YearsSinceLastPromotion (skew data, will be applied log transformation later))
outlier_cols = [
    "NumCompaniesWorked", "TotalWorkExperienceInYears", "TrainingTimesLastYear",
    "ExperienceYearsInCurrentRole", "YearsWithCurrManager"]

# Handle outliers using median replacement
for col in outlier_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    median_value = df[col].median()
    # Replace values above upper bound and below lower bound with median
    df.loc[df[col] > upper_bound, col] = median_value
    df.loc[df[col] < lower_bound, col] = median_value
```

A list of numerical columns with outliers is defined as "outlier_cols". They are replaced with median values, which are calculated by the code "median_value =

df[col].median()". This method is proved not to be sensitive to outliers since it does not remove any, thereby preventing data loss and preserving the overall distribution (Kwak and Kim, 2017). Meanwhile, the authors exclude 'ExperienceYearsAtThisCompany', 'YearsSinceLastPromotion' (skewed data) to process under log transformation later. Notably, 'PerformanceRating', the target of the research, is also not processed in this stage. Unlike features, where extreme values can mislead a model, outliers in the target often indicate genuine performance differences that should be considered. Hence, handling the outliers may cause losing important information about employees who perform exceptionally well or poorly.

3.3. Categorical Data Conversion

When encountering categorical variables which represent qualitative meanings, it is crucial to convert them into numerical format for machine learning models. Proper encoding can prevent model bias and enhance predictability.



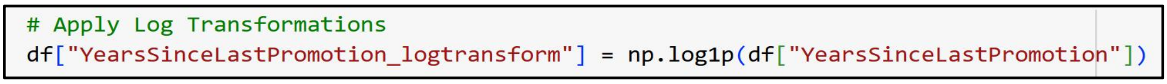
```
categorical_variable = df.select_dtypes(include=['object']).columns
# Encode using LabelEncoder
enc = LabelEncoder()
for col in categorical_variable:
    df[col] = enc.fit_transform(df[col])

df.head()
```

The authors convert categorical values by Label Encoding. An instance of LabelEncoder() from sklearn.preprocessing is created. The code loops through each categorical column in categorical_variables and applies Label Encoding using fit_transform(). Label Encoding assigns numeric values (0, 1, 2, …) to each unique category in a column.

3.4. Feature Transformation

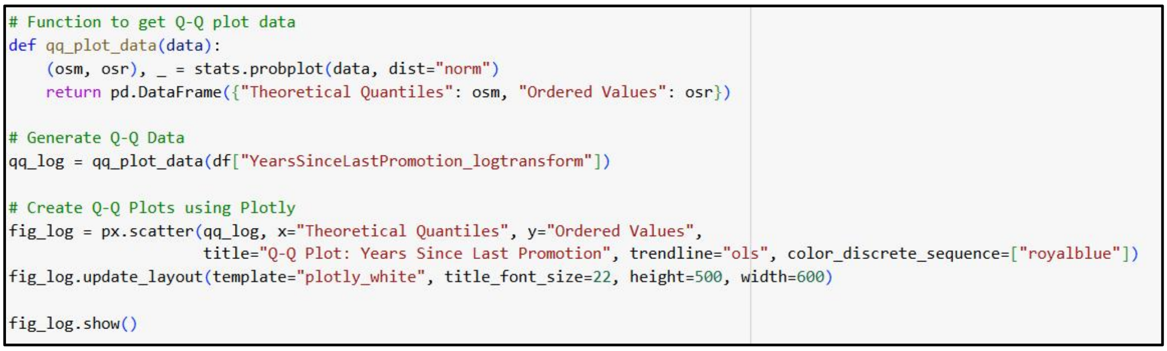
Feature transformation is a vital step to adjust distributions and reduce the skewness of data. Its objective is the model performance optimization through numerical stability between features.

Specifically, in this dataset, according to the EDA and skewness analysis in Chapter 2, the skewness of the two variables 'ExperienceYearsAtThisCompany' and 'YearsSinceLastPromotion' are 1.789055 and 1.974932, respectively. The data exhibiting highly positively skewed (>1.0) should undergo log transformation, rather than square root since square root is only applicable for moderate positive skewed data (0.5 to 1.0) (Higgins, White, & Anzures-Cabrera, 2008). Using the log transformation for YearsSinceLastPromotion, the codes are executed as below.



```
# Apply Log Transformations
df["YearsSinceLastPromotion_logtransform"] = np.log1p(df["YearsSinceLastPromotion"])
```

In the next step, the authors continue to use the Quartile - Quartile (Q-Q) plots to compare the distribution of the YearsSinceLastPromotion against a normal distribution to check normality.



```
# Function to get Q-Q plot data
def qq_plot_data(data):
    (osm, osr), _ = stats.probplot(data, dist="norm")
    return pd.DataFrame({"Theoretical Quantiles": osm, "Ordered Values": osr})

# Generate Q-Q Data
qq_log = qq_plot_data(df["YearsSinceLastPromotion_logtransform"])

# Create Q-Q Plots using Plotly
fig_log = px.scatter(qq_log, x="Theoretical Quantiles", y="Ordered Values",
                     title="Q-Q Plot: Years Since Last Promotion", trendline="ols", color_discrete_sequence=["royalblue"])
fig_log.update_layout(template="plotly_white", title_font_size=22, height=500, width=600)

fig_log.show()
```

The similar process is applied to 'ExperienceYearsAtThisCompany', and the results for the two variables are presented below (Figure 3.1; Figure 3.2).

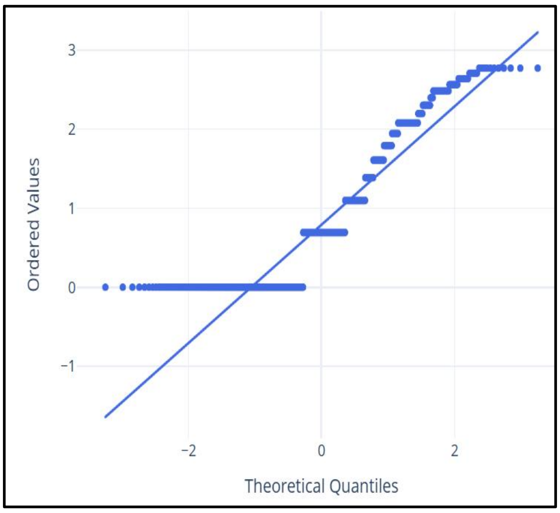


Figure 3.1: The Q-Q plot for YearsSinceLastPromotion

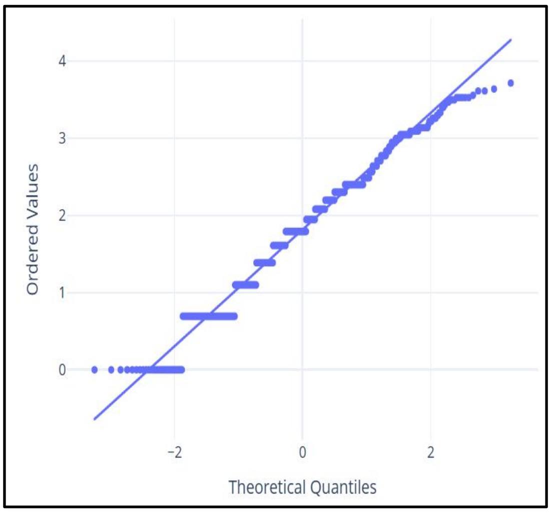


Figure 3.2: The Q-Q plot for ExperienceYearsAtThisCompany

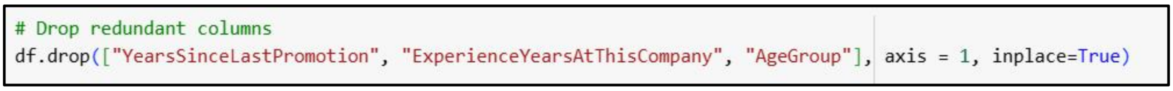
Source: Authors' data own processing (2025)

The closer the points are to the diagonal line, the more normally distributed the data is. As shown in the figures, most points in the middle of the plot nearly align with the diagonal line, indicating that that log transformation has successfully mitigated skewness and brought the data closer to normality in general.

3.5. Feature Selection

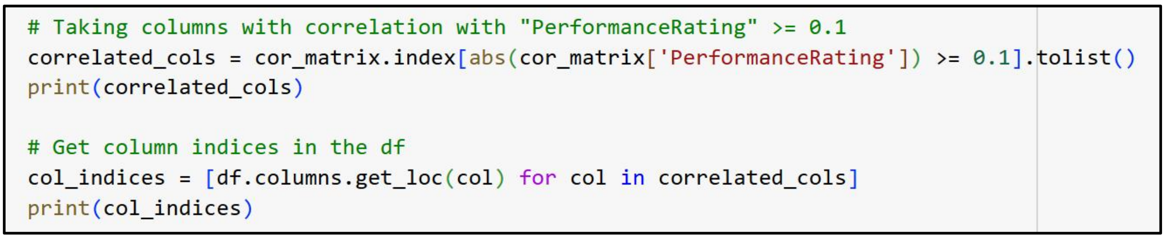
Feature selection is the final step in our data preprocessing, which determines valid data that contribute to the target prediction while removing redundant or irrelevant ones. This aims to improve model interpretability, reduce overfitting, and enhance computational efficiency by eliminating unnecessary features.

Here, by the following code, the authors first eliminate redundant columns 'YearsSinceLastPromotion', 'ExperienceYearsAtThisCompany' (since they have already been processed through log transformation), and 'AgeGroup' (only added by the authors for better illustration in the previous part).



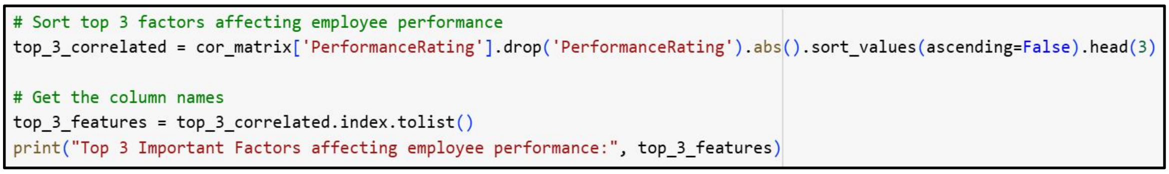
```
# Drop redundant columns
df.drop(["YearsSinceLastPromotion", "ExperienceYearsAtThisCompany", "AgeGroup"], axis = 1, inplace=True)
```

The below code identifies and selects features that have at least a 0.1 absolute correlation with 'PerformanceRating' from a correlation matrix (cor_matrix), and these important features will be used in the models.



```
# Taking columns with correlation with "PerformanceRating" >= 0.1
correlated_cols = cor_matrix.index[abs(cor_matrix['PerformanceRating']) >= 0.1].tolist()
print(correlated_cols)

# Get column indices in the df
col_indices = [df.columns.get_loc(col) for col in correlated_cols]
print(col_indices)
```

Finally, the authors figure out the top three variables most significantly impact employee performance by analyzing their correlation with 'PerformanceRating'. Initially, it extracts correlation values from cor_matrix, removes 'PerformanceRating' itself to avoid self-correlation, and ranks features based on the absolute values of their correlations. The strongest correlations are sorted in descending order. The results indicate that 'EmpEnvironmentSatisfaction' (employee satisfaction with the work environment), 'EmpLastSalaryHikePercent' (percentage of the last salary increase), and 'YearsSinceLastPromotion_logtransform' (log transformed years since the last promotion) are the most influential factors affecting employee performance.



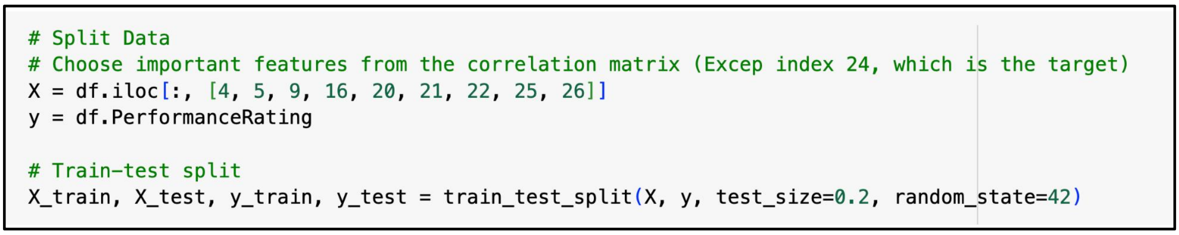
```
# Sort top 3 factors affecting employee performance
top_3_correlated = cor_matrix['PerformanceRating'].drop('PerformanceRating').abs().sort_values(ascending=False).head(3)

# Get the column names
top_3_features = top_3_correlated.index.tolist()
print("Top 3 Important Factors affecting employee performance:", top_3_features)
```

CHAPTER 4: MACHINE LEARNING MODEL CREATION AND EVALUATION

4.1. Data Preparation

The dataset undergoes a well data preparation process to enhance the predictive performance, focusing on feature selection and train-test splitting.

Feature selection is conducted using correlation analysis, where only features exhibiting a correlation coefficient of at least 0.1 with the target variable, 'Performance Rating', are retained. As a result, the selected indices include [4, 5, 9, 16, 20, 21, 22, 25, 26] (corresponding to variables: EmpDepartment, EmpJobRole, EmpEnvironmentSatisfaction, EmpLastSalaryHikePercent, EmpWorkLifeBalance, ExperienceYearsInCurrentRole, YearsWithCurrManager, and two log-transformed versions of YearsSinceLastPromotion and ExperienceYearsAtThisCompany).



```
# Split Data
# Choose important features from the correlation matrix (Excep index 24, which is the target)
X = df.iloc[:, [4, 5, 9, 16, 20, 21, 22, 25, 26]]
y = df.PerformanceRating

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Following feature selection, the dataset is split into training and testing sets using an 80-20 ratio. This reinforces that the training set would be large enough to capture meaningful patterns while maintaining a separate test set for unbiased model evaluation. The independent variables (X) include all selected features except PerformanceRating, which serves as the dependent variable (y). The structured approach guarantees that the dataset is appropriately prepared to facilitate its integration into subsequent machine learning model stages.

4.2. Model Selection

Four models have been considered for this study: Support Vector Machine (SVM), Random Forest, Balanced Random Forest (BRF), and Artificial Neural Network (MLP Classifier). Each model offers unique advantages suited to the complexity and characteristics of Employee performance data.

4.2.1. Support Vector Machine (SVM)

SVM was developed in the 1990s by Vladimir N. Vapnik and his colleagues, specifically used to classify data by finding an optimal line or creating multiple hyperplanes that maximize the distance between each class in an N-dimensional space (IBM, 2023). The model is a powerful algorithm applicable for pattern classification (both linear and nonlinear), regression problems, and outlier detection.

SVMs derive a class decision by identifying the closest points in the training data – called support vectors, then positioning the boundary so that the distance from these points is maximized. Moreover, by minimizing structural risk (which considers both training error and the model's complexity) rather than empirical risk, SVM could efficiently avoid a potential misclassification of testing data (Hong et al., 2005). In other words, the larger the margin, the lower the generalization error of the classifier (Punnoose and Ajit, 2016).

One of the foremost strengths of SVM is its ability to handle high-dimensional datasets efficiently that fit directly with the employee performance dataset, which often comprises numerous features – from quantitative metrics such as total work experience to qualitative variables like education background (Digital Defynd, 2024).

Moreover, employee performance data may be naturally non-linear, with complex interdependencies among various performance indicators. In such cases, SVM can also use a technique called the Kernel trick. Whereas, to map the input data into a higher-dimensional space where the data becomes linearly separable. Thus, allowing SVM to capture non-linear patterns without incurring computational costs, making them highly adaptable to diverse performance prediction scenarios (GeeksforGeeks Organization, 2023).

Some relevance of this model to predict Employee performance lies in similar reports, such as: "Enhancing Employee Performance Management" by Mourad et al. (2024), "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms" by Punnoose and Ajit (2016), or "Unbiased employee performance evaluation using machine learning" by Nayem and Uddin (2024).

4.2.2. Random Forest (RF)

Random Forest is a widely-used machine learning that combines the output of multiple decision trees, to reach a single result. Its ease of use has driven its adoption, as it effectively handles both classification and regression problems (Sruthi, 2024).

The mechanism is to construct a forest of trees based on bagging methods, where each tree is constructed independently from a randomly selected subset of the data (Punnoose and Ajit, 2016), then find the best node among random subsets generated and determine the most important features based on the impurity index calculated during the generation process (Papineni et al., 2021). Moreover, its ability to automatically balance data sets is particularly useful when dealing with imbalanced classes (CFI, 2024) like performance rating metrics.

Similar to SVM, Random Forest performs well with large, high-dimensional datasets, making it a strong choice for analyzing employee performance data (Alex, 2024). However, one key difference is that Random Forest counts every row during cross-validation, thus providing strong training and improving prediction. Yet, the model can suffer from "overfitting" problems, a common challenge in many of the ensemble bagging and boosting algorithms, as it always considers the majority-voted target values to perform the best split at each iteration (Papineni et al., 2021).

With regard to the analysis of employee performance, some relevant reports can be listed like "Unbiased employee performance evaluation using machine learning" by Nayem and Uddin (2024), or "Random Forest algorithm for HR data classification and performance analysis in cloud environments" by Dong (2024).

4.2.3. Balanced Random Forest (BRF)

Machine learning models often face challenges when dealing with imbalanced datasets, where the classifier favors the majority class while underperforming the minority class, leading to biased predictions.

BRF, a variation of the Random Forest, is specifically designed to address the class imbalance problem. Precisely, instead of using the entire imbalanced dataset, BRF performs undersampling of the majority class so that each tree gets a balanced dataset between the majority and minority classes (Fadly, 2024). That is also the key benefit of this model. However, as past research shows, making class priors equal either by down-sampling the majority class or over-sampling the minority class is usually more effective with respect to a given performance measurement, and that downsampling seems to have an edge over over-sampling. However, down-sampling the majority class may result in loss of information, as a large part of the majority class is not used (Chen, Liaw, and Brieman, 2004).

In contrast, BRF has better accuracy for the minority class, by balancing the data, it can pay more attention to them, making predictions more accurate. Besides, BRF also reduced bias compared to the standard models, given that the majority class can dominate, by giving equal weight to both classes (Fadly, 2024).

4.2.4. Artificial Neural Network - ANN (MLP Classifier)

ANN, particularly the Multi-Layer Perceptron (MLP), is a popular choice for predicting employee performance due to its proven ability to capture complex, non-linear relationships inherent in human behavior and organizational dynamics.

MLP, a type of ANN that consists of a class of feedforward artificial neural networks with an input layer, one or more hidden layers, and an output layer, with each layer characterized by multiple interconnected nodes, or neurons. These connections are weighted, and learning occurs by adjusting these weights, thereby allowing the model to learn complex patterns from input data (Moriarty and R. Miikkulainen, 1998).

A key benefit of MLP is the ability to perform hierarchical feature extraction. As data flows through the hidden layers, the network automatically transforms raw inputs into increasingly abstract representations, allowing the MLP to capture even nonlinear patterns and correlations among diverse employee data (Banerjee, 2024). Another significant advantage of MLP is its adaptability through the process of backpropagation (shortened for "backward propagation of errors"). Backpropagation computes gradients of a loss function with respect to the model's parameters and updates the parameters iteratively to minimize the loss. As a result, the model not only enhances performance on both training and testing data but also prevents "overfitting" problems, ensuring that the model remains robust even when the employee training data is noisy or limited (Jaiswal, 2024).

Some studies that support the widespread use of MLP in analyzing Employee performance can be named: "Identifying employee engagement drivers using multilayer perceptron classifier and sensitivity analysis" by Nunez-Sánchez et al., (2024), or "Evaluation of Factors Affecting Employees' Performance Using Artificial Neural Networks Algorithm: The Case Study of Fajr Jam" by Rahmanidoust and Zheng (2019).

4.3. Model Implementation

4.3.1. Class Imbalance Handling

Table 4.1: Overall Performance Ratinog Distribution

Performance Rating	Count	Percentage
3	874	72.8%
2	194	16.2%
4	132	11%

Source: Authors' data own processing (2025)

The presence of class imbalance in the Performance Rating dataset is first identified through the value counts and performance rating distribution ratio (Table 4.1). The analysis reveals that 72.8% of employees receive a Rating 3, while 16.2% receive Rating 2, and only 11% receive Rating 4. This imbalance poses a challenge for model training, as machine learning models tend to favor majority classes, leading to biased predictions and a potential diminishment in the accuracy of underrepresented categories.

To mitigate this issue, a structured approach is implemented to ensure fair representation of all classes in the dataset. The class distribution is first analyzed to quantify the extent of imbalance, followed by the application of appropriate resampling techniques to enhance predictive performance across all rating levels.

Different strategies are employed depending on the selected model. For the SVM and RF models, the class_weight = "balanced" parameter is applied to automatically adjust the weights of each class based on their inverse frequency. This method ensures that the model gives appropriate importance to minority classes without excessively favoring the majority class.

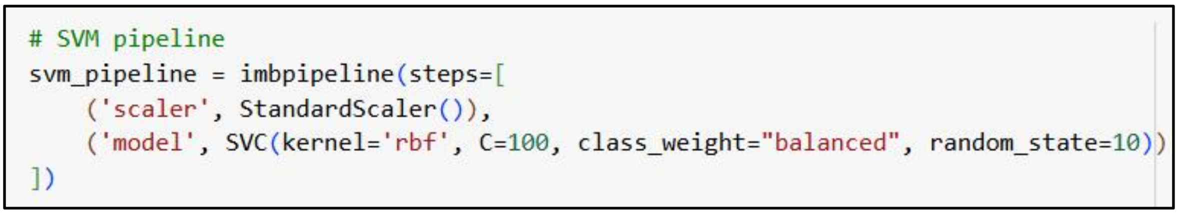
For the BRF model, the sampling_strategy = “not majority” is applied to handle the nature of imbalanced employee performance data more effectively. This approach resamples all minority classes while keeping the majority class unchanged, preventing excessive duplication of data. By curbing the influence of the most frequent category, the model gains a more balanced perspective, reducing the risk of overfitting while still capturing important patterns in underrepresented groups.

However, for the MLP Classifier, "class_weight" is not inherently available. Instead, Synthetic Minority Over-sampling Technique (SMOTE) is employed to handle class imbalance by artificially generating synthetic instances for the minority classes. Unlike random oversampling, which simply duplicates existing samples and increases the risk of overfitting, SMOTE effectively enhances minority class representation by interpolating new data points between existing observations. This method preserves the overall data distribution while improving the representation of underrepresented categories, enabling the model to learn more generalized patterns rather than memorizing repeated instances (Husain et al., 2025). Given that the dataset contained both numerical and categorical features, SMOTE appears to be particularly advantageous as it retains the relationships between features while expanding the training dataset.

By minimizing bias to ensure that the test set remained representative of real-world scenarios, the model's ability to generalize and deliver fair and reliable predictions across diverse employee categories has been significantly improved. Generally, this approach not only enhances the predictive performance across all rating levels but also contributes to a more equitable evaluation of employees, reducing the risk of misclassification due to class imbalances.

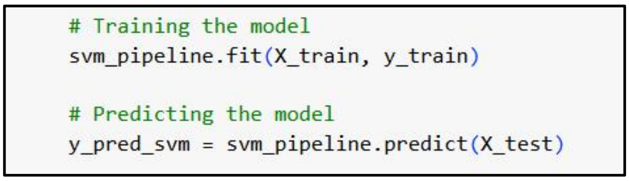
4.3.2. Model Implementation

4.3.2.1. Support Vector Machine (SVM)



```
# SVM pipeline
svm_pipeline = imbpipeline(steps=[
    ('scaler', StandardScaler()),
    ('model', SVC(kernel='rbf', C=100, class_weight="balanced", random_state=10))
])
```

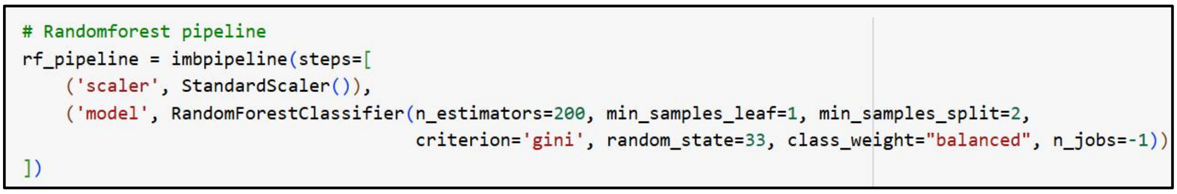
The SVM model constructed a Pipeline, which allows sequential transformations of the data before feeding it into the final estimator. First, StandardScaler is utilized to subtract the mean and divide it by the standard deviation for each feature, which means scaling all features to have approximately zero mean and unit variance, the model can converge more quickly and avoid bias toward features with larger numeric ranges. Moreover, the core classifier, SVC is applied. Within the SVC, several key parameters are presented. The RBF (Radial Basis Function) kernel projects data into a higher-dimensional space, allowing it to capture complex, non-linear boundaries between classes. The C parameter serves as a regularization term, where a higher value forces the model to classify every training point correctly, possibly at the cost of a simpler decision boundary. The setting of class_weight = 'balanced' is used to deal with imbalance as mentioned above. Finally, fixing the random seed with 'random_state = 10' guarantees the reproducibility of the model.



```
# Training the model
svm_pipeline.fit(X_train, y_train)

# Predicting the model
y_pred_svm = svm_pipeline.predict(X_test)
```

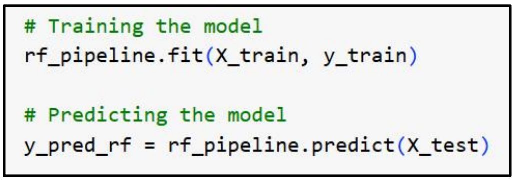
Once the pipeline is set up, the training process is executed by calling the “fit” method with the training data and labels. This method first fits the scaler on the training data, then transforms the data accordingly, and finally trains the normalized data with the RBF kernel, which allows it to map data into a higher-dimensional space. For prediction, the pipeline’s prediction method applies the same scaling transformation, using the parameters learned from the training set, to test the data before feeding it into the trained SVC to generate class predictions.

4.3.2.2. Random Forest (RF)



```
# Randomforest pipeline
rf_pipeline = imbpipeline(steps=[
    ('scaler', StandardScaler()),
    ('model', RandomForestClassifier(n_estimators=200, min_samples_leaf=1, min_samples_split=2,
                                     criterion='gini', random_state=33, class_weight="balanced", n_jobs=-1))
])
```

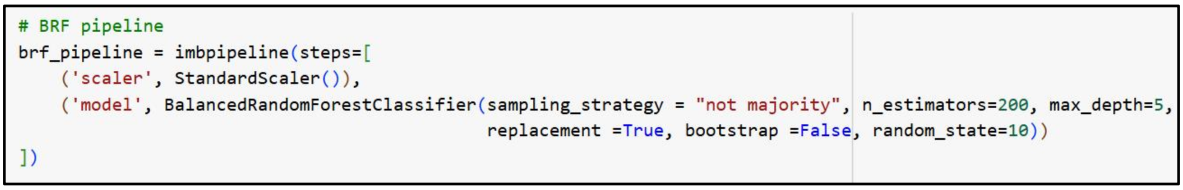
Similar to SMV, the Random Forest built a Pipeline, with 2 key steps. First, StandardScaler standardizes the numerical features, ensuring they have zero mean and unit variance. Second, model training Random Forest Classifier, an ensemble learning model that builds multiple decision trees and aggregates their predictions to enhance accuracy and avoid overfitting. Random Forest Classifier with parameters, including 'n_estimators = 200', meaning the model consists of 200 decision trees built independently on different samples of the training data. The 'min_samples_leaf = 1' and 'min_samples_split = 2' parameters, which control the minimum number of samples required in leaf nodes and for a split, respectively. In other words, a split continues until the stopping criteria are met. The Gini impurity criterion, to determine the quality of splits. A 'random_state = 33' is set for reproducibility, and class_weight = “balanced” helping with imbalanced datasets. The 'n_jobs = -1' setting ensures that all available CPU cores are utilized for faster computation.



```
# Training the model
rf_pipeline.fit(X_train, y_train)

# Predicting the model
y_pred_rf = rf_pipeline.predict(X_test)
```

The training process begins with the standardization of the training data by StandardScaler before passing the data to the Random Forest Classifier, which then constructs an ensemble of decision trees using the training labels y_train. By aggregating predictions from multiple trees, the Random Forest classifier reduces overfitting and improves generalization. Once training is complete, predictions are made following the same steps as training with the final prediction for each instance determined by majority voting. In other words, the class that receives the most votes across all trees is assigned as the ultimate prediction.

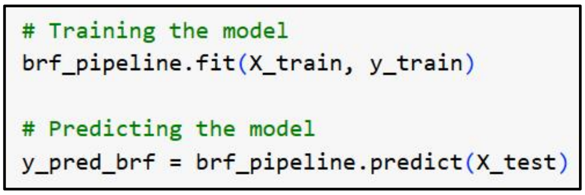
4.3.2.3. Balanced Random Forest (BRF)



```
# BRF pipeline
brf_pipeline = imbpipeline(steps=[
    ('scaler', StandardScaler()),
    ('model', BalancedRandomForestClassifier(sampling_strategy = "not majority", n_estimators=200, max_depth=5,
                                              replacement =True, bootstrap =False, random_state=10))
])
```

Like the above models, BFR pipeline also consists of 2 main parts: scaler (StandardScaler) to standardize features, and model training (BRF Classifier). In which, BRF Classifier, a variant of the traditional Random Forest Classifier, is designed to handle class imbalance effectively by undersampling the majority class before training each tree in the ensemble.

The BRF Classifier in this report is configured with various main parameters that enhance its ability to handle imbalanced datasets. The parameter sampling_strategy = ‘not majority’, is to handle class imbalance. The 'n_estimators = 200', indicating that 200 decision trees are built along with a controlled 'max_depth = 5', restricting the complexity of each tree to prevent overfitting problems. The 'replacement = True' parameter allows for sampling with replacement, meaning that individual data points may appear in multiple trees, which in turn enhances diversity. In contrast to traditional Random Forest implementations, the 'bootstrap = False' parameter is applied, disabling the typical bootstrapping method. This modification ensures that each tree is trained on a distinct random subset of the data, further emphasizing balanced representation across classes. Finally, the 'random_state = 10' ensures that results are reproducible.

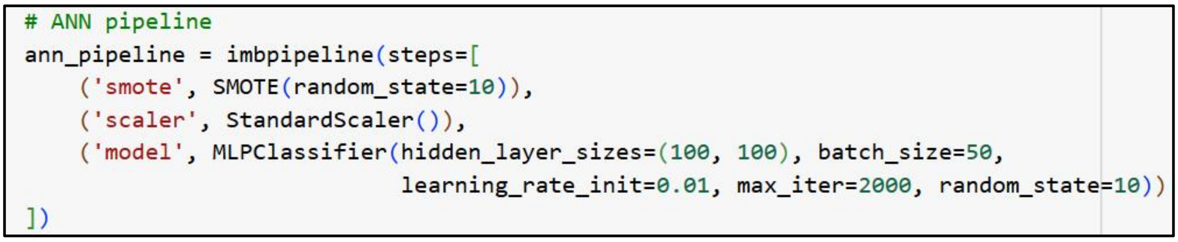


```
# Training the model
brf_pipeline.fit(X_train, y_train)

# Predicting the model
y_pred_brf = brf_pipeline.predict(X_test)
```

After defining the pipeline, the model is trained by passing the input data through StandardScaler. This standardized data is then fed into the BRF Classifier, where an ensemble of decision trees is built based on balanced subsets of the data.

Once training is complete, the model moves into the prediction phase, which undergoes the same processing steps as the training. Yet, unlike the traditional Random Forest Classifier, each tree in the ensemble contributes to the final prediction, and the majority voting mechanism of the Random Forest plus these individual decisions will form a single predicted class. As a result, the method not only reduces the variance but also enhances the overall stability and reliability of the predictions.

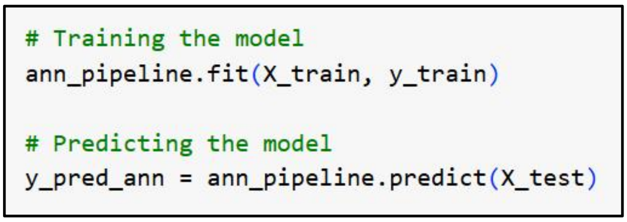
4.3.2.4. Artificial Neural Network - ANN (MLP Classifier)



```
# ANN pipeline
ann_pipeline = imbpipeline(steps=[
    ('smote', SMOTE(random_state=10)),
    ('scaler', StandardScaler()),
    ('model', MLPClassifier(hidden_layer_sizes=(100, 100), batch_size=50,
                            learning_rate_init=0.01, max_iter=2000, random_state=10))
])
```

ANN pipeline consolidates multiple steps, namely SMOTE, scaling, and an MLP Classifier, into a single workflow. The first component of the pipeline is SMOTE as to handling class imbalance. Similar to above models, a StandardScaler is applied, ensuring that all features have a zero mean and a one standard deviation.

The final step in the pipeline is the MLP Classifier, a Multi-Layer Perceptron neural network. The network uses a backpropagation algorithm to adjust the weights of each neuron in its two hidden layers, with 100 neurons per layer. Other parameters such as 'batch_size = 50' and 'learning_rate_init = 0.01' determine how the network updates its weights and processes training examples, respectively. A larger 'max_iter = 2000' is specified to allow the model sufficient epoch to converge on a solution. Besides, setting a 'random_state = 10' ensures that weight initialization and other random factors remain consistent, thus facilitating the reproducibility.



```
# Training the model
ann_pipeline.fit(X_train, y_train)

# Predicting the model
y_pred_ann = ann_pipeline.predict(X_test)
```

The training process begins with a pipeline systematically applying each transformation step in the specified order - SMOTE, scaling, and then fitting to MLPClassifier - using only the training data. First, SMOTE identifies minority classes and synthesizes new examples for those classes. Once the data is rebalanced, StandardScaler standardizes each feature and passes the data to the MLP Classifier.

Once the model has been trained, predictions are generated, go through the same workflow as the training process, however, SMOTE is only fitted on the training data and not applied to the test data in practice, meaning no new synthetic examples are created during this phase. In the end, the processed test data produces class predictions based on the learned weights and biases.

4.4. Model Evaluation

4.4.1. Support Vector Machine (SVM)

Table 4.2: SVM Classification Report

	precision	**recall**	**f1-score**	**support**
2	0.76	0.86	0.81	29
3	0.93	0.90	0.91	184
4	0.63	0.70	0.67	27
SVM Accuracy	**0.87 (0.8708)**			240
Macro avg	0.77	0.82	0.80	240
Weighted avg	0.88	0.87	0.87	240

Source: Authors' data own processing (2025)

The model achieves an accuracy of 87%, indicating that it correctly classifies the majority of samples in the test set. First of all, Class 2 achieves a precision of 0.76 and a recall of 0.86, indicating some difficulties in capturing this class correctly. In contrast, Class 3 demonstrates particularly strong performance, reflected in a precision of 0.93 and a recall of 0.90, suggesting that the model rarely confuses this class with others, given the fact that Class 3 is supported by a substantial portion of the dataset, which provides SVM with more training data to learn from and improve its predictions. Lastly, Class 4 appears to overlap with Class 3, resulting in a worse performance within the SVM model with a 0.70 recall and a 0.63 precision.

Observing the SVM Confusion Matrix, the majority of Class 2 are correctly labeled (25), but a few (4) are misclassified as Class 3. In contrast, Class 3, having the largest number of samples (165), maintains a high recall; however, some are mislabeled as either Class 2 (8) or Class 4 (11). Last, Class 4 shows no disorder with Class 2 (0) but does exhibit some overlap with Class 3 (8).

These findings confirm that the model mainly struggles to differentiate between certain instances of Class 2 and Class 3, and to a lesser extent between Class 3 and Class 4, while there is no confusion between Class 2 and Class 4.

4.4.2. Random Forest (RF)

Random Forest Classifier achieved an overall accuracy of 95.42%, indicating that the model is highly effective at classifying employee performance datasets.

Table 4.3: Random Forest Classification Report

	precision	recall	f1-score	support
2	0.93	0.93	0.93	29
3	0.95	0.99	0.97	184
4	1.00	0.74	0.85	27
RF Accuracy	**0.95 (0.9542)**			240
Macro avg	0.96	0.89	0.92	240
Weighted avg	0.96	0.95	0.95	240

Source: Authors' data own processing (2025)

For each Class performance, Class 2, had a balanced precision and recall of 0.93, meaning it was correctly identified when present and not frequently misclassified as another class. Class 3 (with the highest number of samples, 184) had the best recall score at 0.99, meaning nearly all instances of Class 3 were correctly identified. However, Class 4, had the highest precision (1.00) but a lower recall (0.74) meaning although the model predicted Class 4 correctly every time, it failed to capture all actual instances of this class, thus leading to misclassification.

From observation of the above Confusion Matrix, the majority of misclassifications occurred in Class 4, where 7 instances were misclassified as Class 3, likely due to feature similarities between the two classes. Additionally, Class 2 had 2 samples misclassified as Class 3, but no misclassification into Class 4.

4.4.3. Balanced Random Forest (BRF)

The BRF model demonstrates a strong overall performance, as reflected by its accuracy of 95.42%, the same result as the traditional Random Forest learning. By observing, Class 2 achieves a precision of 0.96 and a recall of 0.90, indicates the model corrects 96% of the time prediction and identifies 90% of the actual instances of Class 2 in the dataset. The performance in Class 3 is noteworthy, with a precision of 0.95 and an impressive recall of 0.99, demonstrating an effective capture of nearly all instances, given that Class 3 is the majority class with 184 samples. Class 4 yields a perfect precision of 1.00 implying that whenever the model predicts Class 4, it results in no falsehood. Yet, the recall of 0.74 shows that about ¼ of actual samples are missed.

Table 4.4: Balanced Random Forest Classification Report

	precision	**recall**	**f1-score**	**support**
2	0.96	0.90	0.93	29
3	0.95	0.99	0.97	184
4	1.00	0.74	0.85	27
BRF Accuracy	**0.95 (0.9542)**			240
Macro avg	0.97	0.88	0.92	240
Weighted avg	0.96	0.95	0.95	240

Source: Authors' data own processing (2025)

Regarding the Confusion Matrix, Class 2, 26 instances were correctly classified, with only 3 instances misclassified as Class 3, demonstrating strong performance. Class 3 stands out with 183 instances correctly predicted and only 1 misclassified as Class 2, indicating exceptional accuracy for this class. In contrast, Class 4 shows a degree of difficulty for the model, with 20 instances correctly identified and 7 instances misclassified as Class 3. These misclassifications highlight that while the model performs well overall, it tends to confuse Class 4 with Class 3, suggesting areas where further tuning might be needed.

4.4.4. Artificial Neural Network - ANN (MLP Classifier)

Table 4.5: MLP Classification Report

	precision	**recall**	**f1-score**	**support**
2	0.85	0.79	0.82	29
3	0.94	0.94	0.94	184
4	0.66	0.70	0.68	27

MLP Accuracy	**0.90 (0.8958)**			240
Macro avg	0.82	0.81	0.81	240
Weighted avg	0.90	0.90	0.90	240

Source: Authors' data own processing (2025)

Overall, the model achieves an accuracy of roughly 0.90 (or 89.58%), suggesting that it makes correct predictions on the majority of the test samples. In detail, Class 3, has the highest support (184 samples), achieves the strongest performance, yielding a precision and recall of 0.94. Classes 2 and 4, by contrast, have far fewer samples (29 and 27, respectively). While Class 2 still demonstrates relatively strong metrics, 0.85 precision and 0.79 recall, Class 4 shows noticeably weaker performance, with a precision of 0.66, and a recall of 0.70. This pattern may indicate some confusion between Classes 3 and 4, as the model appears to struggle in distinctly identifying Class 4 samples.

Focusing on Class 2 for the ANN model, the result shows that 23 instances are correctly identified, while 3 are incorrectly classified as Class 3 and another 3 as Class 4. Such confusion can arise from overlapping feature distributions or insufficient representation in the training process. In contrast, Class 3 correctly classifies 173 instances. Still, there are 4 instances incorrectly labeled as Class 2 and 7 incorrectly labeled as Class 4. Lastly, Class 4 with 19 instances accurately classified, yet, 8 instances were misclassified as Class 3. This pattern is often observed in situations where Class 4 is less frequent in the dataset, limiting the model to distinct patterns and leading to confusion with the more common Class 3.

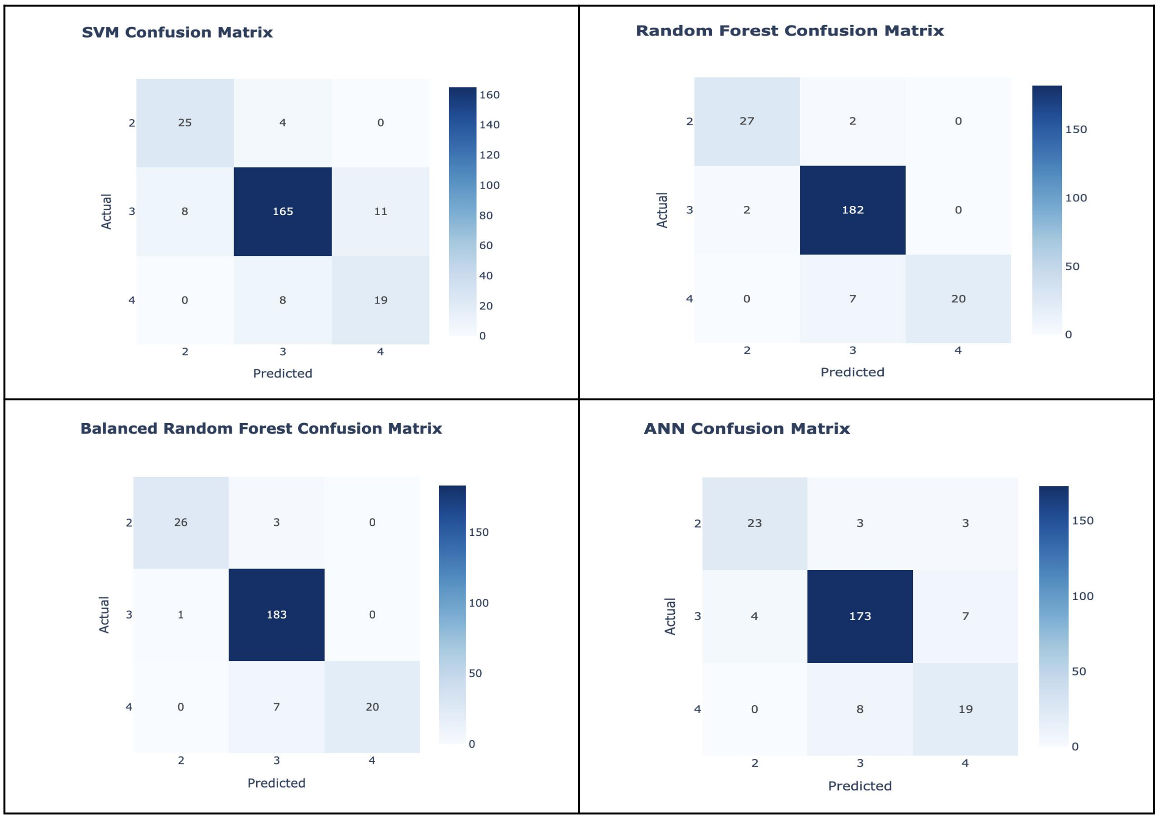


Figure 4.1: Summary of the 4 Models' Confusion Matrix

Source: Authors' data own processing (2025)

4.5. Model Comparison

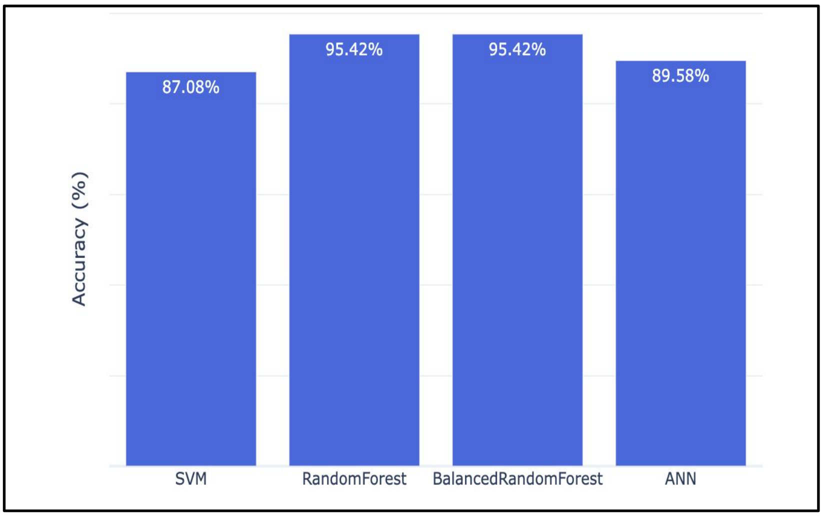


Figure 4.2: Accuracy Comparison of 4 Models

Source: Authors' data own processing (2025)

In evaluating the employee performance, both Random Forest and BRF emerged as the top-performing algorithms with an accuracy of 95.42%. These models outperformed MLP/ ANN (89.58% accuracy), and the SVM (87.08%). From a high-level perspective, the Random Forest family's success stems from its ensemble approach of combining multiple decision trees, leading to strong prediction and preventing overfitting. This makes it particularly suitable for complex datasets where employee performance is influenced by both categorical and numerical variables.

Delving into the classification reports, Random Forest and BRF not only produce high overall accuracy but also exceed in class-level metrics with strong precision and recall scores across the different classes (Class 2, 3, and 4). In a business context, it implies that the model is reliable for identifying both high and low performers. The Balanced Random Forest variant, in particular, is designed to address potential class imbalance by down-sampling the majority class. In an employee performance scenario, where “excellent” or “low” performers might be underrepresented, BRF ensures that these classes receive due attention, leading to more equitable predictions.

From a human resources perspective, these findings underscore the importance of using ensemble methods to gain nuanced insights into workforce productivity. The high recall for certain classes (notably Class 3 in both Random Forest models) means that the model captures a large proportion of employees truly belonging to that performance level. High precision, on the other hand, ensures that when the model predicts a certain performance level, it is correct most of the time. This balance of precision and recall translates into more confident decision-making, whether the goal is to identify high-potential employees for leadership programs or to detect underperformers who may benefit from targeted interventions and support.

CHAPTER 5: KEYS FINDINGS AND RECOMMENDATIONS

5.1. Key Findings

The above evaluation of employee performance data provided key insights into the factors impacting productivity and engagement, with the top 3 variables with the highest correlation to performance are EmpEnvironmentSatisfaction, EmpLastSalaryHikePercent, and YearsSinceLastPromotion. These variables will be reviewed to summarize their significance on the overall employee performance.

5.1.1. Employee Environment Satisfaction

Employee environment satisfaction has the strongest correlation with performance ratings (correlation coefficient equals 0.4), emphasizing the importance of a positive work environment in driving employee engagement and overall productivity. This result is reasonable as a conducive work environment provides employees with a sense of security, thus enabling them to perform optimally (Badrianto et al., 2020). Moreover, the working environment can affect employee's emotions. Therefore, when the employees are satisfied with their working environment, they tend to feel comfortable, which allows them to perform their tasks efficiently. As a result, they make optimal use of their working hours, leading to increased productivity and enhanced performance (Ramli et al., 2019).

5.1.2. Last Salary Hike Percentage

Employees who have recently received salary increases also demonstrate significant improvements in their work performance levels (correlation coefficient equals 0.33). This result suggests that financial incentives act as a powerful working motivator, emphasizing the importance of maintaining a competitive compensation structure to enhance employee motivation and retention. Employees who receive a sufficient salary to cover their daily expenses are more likely to work efficiently, particularly when their contributions are acknowledged through bonuses, incentives and salary increases (AI Mehrzi et al., 2016).

5.1.3. Years Since Last Promotion

The third important feature is Years Since Last Promotion, in which the result found that employees who receive recent promotions are more likely to demonstrate higher working performance, with a correlation coefficient equal to -0.24 (after categorical encode and log transformation), showing that career promotion chances enhance employee motivation and engagement. Job promotions offer employees opportunities for personal development, greater responsibilities and power, and also enhanced social status. The desire for a promotion motivates high-performing employees to put in extra effort to reach their performance goals (Nguyen et al., 2015). When promotions are effectively implemented, they contribute to increased job satisfaction, thus improving employee performance (Robbins and Judge, 2011). Especially in high power distance cultures, people may value promotions more than bonuses, as promotions signify higher social status and success (Nguyen et al., 2015).

5.2. Strategies to Improve Employee Performance

To ensure a positive work environment, companies should conduct periodic employee satisfaction surveys and try to improve based on that feedback. For example, Microsoft has actively gathered feedback from employee engagement surveys and adjusted accordingly to improve employee satisfaction and the organization's overall health (Microsoft, 2018). Moreover, it is necessary to encourage open communication and foster a culture of inclusivity and diversity, which can help to build a more positive and collaborative atmosphere. Google, for instance, has successfully implemented open office designs, flexible working arrangements, and intentionally flat organizational structures to enhance creativity and job satisfaction, which also serve as an attractive feature for top talent (Abirami, n.d.). Finally, prioritizing ergonomic workspace design, cutting-edge technology, and comprehensive wellness programs also significantly contribute to productivity, and enhance employee satisfaction and overall well-being.

In terms of salary and bonuses, a well-defined performance-based salary increment framework should be established to maintain workforce motivation. Beyond salary increases and financial incentives, companies can also offer non-monetary benefits such as work-from-home days, professional development opportunities, employee recognition programs, or travel packages. It is also crucial for companies to regularly conduct industrial benchmarks to ensure a competitive and reasonable salary structure, adding to talent retention and motivation. By addressing both financial and non-financial needs, firms can create a more balanced and supportive working environment, which will ultimately drive higher productivity, innovation, and employee loyalty.

Regarding career development, transparent career progression pathways and promotion criteria should be implemented to provide employees with clear growth opportunities. Leadership development programs can play a crucial role in preparing employees for higher position responsibilities, thereby ensuring a steady flow of skilled employees ready to take on further career ladder. Moreover, companies should also conduct cross-functional mobility and promote job diversification through rotation programs to broaden staff's competencies. For instance, companies such as Unilever or Nestle have successfully implemented rotational leadership programs with clear career paths to equip employees with hands-on experience across various functions and accelerate their career progressions.

CONCLUSION

In conclusion, the authors have identified the most influential factors impacting employee performance, including employee environment satisfaction, last salary hike percentage, and years since last promotions. Several strategies to enhance workplace performance ratings and productivity have also been suggested. Specifically, by fostering a positive working environment, implementing performance-based salary systems, and providing transparent career development opportunities for employees, companies can create a more motivated, innovative and loyal workforce. While financial incentives remain important, non-monetary benefits, suitable organizational structure and job promotions also significantly enhance employee efficiency.

The constraint comes from the dataset itself, which contains only 1200 records. This sample size is relatively small, which may potentially fail to fully reflect the diversity and complexity of a broader workplace. Additionally, this dataset is missing some other important details that could also affect employee performance, such as the actual amount of salary (since in the dataset, it only includes percentage raises), working conditions like hours or workload, and some outside factors like company culture or organizational structure. The absent information limits the depth of the analysis and its ability to comprehensively explain performance variations among employees.

REFERENCE

Abirami, D. M. Open Office Design and its Impact on Employees: A Review of Research and Perspectives.

Alex, S. (2024, September 30). What are the Advantages and Disadvantages of Random Forest? Retrieved from Pickl.AI website: https://www.pickl.ai/blog/advantages-and-disadvantages-random-forest/

Al Mehrzi, N., & Singh, S. K. (2016). Competing through employee engagement: a proposed framework. *International Journal of Productivity and Performance Management, 65*(6), 831-843.

Badrianto, Y., & Ekhsan, M. (2020). Effect of work environment and job satisfaction on employee performance in pt. Nesinak industries. *Journal of Business, Management, & Accounting, 2*(1).

Banerjee, S. (2024, April 6). Exploring the Power and Limitations of Multi-Layer Perceptron (MLP) in Machine Learning. Retrieved from Medium website: https://shekhar-banerjee96.medium.com/exploring-the-power-and-limitations-of-multi-layer-perceptron-mlp-in-machine-learning-d97a3f84f9f4.

Baral, R. (2009). Examining antecedents of work-family enrichment and its effect on individual, family and organisational outcomes. *Unpublished Doctoral Dissertation, IIT Bombay*.

Chen, C., Liaw, A., & Brieman, L. (2004, January). Using Random Forest to Learn Imbalanced Data | Department of Statistics. Retrieved from statistics.berkeley.edu website: https://statistics.berkeley.edu/tech-reports/666

Dahkoul, Z. M. (2018). The determinants of employee performance in Jordanian organizations. *Pressacademia*, *5*(1), 11–17. https://doi.org/10.17261/pressacademia.2018.780.

Danesh, M., & Nourdad, N. (2017). On the Relationship between Creative Problem Solving Skill and EFL Reading Comprehension Ability. *Theory and Practice in Language Studies*, *7*(3), 234. https://doi.org/10.17507/tpls.0703.10

Digital Defynd. (2024, July 6). 10 Pros & Cons of Support Vector Machines [2024]. Retrieved from DigitalDefynd website: https://digitaldefynd.com/IQ/pros-cons-of-support-vector-machines/.

Dong, F. (2024). Random Forest Algorithm for HR Data Classification and Performance Analysis in Cloud Environments. *International Journal of Advanced Computer Science and Applications*, *15*(11). https://doi.org/10.14569/ijacsa.2024.0151147.

Dubin, R. (1956). Industrial Workers' Worlds: A Study of the "Central Life Interests" of Industrial Workers. *Social Problems*, *3*(3), 131–142. https://doi.org/10.2307/799133.

Fadly, M. T. (2024, July 3). BALANCED RANDOM FOREST - Fadly Mochammad Taufiq - Medium. Retrieved March 22, 2025, from Medium website: https://medium.com/@fadleemt/balanced-random-forest-d5dc9c896bb4.

GeeksforGeeks Organization. (2023, May 7). Support Vector Machine in Machine Learning. Retrieved from GeeksforGeeks website: https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/

Greenhalgh, L., & Rosenblatt, Z. (1984). Job insecurity: Toward conceptual clarity. *Academy of Management review*, *9*(3), 438-448.

Higgins, J. P. T., White, I. R., & Anzures-Cabrera, J. (2008). Meta-analysis of skewed data: Combining results reported on log-transformed or raw scales. *Statistics in Medicine*, *27*(29), 6072–6092. https://doi.org/10.1002/sim.3427.

Hngoi, C. L., Abdullah, N. A., Wan Sulaiman, W. S., & Zaiedy Nor, N. I. (2024). Examining job involvement and perceived organizational support toward organizational commitment: job insecurity as mediator. *Frontiers in Psychology, 15,* 1290122.

Hong, W.-C., Pai, P.-F., Huang, Y.-Y., & Yang, S.-L. (2005). Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance. *Lecture Notes in Computer Science*, *3610*(978-3-540-31853-8), 668–674. https://doi.org/10.1007/11539087_85.

IBM. (2023, December 12). Support Vector Machine. Retrieved from IBM website: https://www.ibm.com/think/topics/support-vector-machine

Jaiswal, S. (2024, February 7). Multilayer Perceptrons in Machine Learning: A Comprehensive Guide. Retrieved from Datacamp.com website:

https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning.

Janssen, P. P., Schaufelioe, W. B., & Houkes, I. (1999). Work-related and individual determinants of the three burnout dimensions. *Work & stress, 13*(1), 74-86.

Kanwar, Y. P. S., Singh, A. K., & Kodwani, A. D. (2009). Work—life balance and burnout as predictors of job satisfaction in the IT-ITES industry. *Vision*, *13*(2), 1-12.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407.

Lasa, A. N., Pedroni, A., & Komm, A. (2024, May 15). Performance management that puts people first. Retrieved from www.mckinsey.com website: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/in-the-spotlight-performance-management-that-puts-people-first

Lodahl, T. M., & Kejnar, M. (1965). The definition and measurement of job involvement. *Journal of applied psychology*, *49*(1), 24.

Mesiya, A. Y. (2019, June 30). Factors Affecting Employee Performance: An Investigation on the Private School Sector. Retrieved from papers.ssrn.com website: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645036.

Microsoft. (2018). Improving employee engagement in the workplace. Retrieved from Microsoft.com website: https://www.microsoft.com/en-us/microsoft-365/business-insights-ideas/resources/employee-engagement

Moriarty, D. E., & R. Miikkulainen. (1998). Hierarchical evolution of neural networks. *World Congress on Computational Intelligence*, *98TH8360*(pp. 428-433). https://doi.org/10.1109/icec.1998.699793.

Mourad, Z., Noura, A., Mohamed, C., & Abdelhamid, B. (2024). Enhancing Employee Performance Management. *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, *15*(3). https://doi.org/10.14569/ijacsa.2024.01503100.

Nayem, Z., & Uddin, A. (2024). Unbiased Employee Performance Evaluation Using Machine Learning. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(1), 100243. https://doi.org/10.1016/j.joitmc.2024.100243.

Nguyen, P. D., Dang, C. X., & Nguyen, L. D. (2015). Would better earning, work environment, and promotion opportunities increase employee performance? An investigation in state and other sectors in Vietnam. *Public Organization Review, 15,* 565-579.

Núñez-Sánchez, J. M., Jesús Molina-Gómez, Pere Mercadé-Melé, & Fernández-Miguélez, S. M. (2024). Identifying employee engagement drivers using multilayer perceptron classifier and sensitivity analysis. *Eurasian Economic Review*, *14*. https://doi.org/10.1007/s40821-024-00283-6.

Papineni, S. lakshmi, Reddy, M., Yarlagadda, S., Yarlagadda, S., & Akkineni, H. (2021). An Extensive Analytical Approach on Human Resources using Random Forest Algorithm. *International Journal of Engineering Trends and Technology*, *69*(5), 119–127. https://doi.org/10.14445/22315381/ijett-v69i5p217.

Perry-Smith, J. E., & Blum, T. C. (2000). Work-family human resource bundles and perceived organizational performance. *Academy of management Journal*, *43*(6), 1107-1117.

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9). https://doi.org/10.14569/ijarai.2016.050904.

Rahmanidoust, M., & Zheng, J. (2019). Evaluation of Factors Affecting Employees' Performance Using Artificial Neural Networks Algorithm: The Case Study of Fajr Jam. *International Business Research*, *12*(10), 86. https://doi.org/10.5539/ibr.v12n10p86.

Ramli, A. H. (2019). Work environment, job satisfaction and employee performance in health services. *Business and Entrepreneurial Review, 19*(1), 29-42.

Reginaldo G. de S. Neto, Jatobá, M. N., Santana, M., Fernandes, P. S., Ferreira, J. J., Foleis, J. H., & Teixeira, J. P. (2025). Human Resources Sptimization with

MultiLayer Perceptron: An Automated Selection Tool. *Procedia Computer Science*, *256*(1877-0509), 238–245. https://doi.org/10.1016/j.procs.2025.02.117.

Salessi, S. M., & Omar, A. G. (2019). Job involvement in current research: Update and state of the art.

Schermerhorn Jr, J. R., Osborn, R. N., Uhl-Bien, M., & Hunt, J. G. (2011). Organizational behavior. John Wiley & Sons.

Scholarios, D., & Marks, A. (2004). Work-life balance and the software worker. *Human Resource Management Journal*, *14*(2), 54-74.

Sheng, X., Wang, Y., Hong, W., Zhu, Z., & Zhang, X. (2019). The curvilinear relationship between daily time pressure and work engagement: The role of psychological capital and sleep. *International Journal of Stress Management*, *26*(1), 25–35. https://doi.org/10.1037/str0000085.

Singh, A., & Gupta, B. (2015). Job involvement, organizational commitment, professional commitment, and team commitment: A study of generational diversity. *Benchmarking: An International Journal*, *22*(6), 1192-1211.

Sulistiani, L., & Faozanudin, M. (2022). Effectiveness Analysis of the Employee Work Performance Assessment System – A Critical Three-Component Approach. *Expert Journal of Business and Management*, *10*(2). Retrieved from https://business.expertjournals.com/23446781-1005/.

Zhenjing, G., Chupradit, S., Ku, K. Y., Nassani, A. A., & Haffar, M. (2022). Impact of employees' Workplace Environment on employees' performance: a multi-mediation Model. *Frontiers in Public Health*, *10*(890400). NCBI. https://doi.org/10.3389/fpubh.2022.890400.

APPENDICES

PerformanceRating	2	3	4
Gender			
Female	75	349	51
Male	119	525	81
AgeGroup			
<25	12	54	11
25-34	69	336	47
35-44	66	300	48
45-54	36	151	18
55+	11	33	8

Appendix 1: Average Employee Performance Ratings by Gender and Age

Source: Authors' data own processing (2025)

Code: df.groupby(['Gender', 'PerformanceRating']).size().unstack().fillna(0)

df.groupby(['AgeGroup', 'PerformanceRating']).size().unstack().fillna(0)

	EmDepartment	PerformanceRating
0	Development	1114
1	Sales	1067
2	Research & Development	1002
3	Human Resources	158
4	Finance	136
5	Data Science	61

	EmDepartment	Gender	PerformanceRating
0	Data Science	Female	3.000000
1	Data Science	Male	3.083333
2	Development	Female	3.098592
3	Development	Male	3.077626
4	Finance	Female	2.681818
5	Finance	Male	2.851852
6	Human Resources	Female	3.058824
7	Human Resources	Male	2.864865
8	Research & Development	Female	2.945736
9	Research & Development	Male	2.906542
10	Sales	Female	2.840764
11	Sales	Male	2.875000

Appendix 2: Distribution of Employee Performance Ratings Across Departments, by Gender

Source: Authors' data own processing (2025)

Code: df.groupby('EmpDepartment')['PerformanceRating'].sum()

df.groupby(['EmpDepartment', 'Gender'])['PerformanceRating'].mean()

	YearsSinceLastPromotion	**PerformanceRating**
0	0	3.123667
1	1	2.898990
2	2	2.763780
3	3	2.777778
4	4	2.773585
5	5	2.828571
6	6	2.833333
7	7	2.854839
8	**8**	**2.545455**
9	9	2.687500
10	10	2.600000
11	11	2.826087
12	12	2.666667
13	**13**	**3.125000**
14	14	3.000000
15	15	2.909091

Appendix 3: Average Performance Rating by Years Since Last Promoted

Source: Authors' data own processing (2025)

Code:

df.groupby(['YearsSinceLastPromotion'])['PerformanceRating'].mean().reset_index()

	EmpLastSalaryHikePercent	Gender	PerformanceRating
0	11	Female	2.828571
1	11	Male	2.848485
2	12	Female	2.846154
3	12	Male	2.800000
4	13	Female	2.793651
5	13	Male	2.895238
6	14	Female	2.843750
7	14	Male	2.870370
8	15	Female	2.928571
9	15	Male	2.907407
10	16	Female	2.812500
11	16	Male	2.888889
12	17	Female	3.000000
13	17	Male	2.869565
14	18	Female	2.928571
15	18	Male	2.822222
16	19	Female	2.892857
17	19	Male	2.857143
18	20	Female	3.440000
19	20	Male	3.280000
20	21	Female	3.454545
21	21	Male	3.652174
22	22	Female	3.444444
23	22	Male	3.413793
24	23	Female	3.363636
25	23	Male	3.700000
26	**24**	**Female**	**4.000000**
27	24	Male	3.400000
28	25	Female	3.375000
29	25	Male	3.200000

Appendix 4: Average Performance Rating by Salary Hike and Gender

Source: Authors' data own processing (2025)

Code: df.groupby(['EmpLastSalaryHikePercent','Gender'])['PerformanceRating'].mean().reset_index()

	EmpDepartment	OverTime	PerformanceRating
0	Data Science	No	3.000000
1	Data Science	Yes	3.333333
2	Development	No	3.107884
3	Development	Yes	3.041667
4	Finance	No	2.810811
5	Finance	Yes	2.666667
6	Human Resources	No	2.850000
7	Human Resources	Yes	3.142857
8	Research & Development	No	2.894737
9	Research & Development	Yes	2.989583
10	Sales	No	2.830189
11	Sales	Yes	2.935185

Appendix 5: Average Performance Rating Across All Departments, by Overtime

Source: Authors' data own processing (2025)

Code:

```
df.groupby(['EmpDepartment','OverTime'])['PerformanceRating'].mean().reset_index()
```