2022-3_NSYR APPLIED ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS CAPSTONE
(MAL7504-E)

TOPIC: PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING

PRESENTED BY: CHUKWU GODSON UDE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MSC IN APPLIED ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS

WORD COUNT: 11,321

DATE: MARCH 2024

STATEMENT OF AUTHENTICITY

I have read the university regulations relating to plagiarism. In submitting this, I certify that this dissertation is all my own work and does not contain any unacknowledged work from any other sources.

Signature: 

Date: March 2024

ACKNOWLEDGEMENT

I want to especially acknowledge my supervisor, Dr. Raja Sreedharan, and all my lecturers at the University of Bradford for their immeasurable guidance.

DEDICATION

I dedicate this work to my caring family for their sacrifice, my supportive friends, classmates, and former colleagues.

ABSTRACT

The primary objective of this research is to predict employee attrition using machine learning techniques. In achieving this goal, the study addresses the following research questions: 1) To what extent can machine learning predict employee attrition? 2) Which machine learning model produces the most accurate predictions? 3) Which features or variables are most important in predicting employee attrition?

Employing the CRISP-DM research design, crucial findings have emerged. The Extra Tree classifier outperformed other models, achieving an impressive accuracy and predictability score of 95.14%. Furthermore, the study highlights the significance of applying appropriate sampling techniques to enhance model performance. Notably, Overtime, job level, marital status, years in current role, and total working years emerge as the top five predictors of attrition.

These findings hold significant implications for workforce management stakeholders, enabling them to implement proactive employee retention programmes well in advance before attrition occurs. Additionally, the research contributes to the scholarly community by refining existing literature and offering recommendations for future research endeavours.

KEYWORD: Employee attrition, machine learning, workforce, model, dataset.

TABLE OF CONTENTS

STATEMENT OF AUTHENTICITY i

ACKNOWLEDGEMENT ii

DEDICATION iii

ABSTRACT iv

KEYWORD iv

LIST OF FIGURES ix

LIST OF TABLES x

CHAPTER ONE: INTRODUCTION 1

1.1 BACKGROUND OF THE STUDY 1

1.2 PROBLEM STATEMENT 1

1.3 RESEARCH AIM 2

1.4 RESEARCH OBJECTIVES 2

1.5 RESEARCH QUESTIONS 2

1.6 SIGNIFICANCE OF THE STUDY 3

1.7 STRUCTURE OF THE STUDY 3

CHAPTER TWO: LITERATURE REVIEW 5

2.1 INTRODUCTION 5

2.2 OVERVIEW OF EMPLOYEE ATTRITION 5

2.3 MEASURING ATTRITION 5

2.4 FACTORS CONTRIBUTING TO EMPLOYEE ATTRITION 6

2.5 EFFECTS OF EMPLOYEE ATTRITION 7

2.6 EMPLOYEE ATTRITION TRENDS 7

2.7 EMPLOYEE RETENTION STRATEGIES 8

2.8 TRADITIONAL ANALYTICS TECHNIQUES 9

2.9 MACHINE LEARNING TECHNIQUES 9

2.9.1 TYPES OF MACHINE LEARNING 10

2.10 REVIEW OF RELATED WORK 11

2.11 RESEARCH GAP 14

CHAPTER THREE: METHODOLOGY 15
3.1 INTRODUCTION 15
3.2 RESEARCH PHILOSOPHY 15
3.3 RESEARCH APPROACH 15
3.4 RESEARCH DESIGN 16
3.4.1 BUSINESS UNDERSTANDING 17
3.4.2 DATA UNDERSTANDING 18
3.4.2.1 DATA COLLECTION 18
3.4.2.2 DATA DESCRIPTION 18
3.4.2.3 DATA EXPLORATION 19
3.4.3 DATA PREPARATION 19
3.4.3.1 HANDLING MISSING VALUES 19
3.4.3.2 HANDLING INCONSISTENT DATA 19
3.4.3.3 FEATURE SELECTION 20
3.4.3.4 ENCODING 20
3.4.3.5 NORMALIZATION 20
3.4.3.5 DATA SAMPLING 20
3.4.3.6 DATA SPLIT 21
3.4.4 MODELLING 21
3.4.4.1 LOGISTIC REGRESSION 21
3.4.4.2 RANDOM FOREST CLASSIFIER 22
3.4.4.3 SUPPORT VECTOR MACHINE (SVM) 22
3.4.4.4 EXTRA TREE CLASSIFIER 22
3.4.5 EVALUATION 23
3.4.5.1 ACCURACY 24
3.4.5.2 PRECISION 24
3.4.5.3 RECALL 24
3.4.5.4 F1- SCORE 24
3.4.5.5 K-FOLD CROSS-VALIDATION 24

3.4.6 DEPLOYMENT 25
CHAPTER FOUR: RESULTS AND FINDINGS 26
4.1 INTRODUCTION 26
4.2 DATA UNDERSTANDING 26
4.2.1 UNIVARIATE ANALYSIS 28
4.2.2 BIVARIATE ANALYSIS 33
4.2.3 MULTIVARIATE ANALYSIS 36
4.3 DATA PREPARATION 41
4.3.1 HANDLING INCONSISTENT DATA 42
4.3.2 NORMALIZATION 42
4.3.3 ENCODING 43
4.3.4 DATA SAMPLING 43
4.3.5 FEATURE SELECTION 45
4.4 MODEL RESULTS AND EVALUATION 48
4.4.1 RANDOM FOREST CLASSIFIER 48
4.5.2 LOGISTIC REGRESSION MODEL 49
4.5.3 SUPPORT VECTOR MACHINE MODEL 50
4.5.4 EXTRA TREE CLASSIFIER MODEL 51
4.5.5 K-FOLD CROSS-VALIDATION 53
4.5.6 EVALUATION OF MODEL WITH IMBALANCED DATA 54
CHAPTER FIVE – CONCLUSION 56
5.1 DISCUSSION OF FINDINGS 56
5.1.1 PERFORMANCE OF MACHINE LEARNING 56
5.1.2 IDENTIFICATION OF KEY ATTRITION PREDICTORS AND RISKS 56
5.2 CONTRIBUTIONS OF RESEARCH 56
5.3 REFLECTION 57
5.4 LIMITATIONS OF RESEARCH 57
CHAPTER SIX: RECOMMENDATIONS 58
6.1 SUMMARY OF FINDINGS 58

6.2 RECOMMENDATION TO HUMAN RESOURCE MANAGERS 58
6.3 RECOMMENDATIONS FOR FUTURE RESEARCH 58
REFERENCES 59
APPENDICES 64
1. PYTHON CODE 64
2. RESEARCH PROPOSAL 65

LIST OF FIGURES

Figure 1: Employee attrition of professional services organisations worldwide 2013-2022 (Statista 2023)..... 8
Figure 2 Types of machine learning (Geeksforgeeks 2023)..... 10
Figure 3: CRISP- DM methodology (Shearer 2000) 17
Figure 4: Random Forest – Banerjee (2020) 22
Figure 5: Confusion Matrix 23
Figure 6: Histogram of Age..... 29
Figure 7: Histogram of DailyRate 29
Figure 8: Histogram of DistanceFromHome 30
Figure 9: Histogram of MonthlyIncome..... 31
Figure 10: Bar plot of Attrition..... 31
Figure 11: Bar Plot of Department..... 32
Figure 12: Bar Plot of Education 33
Figure 13 Bar plot Gender vs Attrition 34
Figure 14: Bar plot WorkLifeBalance vs Attrition..... 35
Figure 15: Histogram of PercentageSalaryHike vs Attrition 35
Figure 16 Boxplot of Attrition vs Gender and YearsSinceLastPromotion..... 36
Figure 17: Bar plot of Education vs Gender and PerformanceRating 37
Figure 18: Scatter plot of Monthly income and Age vs Attrition 38
Figure 19: Correlation Matrix Heatmap 39
Figure 20: Box plot for numerical variables 40
Figure 21: Scatter plot for Attrition vs MonthlyIncome and EmployeeNumber..... 41
Figure 22 Unique values in Education variable..... 42
Figure 23: Distribution of Attrition before applying SMOTE..... 44
Figure 24: Distribution of Attrition after applying SMOTE..... 45
Figure 25: Feature Important scores of numerical variables..... 46
Figure 27: Correlation of input features with Attrition 47
Figure 28: Random Forest Classifier model result..... 49
Figure 29: Logistic Regression model result..... 50
Figure 30: Support Vector Machine model result..... 51
Figure 31: Extra Tree Classifier model result..... 52
Figure 32: Model accuracy on test data 53
Figure 33 K-fold cross-validation mean accuracy score 54
Figure 34: Model accuracy on test data before applying SMOTE..... 55

LIST OF TABLES

Table 1: Review of related work 12
Table 2: Description of dataset variables 27
Table 3: Descriptive statistics of numerical variables 28
Table 4: Feature importance scores of categorical variables 46

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Employees are invaluable assets within organisations, capable of shaping brand perception and driving profitability (Hobson 2019). However, factors such as neglect and other reasons often lead to employees voluntarily or involuntarily leaving their jobs, a phenomenon known as employee attrition (Gartner 2023). This departure poses significant challenges to organisational performance, as emphasised by the CIPD (2023), especially when skills are scarce, and recruitment is costly. Wallace (2023) further highlights the detrimental effects of employee attrition, including loss of productivity, employee burnout, expertise depletion, and the resources expended in hiring and training new staff. Such costs create disruptions within organisations, affecting service delivery and operational efficiency.

In the aftermath of the COVID-19 pandemic, Deberry (2021) reveals that one in four employees plans to seek alternative employment, adding to the complexity of workforce management. Consequently, employee attrition becomes a pressing concern for organisations, prompting employers to devise retention strategies. Understanding the causes, effects, and costs associated with employee attrition is essential for organisations to mitigate its impact effectively.

Addressing these challenges requires organisations to adopt proactive measures, including machine learning techniques for predicting employee attrition. As noted by Spain and Groysberg (2016), machine learning offers a data-driven approach that surpasses traditional methods like exit interviews, which rely on retrospective analysis and may be less effective in managing attrition. Therefore, this study aims to contribute to organisational solutions by exploring the potential of machine learning in predicting employee attrition, aiming to prevent its adverse effects on organisational performance.

1.2 PROBLEM STATEMENT

The surge in voluntary employee attrition rates across various industries in the US, such as wholesale trade (25%), leisure and hospitality (3.5%), and arts, entertainment, and recreation (12.73%), as reported by the US Bureau of Labour Statistics (BLS 2023), underscores the critical challenge of employee attrition faced by organisations today. Additionally, the global workforce's increasing inclination towards changing jobs, as highlighted by Microsoft's Work Trend Index (Pellegrini 2023), further emphasizes this pressing issue.

Factors such as the enduring impact of the COVID-19 pandemic and the shifting dynamics of modern work culture have further exacerbated this issue, resulting in significant talent loss and disruptions in workforce stability. For example, recent surveys, including those among US healthcare workers, have shown that a notable percentage of employees attribute their exits

to pandemic-related factors (Galvin, 2021). Additionally, emerging trends like employees prioritising factors beyond salary, such as remote or hybrid work arrangements (Miranda 2023), underscore the complex drivers of attrition in today's workplace.

This escalating trend of employee attrition poses substantial risks to organisational performance and service delivery, especially in critical sectors like healthcare, where workforce stability directly affects patient care outcomes. Mitigating these risks requires a more resilient workforce, and proactive measures such as predicting employee attrition using machine learning. Such efforts aim to identify the root causes of attrition and predict employee departures before they occur.

1.3 RESEARCH AIM

This research aims to leverage machine learning (ML) predictive models and available data to forecast employee attrition, empowering employers and HR managers with precise insights for effective workforce management. Building upon the observations of Najafi-Zangeneh et al. (2021) regarding inadequate analytics on employee attrition, ML techniques offer a promising avenue for more accurate predictions. In addressing the limitations of traditional methods like exit interviews and continuous feedback, known for inconsistent predictions, this study seeks to leverage ML algorithms for precise attrition prediction.

Furthermore, the research aims to identify and interpret factors contributing to employee attrition, enabling organisations to understand the underlying causes and implement targeted retention strategies.

1.4 RESEARCH OBJECTIVES

Achieving the research aim requires meeting the following objectives:

1. Use machine learning algorithms/models and available data to predict employees likely to leave the organization.
2. Evaluate relevant machine learning models and recommend the most suitable model based on their performance.
3. Analyse data variables or features to ascertain their importance in determining employee attrition and to understand the key factors contributing to attrition.

1.5 RESEARCH QUESTIONS

The following research questions address the research objectives of the work:

1. To what extent can machine learning predict employee attrition?
2. Which machine learning model produces the most accurate predictions?
3. Which features or variables are most important in predicting employee attrition?

1.6 SIGNIFICANCE OF THE STUDY

This study holds significant importance in the following ways:

1. It promotes proactive workforce management and enhances the retention of valuable employees, surpassing the limitations of reactive traditional approaches.
2. Employers and HR managers can utilize the research findings to reduce costs associated with losing talented employees and minimize expenses related to hiring or mis-hiring new personnel.
3. The study provides valuable insights to stakeholders, enabling them to make data-driven decisions that enhance strategic planning and decision-making.

1.7 STRUCTURE OF THE STUDY

This section provides an outline of the structure of the project, which consists of six chapters as follows:

Chapter One: Introduction

This chapter serves as an introduction to the research work, establishing the background, aim, objectives, research questions, and significance of the study.

Chapter Two: Literature Review

This section reviews relevant literature on employee attrition prediction using machine learning. It covers various aspects such as the overview of employee attrition, types of attrition, methods of measuring attrition, factors contributing to employee attrition, effects of attrition, trends in employee attrition, strategies for employee attrition management, as well as general ethical and theoretical concepts relevant to employee attrition, machine learning, and HR. Similarly, this chapter explores machine learning models and traditional approaches to predicting employee attrition. Furthermore, this chapter identifies the research gap to be filled by this research.

Chapter Three: Methodology

Chapter three introduces the methodology employed in this study, encompassing the research philosophy, research design, sampling approach, sampling size, data analysis techniques, machine learning model selection and justification, model training, model evaluation, model interpretations, and ethical considerations.

Chapter Four: Results and Findings

This chapter presents the findings, including the data analysis and model results. It also correlates the findings with the research questions posed in Chapter One.

Chapter Five: Conclusion

Chapter five focuses on the evaluation criteria, performance metrics, and interpretation of findings derived from the study. The chapter further emphasizes the contributions, practical implications, challenges, and limitations of the research.

Chapter Six: Recommendations

This chapter highlights the findings and provides appropriate recommendations.

CHAPTER TWO: LITERATURE REVIEW

2.1 INTRODUCTION

The objective of this section is to conduct a thorough examination and synthesis of the existing body of knowledge concerning employee attrition prediction using machine learning. This task involves gathering insights from experts and established authorities in the field. According to Bennett (2020), conducting a literature review is crucial for consolidating current knowledge and assessing various viewpoints from diverse domains. Reviewing related literature will provide this research with a comprehensive understanding of past scholarly work. Additionally, it will enable the expansion of previous findings and identification of research gaps.

2.2 OVERVIEW OF EMPLOYEE ATTRITION

Employee attrition refers to the reduction in the workforce resulting from resignations, retirements, or layoffs that are not followed by hiring replacements (D'Allessandro, 2023).

Alduayi and Rajpoot (2018) explain employee attrition as either a voluntary or involuntary departure of employees due to unfavourable work conditions, personal reasons, or job dissatisfaction. Alduayi and Rajpoot (2018) further distinguished between voluntary attrition, where valued employees leave despite attempts to retain them, and involuntary attrition, where employers terminate contracts due to business requirements or poor employee performance.

Expanding on this perspective, Chemuturi and Chemuturi (2019) categorize attrition as planned or unplanned. A planned attrition is predictable and allows organizations to make necessary preparations. Examples of planned attrition include downsizing and retirement. However, unplanned attrition can pose challenges for organizations, especially if they lose talented employees whose services are still required.

The definition provided simplifies employee attrition as the exit of personnel due to retirement, resignation, or termination without immediate replacement. Another perspective distinguishes between voluntary and involuntary attrition, highlighting the different factors contributing to employees leaving organizations.

2.3 MEASURING ATTRITION

The attrition rate refers to the rate of exited employees relative to the average number of workers employed for the period. (Jones 2023).

That is attrition rate = number of exited employees ÷ average number of workers employed for the period × 100.

$$\text{Attrition Rate (\%)} = \frac{\text{No. of Employees That Left During Period}}{\text{Average No. of Employees For Period}} \times 100$$

Equation 1: Calculation of attrition rate (Jones 2023)

For instance, if the number of employees that left company A in the year 2023 is 50 and the average number of employees for the same company in the same year is 2000, then,

Attrition rate (%) = 50 ÷ 2000 * 100

= 2.5 %

Employee attrition rate can be a mirror of an organization's work culture, hiring policies, and overall workforce management (Shweta 2022). Understanding attrition rates and comparing same to industry standards can help businesses measure and improve workforce engagement.

Hayes (2023) also emphasized the importance of measuring attrition by admitting that it aids in identifying and addressing the underlying causes of attrition. By measuring attrition, organizations can better understand its root causes and take timely action to mitigate negative impacts.

2.4 FACTORS CONTRIBUTING TO EMPLOYEE ATTRITION

Employee attrition can stem from various factors, as outlined by Kumar (2023), including poor leadership, inadequate working conditions, limited career advancement opportunities, a lack of recognition, and subpar benefits and salaries. These factors constitute the features or inputs analysed in this study. Robinson (2023) echoes similar sentiments, citing causes such as retirement, financial motives, life events, career growth, organisational shifts, and skill mismatches. It's crucial to differentiate between employee-caused and employer-caused attrition; the former is initiated by employees, while the latter is influenced by employers. Lucas (2023) supplements these findings by highlighting additional triggers for attrition, such as workforce demographics, voluntary resignations, company restructuring, financial challenges, strategic shifts, technological progress, and outsourcing. Demographic characteristics can affect how quickly skill gaps are filled post-departure, particularly if few individuals are

available to assume certain roles. Moreover, restructuring, strategic changes, technological advancements, and outsourcing may result in redundancies, contributing to attrition. Economic and political factors may also come into play, with organisations opting to cease operations due to unfavourable market conditions or government policies, leading to employee departures.

2.5 EFFECTS OF EMPLOYEE ATTRITION

According to Singh and Singh (2019), employee turnover may lead to both advantageous and unfavourable outcomes, such as elevated expenses for staffing and training, and lowered production standards. Shweta (2022) identified the negative consequences of attrition, including additional recruitment expenses, shortages of skilled employees, and diminished team cohesion. On the other hand, employees may have to deal with inadequate work-life balance, emotional distress, and stress. However, one positive effect of attrition is the infusion of new technical expertise and ideas by new replacement employees.

Therefore, employers and managers must discern when to implement employer-caused (planned) attrition or manage employee-caused attrition. While employer-caused and voluntary attrition may pose minimal concerns, addressing voluntary and employee-caused attrition demands proactive measures, with which this research aims to assist.

2.6 EMPLOYEE ATTRITION TRENDS

This section delves into the trajectory of attrition rates over time, emphasising the urgency for proactive measures such as machine learning adoption. In the UK, a PwC (2022) workforce survey disclosed that 20% of employees expressed intentions to switch employers within twelve months. This trend persists, with attrition rates projected at 20% for 2024 in the UK, 18% in the US, 19% in Australia, and 23% in Germany (Moss, 2023), illustrating a global phenomenon. The upsurge in attrition rates is attributed partly to COVID-19 and the dynamic job market landscape.

De Smet et al. (2022) noted a surge in job openings in the United States, reaching 11.3 million by May 2022, compared to 9.3 million in April 2021, underscoring attrition's profound impact on corporate entities and the broader economy. Additionally, CIPD (2023) revealed that over one-third of UK employees depart from their jobs annually. Similarly, Statista’s 2023 study depicted a consistent rise in the attrition rate among professional service organizations globally from 2013 to 2022, as shown below.

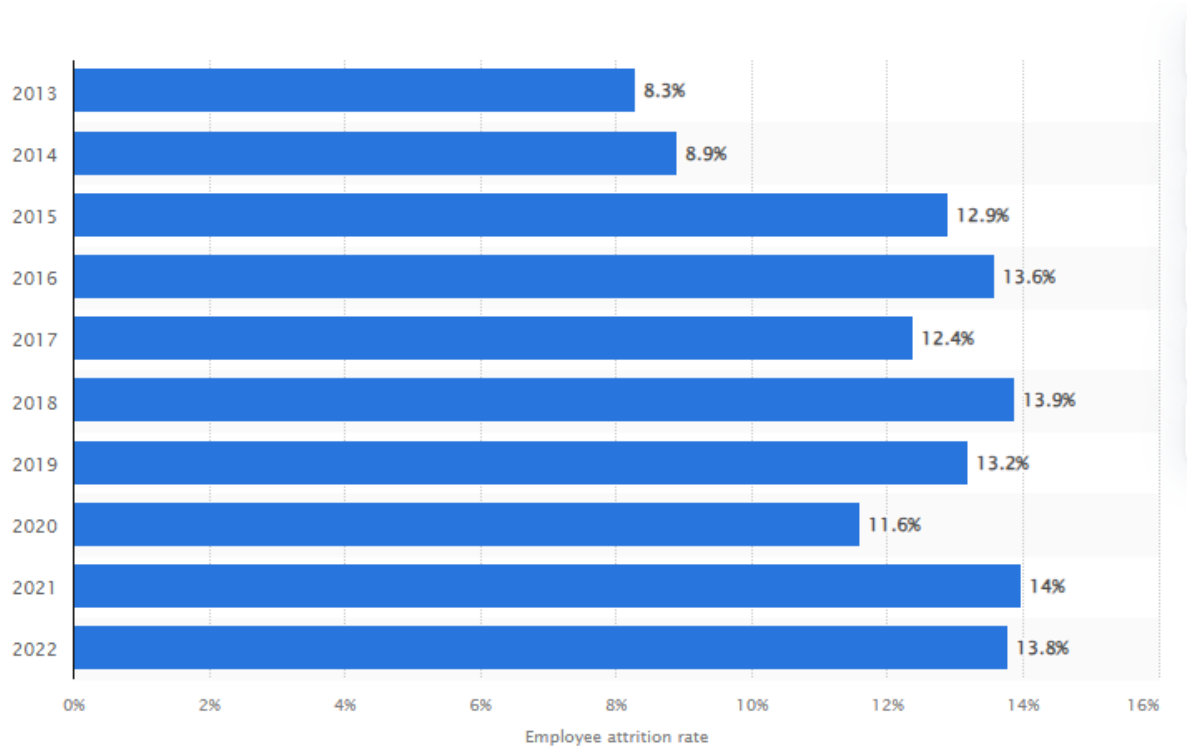

Figure 1: Employee attrition of professional services organisations worldwide 2013-2022 (Statista 2024)

Employee attrition trend reveals a global phenomenon with significant implications for organizations worldwide. Surveys and studies conducted by various sources indicate the pervasive nature of attrition, with projected rates showing a consistent increase across various countries and industries. The impact of COVID-19 and evolving job market dynamics further exacerbates the trends, emphasizing the urgent need for proactive measures.

2.7 EMPLOYEE RETENTION STRATEGIES

Masese (2016) observed that managing attrition requires fostering a conducive work environment, implementing effective leadership, trusting employees' judgement, empowering staff through professional growth opportunities, offering competitive benefits, promoting work-life balance, and enhancing engagement.

Similarly, Crail (2023) proposed several impactful employee retention strategies, including offering competitive salaries, adopting flexible work patterns (such as work-from-home or hybrid arrangements), cultivating a welcoming culture, supporting work-life balance, acknowledging and appropriately rewarding employees, facilitating personal and professional development, and discerning the right time to part ways with an employee.

Roberthalf (2023) emphasised the importance of initiating retention efforts from onboarding, advocating for employee training and development, and maintaining a system of continuous

employee feedback. These strategies collectively aid employers in managing their workforce effectively.

The diverse perspectives showcase the importance of implementing varied strategies to manage employee attrition effectively and enhance retention. By cultivating a positive work environment, providing strong leadership, and empowering employees through growth opportunities, organisations can foster a culture where employees feel valued and motivated to remain. Additionally, offering competitive salaries, flexible work arrangements, and recognising employees' contributions further contributes to job satisfaction and loyalty.

2.8 TRADITIONAL ANALYTICS TECHNIQUES

Holtom (2019) identified traditional techniques for predicting employee attrition, such as exit interviews and annual employee surveys, as valuable for understanding why employees leave. Workday (2021) supported this view but suggested that while exit interviews provide insights, regular employee feedback is more effective for managing attrition. Both sources agreed that data from exit interviews and periodic feedback can help identify attrition factors and speculate on employee exits. For example, exit interviews allow employees to disclose reasons for their departure, aiding in identifying common causes of attrition and areas for retention improvement. However, exit interviews are reactive and may be biassed (Turits 2021). The data collected may not provide honest responses, and making predictions with such data can become complex and challenging when a large volume of interview data is involved. Spain and Groysberg (2016) also argued that exit interviews often do not improve retention rates. Similarly, Mizar (2018) notes that traditional analytics are primarily descriptive and cannot predict attrition for each employee, unlike predictive analytics, which leverages evidence and advanced techniques such as machine learning.

In summary, traditional attrition prediction techniques are typically reactive and less effective at improving employee retention compared to predictive analytics, which rely on statistical models for predictions.

2.9 MACHINE LEARNING TECHNIQUES

Machine learning, a branch of computer science and artificial intelligence, mimics human learning processes and improves accuracy over time through data and algorithms (IBM, 2023). In employee attrition prediction, machine learning utilises data, algorithms, and computations to discern patterns from employee data, like human cognitive processes, thereby gaining insights into employee attrition trends. With technology, predictive analytics has gained prominence, enabling organisations to forecast potential attrition trends (Sghir et al., 2022). Unlike in the past, where employee attrition analytics lacked a data-driven approach (Rane and Narvel 2022), today's landscape integrates data-driven methodologies for effective

workforce management (Varma and Dutta 2023). This evolution has significantly enhanced the understanding of employee attrition trends and facilitated proactive measures to address potential attrition issues, marking a departure from traditional practices.

2.9.1 TYPES OF MACHINE LEARNING

The four types of machine learning outlined by Coursera (2024) include:

1. Supervised machine learning
2. Unsupervised machine learning
3. Semi-supervised learning
4. Reinforcement learning

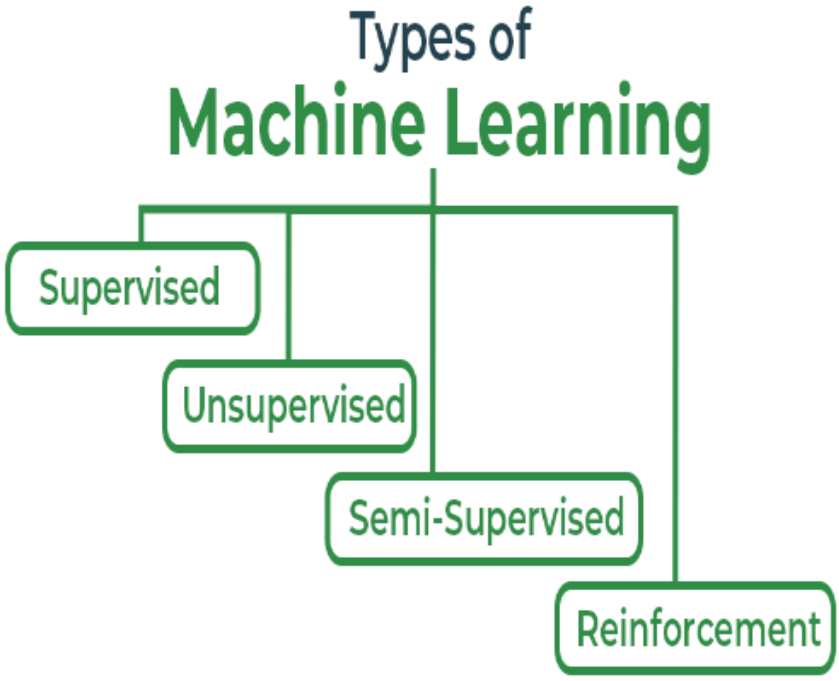

Figure 2 Types of machine learning (GfG 2023)

1. SUPERVISED MACHINE LEARNING

According to Ali (2022), supervised machine learning algorithms learn data patterns from labelled or known input and output data. The supervised machine learning algorithms include regression and classification algorithms. For instance, supervised machine learning algorithms are used for predicting employee attrition because employee input data (age, year of service, grade, etc.) as well as the output (attrition or stay) are labelled and need a supervisor (human) to instruct the algorithm to generate output using the input features.

2. UNSUPERVISED MACHINE LEARNING

Pasquarella (2023) argued that unsupervised machine learning is unsuitable for predicting employee attrition due to the labelled nature of employee data. Wood (2023), on the other hand, defines unsupervised machine learning as a method where models discover patterns and draw conclusions from unlabelled datasets. In unsupervised machine learning, algorithms identify patterns from unlabelled data and generate results. Types of unsupervised machine learning techniques include clustering and association.

3. SEMI-SUPERVISED MACHINE LEARNING

Semi-supervised machine learning, which utilises labelled and unlabelled data, is positioned between supervised and unsupervised learning methods (Stomps et al., 2023). Semi-supervised learning aims to address the limitations of both supervised and unsupervised approaches.

4. REINFORCEMENT MACHINE LEARNING

Reinforcement learning describes a machine learning approach where an agent learns by interacting with an environment, making decisions through trial and error, and receiving feedback in the form of rewards or penalties to achieve a defined objective (Wang et al. 2023). The environment represents the context in which the agent operates, with the agent being the algorithm that learns to accomplish its goal. Reinforcement learning algorithms comprise techniques like Monte Carlo and Q-learning. However, reinforcement learning does not apply to this research as it pertains to a supervised machine-learning problem.

2.10 REVIEW OF RELATED WORK

This research reviewed related literature and provided the summary of their results below.

RELATED WORK ON EMPLOYEE ATTRITION			
Author (s)	**Research objective**	**Machine Learning (ML) Model used**	**Recommended ML Model**
Mansor et al. (2021)	To predict employee attrition.	Decision Tree Classifier, Support Vector Machines (SVM) Classifier, Artificial Neural Network (ANN) Classifier.	SVM classifier
Muneera et al. (2023)	To investigate the effect of ensemble learning techniques on predicting employee attrition.	Naïve Bayes Classifier, Support Vector Machines (SVM), Random Forest (RF), Stacking ensemble learning technique, Voting ensemble learning technique.	Random Forest Classifier
Raza et al (2022)	To analyze factors that causes employee attrition and predict employee attrition using machine learning techniques.	Logistic Regression Classifier, Decision Tree Classifier, Extra Trees Classifier.	Extra Trees Classifier
Guerranti et al. (2023)	To apply machine learning methodologies for employee attrition prediction.	Logistic Regression Classifier, Decision Tree Classifier, Random Forest Classifier, Artificial Neural Network (ANN) Classifier, Naïve Bayes Classifier.	Logistic Regression Classifier and Random Forest Classifier
Repaso et al. (2022)	To determine factors resulting to employee attrition using data mining technique.	Naïve Bayes Classifier	Naïve Bayes Classifier
Raman et al. (2019)	To ascertain if email communication affects employee attrition.	correlation coefficient	correlation coefficient
Bennett B. (2020)	To understand whether employees would leave a company within the first three years of being employed.	Gradient Boosting, Random Forest, Linear Discriminant Analysis and Classification.	Random Forest
Ali et al. (2022	To predict employee attrition and find its causes.	Decision Tree Classifier, Logistic Regression, Random Forest, Support Vector Machines and Extra Trees Classifier (ETC).	Extra Trees Classifier (ETC)
Pasquarella C. (2023)	To ascertain the factors that influence employees to leave an organization.	K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression, Support Vector Machine, Random Forest Decision Tree, Gradient Boosting, Etra Trees Classifier etc.	Extra Trees Classifier (ETC)
Yang et al (2020)	To find the reasons behind employee resignation.	Random Forest Classifier, K-means Clustering, Random Forest and K-means Clustering	Logistics Regression

Table 1: ***Review of related work***

The related works on employee attrition prediction using machine learning present a diverse range of objectives, methodologies, and recommendations for future research. These related works are sourced from reputable academic journals published on various platforms, like ProQuest and ResearchGate.

Mansor et al. (2021) explored the use of Decision Tree Classifier, Support Vector Machines (SVM) Classifier, and Artificial Neural Network (ANN) Classifier for predicting attrition using the same IBM HR Analytics Employee Attrition and Performance dataset as used in this work, ultimately favouring the SVM classifier for predicting employee attrition with an accuracy score of 88.87%. Alshiddy and Aljabera (2023) using the same dataset, proposed ensemble learning techniques, and equally focused on Random Forest Classifier, Naïve Bayes, and Support Vector Machine and favoured the proposed model, which achieved an accuracy of 94.53%. Also, they used SMOTE to address the class imbalance of the no attrition class. Raza et al. (2022) analysed attrition factors using Logistic Regression (LR) Classifier, Decision Tree Classifier (DTC), and Extra Trees Classifier, with a preference for the Extra Trees Classifier having achieved an accuracy score of 93% in predicting employee attrition. Similarly, Guerranti and Dimitri (2023) employed various machine learning methodologies, namely neural networks, random forests, logistic regression, and classification trees, with the IBM HR Analytics Employee Attrition and Performance dataset and recommended the logistic regression classifier as the best-performing model, having achieved an accuracy score of 85%.

Unlike other authors, Repaso et al. (2022) identified age, total working years, and marital status as the topmost important predictors of attrition, and they also found out that the model used, which is Naïve Bayes, can predict attrition with 84.69% cross-validation accuracy. Ali et al. (2022) found out that hourly rate, monthly income, age, and job level are the major causes of attrition, while also recommending the Extra Tree Classifier as the best-performing model with an accuracy score of 93% when compared to SVM, DTC, and LR and suggesting the use of deep learning techniques for future research. Pasquarella C. (2023) equally identified key factors influencing attrition, including years with the agency, years in assignment, and age, among others, as the most important features in predicting employee attrition, while also advocating for further exploration of datasets with external factors as variables, and equally proposed the Extra Trees Classifier as the best-performing model, having gained an accuracy score of 95.13%.

Yang et al. (2020) also investigated the IBM HR Analytics Employee Attrition and Performance dataset and found that age, monthly income, and the number of companies worked have a significant impact on employee attrition. They used random forest and logistic regression, with the former performing better with an accuracy score of 88.43%. Bennett (2020) identified factors like relocation, site experience, and rehire status as potential indicators that employees may voluntarily leave their organisation within three years. And among the multiple models used, Random Forest performed best with an accuracy score of 86.80%. Bennett (2020) also suggested further study be carried out using logistic regression. Raman et al. (2019) focused

on email communication patterns as they affect employee attrition and suggested expanding the dataset to include multiple business schools for comprehensive analysis. They disclosed that sentiment exhibited by employees in their official email can be an attrition predictor. They used the R programming language, unlike most of the other authors who used Python programming language in their analysis. The work of Alkhateeb (2023) was also valuable in this research as he explored various machine learning models using the same dataset and favoured the KNN classifier with 93.80 accuracies as the best-performing model.

Looking forward, future research directions proposed by the authors offer valuable insights. Suggestions include expanding datasets to include multiple institutions, developing real-time workplace tools for candidate evaluation, exploring nested ensemble techniques, and applying advanced deep learning methods like graph neural networks. Additionally, there's a call for identifying key features leading to attrition and incorporating predictive models into real-world systems for enhanced decision-making. Overall, these studies collectively contribute to the ongoing efforts to improve attrition prediction models and understand the underlying dynamics of employee turnover.

2.11 RESEARCH GAP

Recognising the significant contributions made by various authors in employee attrition prediction using machine learning, this work aims to address a specific gap in the existing literature. While previous studies have provided valuable insights, there is a need to enhance the accuracy of commonly used machine learning algorithms and to conduct a thorough analysis of the predictors of employee attrition.

In filling this gap, this research will explore alternative approaches beyond those commonly employed in the literature. By employing verifiable methodologies and considering various factors influencing attrition, the aim is to achieve a more comprehensive understanding of the phenomenon and improve the predictive capabilities of machine learning algorithms.

CHAPTER THREE: METHODOLOGY

3.1 INTRODUCTION

This chapter delves into the methodology and design utilised in this research, providing a detailed overview of the steps undertaken and the rationale behind the chosen approach. It outlines the systematic process employed to address the research questions.

The methodology section commences with a discussion of the research philosophy, shedding light on the underlying assumptions and elucidating the chosen research approach and design.

This study relies on existing quantitative data (sourced from Kaggle, a popular platform for data science competitions) for analysis and interpretation. The Sci-kit Learn libraries were used to implement the task in the Python programming language. Various tools and environments, such as Kaggle, Google Scholar, University of Bradford's Summon, ProQuest, ResearchGate, Scikit-learn, and Stack Overflow, were also leveraged to carry out this research.

3.2 RESEARCH PHILOSOPHY

Research philosophy refers to the set of assumptions and beliefs that guide and shape the direction of the research process (Saunders et al. 2019). According to Bryman and Bell (2015), there are three primary philosophical assumptions: axiology, ontology, and epistemology, each used to define research philosophies. Axiological assumptions refer to how one's values influence the research process, while ontological assumptions relate to understanding the realities encountered during research and guiding effective research design. Epistemology concerns the theory of knowledge and the interpretation of business data. Researchers' assumptions play a crucial role in shaping their research objectives, methods, analysis of findings, and research questions. Saunders et al. (2019) have identified positivism, interpretivism, or constructivism, and realism as three primary research philosophies. There is no one-size-fits-all approach to selecting a research philosophy for business research. However, each philosophy provides a distinct and valuable perspective on the world of organizations. This research adopts a positivist philosophy, wherein researchers seek correlations between variables using a deductive approach, as emphasised by Bryman and Bell (2015).

3.3 RESEARCH APPROACH

There are three main research approaches: induction, abduction, and deduction. According to Saunders et al. (2019), the induction approach involves subjective interpretations of qualitative data, whereas the deductive approach aims to explain causal relationships between variables, primarily utilising quantitative data. This approach aligns with the positivist research

philosophy, emphasising structure, quantification, and testable hypotheses. Conversely, the abductive approach integrates aspects of both deduction and induction.

The deductive approach has been used in this research to establish relationships between the output variable (attrition) and input variables using quantitative data. Conversely, the abductive approach integrates aspects of both deduction and induction.

3.4 RESEARCH DESIGN

Analysing a large volume of data to identify patterns and gain insights requires appropriate data mining techniques (Twin 2023).

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is the design for this research. The rationale behind this choice is that the methodology is all-encompassing, can be applied to any industry, and has been widely accepted (Martinez-Plumed et al., 2021). This standardised approach to data mining, CRISP-DM, breaks the process of data mining into six major stages, including business understanding, data understanding, data preparation, modelling, evaluation, and deployment. This approach is also illustrated in the diagram below and described in subsequent sections.

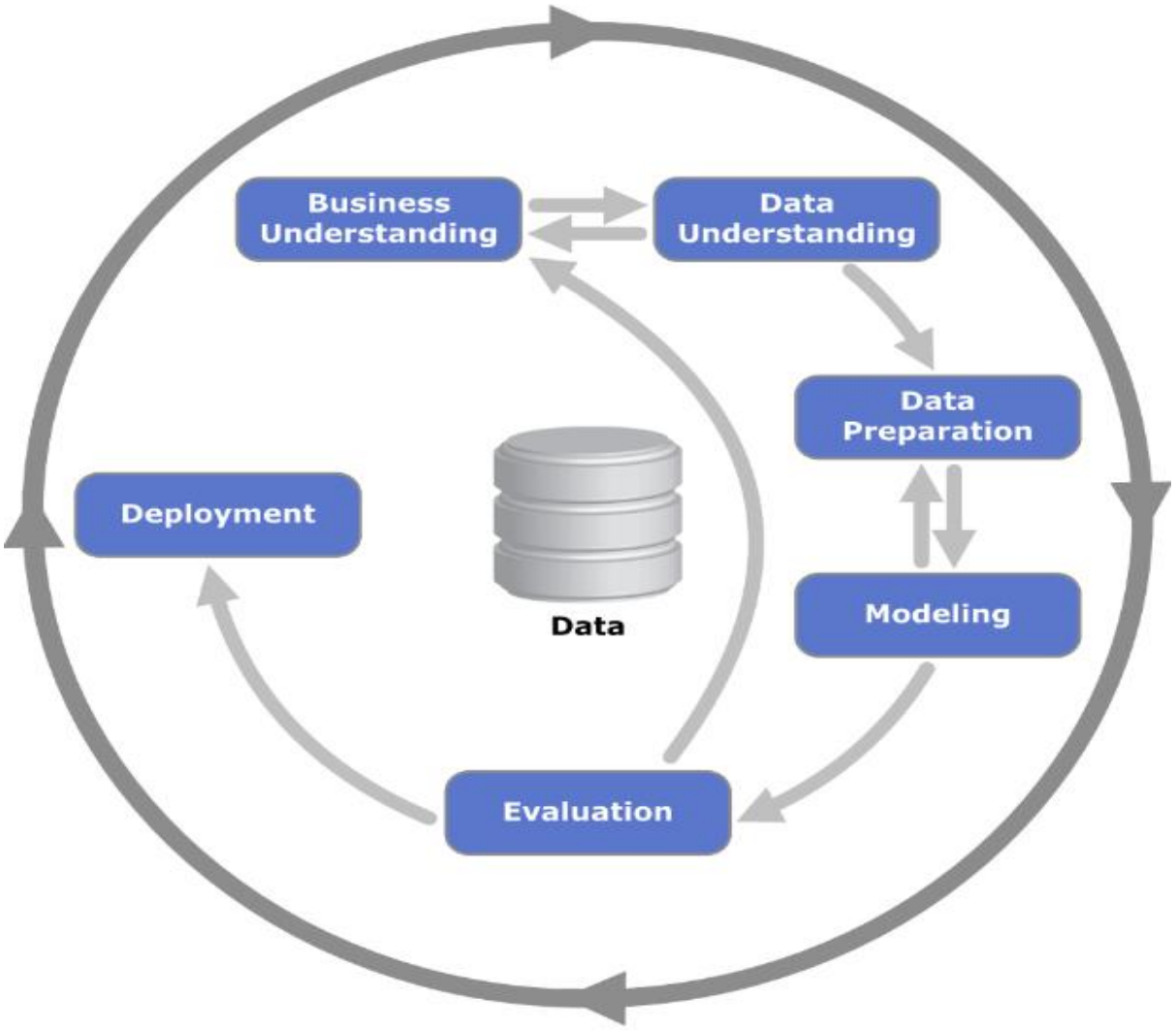

Figure 3: CRISP- DM methodology (Shearer 2000)

3.4.1 BUSINESS UNDERSTANDING

Business understanding marks the first stage of the CRISP-DM methodology and begins with a comprehensive understanding of customer needs, the nature of the business, and overarching business objectives (Hotz 2023).

This phase lays the foundation for guiding subsequent data mining processes, ensuring alignment between analytical efforts and organisational goals.

The dataset used for this research was developed by IBM, a research company, to facilitate this work and similar research.

This research aims to provide valuable insights to any organisation looking to improve employee retention and achieve its business goals.

In this phase, the research focuses on understanding the organisation's business objectives related to employee retention.

These objectives may include reducing attrition rates, improving employee satisfaction and engagement, and optimising workforce management strategies. Predictive analytics efforts are aligned with these objectives to help organisations achieve their goals.

3.4.2 DATA UNDERSTANDING

O'Hara et al. (2023) argue that data understanding typically involves data collection, exploration, description, and other activities that enhance data quality, accuracy, and validity. This process helps in understanding the source of the data and its attributes, especially as they relate to business objectives and solutions to relevant business problems. Data understanding is instrumental in interpreting the meaning of each variable in the dataset. Additionally, the use of the Python programming language for data understanding is discussed further in the following sections.

3.4.2.1 DATA COLLECTION

This research utilises secondary data from Kaggle, an external platform renowned for hosting datasets tailored for data science projects. IBM, a research company, created the dataset for researchers aiming to develop machine learning models for predicting employee attrition. The data is imported into the Google Colab environment using relevant Python libraries. Most of the datasets used in the literature are secondary data, like the one used in this research. However, there are a few exceptions. For instance, Bennett (2023) used data from a manufacturing firm in the US Southeastern region. Pasquarella (2023) used a dataset from the US Department of State.

3.4.2.2 DATA DESCRIPTION

Python libraries like Pandas, NumPy, and Seaborn were employed to describe the data and visualise its type and composition. The df.describe() function was utilised to generate summary statistics, including row count, mean, standard deviation, minimum and maximum values, and quartile values for each variable. This methodology is consistent with approaches used by other researchers in the literature. However, it is worth noting that Bennett (2022), Raman et al. (2022), and Salunkhe (2018) utilised the R programming language for their analyses.

During the analysis, the data was examined to determine the number of rows and columns in the dataset. Checks were also carried out to identify missing or duplicate values and to ascertain the data type of each variable, whether categorical (object) or numerical (integer and float). For categorical variables, further analysis was performed to identify unique values. This process revealed that the "Attrition" variable had unique values "Yes" and "No," while the "Marital Status" variable had unique values "single," "married," and "divorced." Additionally, the data description uncovered instances where variables that were supposed to be

categorical were coded as numerical in the dataset, such as "EnvironmentSatisfaction," "Education," and "JobLevel."

In summary, the data description helped comprehend the dataset and determine the appropriate analytical techniques and data preprocessing steps required for further analysis.

3.4.2.3 DATA EXPLORATION

Univariate, bivariate, and multivariate analyses investigated the dataset to identify patterns, trends, relationships, and anomalies within the data. In bivariate analysis, bar plots compare attrition against each categorical and numerical variable. Multivariate analysis compares the relationships between different variables. Correlation heatmaps and scatter plots assisted in identifying the relationships among various variables. For instance, bar plots visualise categorical variables like "attrition," indicating the count of employees who either left or stayed in the organization. Histograms displayed the distribution of numerical variables such as "age" and "monthly income." Additionally, boxplots and scatterplots identified the outliers and the data distribution.

3.4.3 DATA PREPARATION

This phase of CRISP-DM involves merging datasets, sorting data, splitting into training and test data, and replacing or removing missing and blank values (IBM 2021). When preparing data for analysis and modelling, it is crucial to ensure that it is accurate, complete, consistent, and reasonable (Fleckenstein and Fellows 2018). Data preparation for analysis and modelling involved the activities outlined below.

3.4.3.1 HANDLING MISSING VALUES

The Python code df.isnull().sum() was employed to check for missing values in the dataset. Identifying and rectifying missing values are crucial for maintaining the data sample size and enhancing model performance. For example, this check revealed no missing values in the dataset.

3.4.3.2 HANDLING INCONSISTENT DATA

During the data preprocessing stage, certain numerical features were converted into categorical features using Python's mapping function because the affected features possess characteristics of categorical variables, which suggests that they should have unique values.

The features subjected to this transformation include WorkLifeBalance, StockOptionLevel, RelationshipSatisfaction, PerformanceRating, JobSatisfaction, JobInvolvement, EnvironmentSatisfaction, Education, and JobLevel. Analysis of their minimum and maximum values showed the range of unique values each feature can have. For example, the statistics showed that JobLevel's minimum and maximum values were 1 and 5, respectively. The

variable "JobLevel" was transformed into a categorical variable with five unique values: Junior, Entry, Mid-Level, Senior, and Executive. Integer labels were assigned to each category using mapping dictionaries and then applied to the Data Frame. This conversion was necessary to maintain consistency in the data type across the entire dataset.

3.4.3.3 FEATURE SELECTION

Feature selection aimed to identify features for prediction while eliminating those exhibiting high multicollinearity and those deemed unimportant for the study, such as EmployeeNumber, StandardHour, Over18, and EmployeeCount. The ANOVA F-test was employed to rank numerical features based on their significance to the study. According to Siraj et al. (2022), selecting optimal features is crucial to enhancing model performance. The ANOVA F-test and the Chi-Square test were used with SelectKBest (a feature selection method in Sci-kit Learn) to determine the importance of categorical features. This method selects the top k features based on a specified scoring function. Feature importance scores of 30 and above were considered highly important for predicting attrition.

3.4.3.4 ENCODING

Some machine learning models, such as Logistic Regression and Support Vector Machine, are only compatible with numeric data (Wadikar 2020). Therefore, it was imperative to convert categorical variables into numerical variables to make them compatible with these algorithms. To achieve this, the Sci-kit Learn library was utilised for techniques like one-hot encoding, label encoding, and ordinal encoding, facilitating the transformation of categorical variables into numerical format.

3.4.3.5 NORMALIZATION

Skewed and kurtotic variables like YearsAtCompany and Monthly Income were normalized through StandardScaler and Yeo-Johnson's PowerTransformer to obtain a normal distribution.

Alkhateeb (2023) used power transformation and standard scaling to pre-process the data, making it more suitable for machine learning algorithms that perform better with normally distributed features on the same scale. Normalisation helped optimise outliers, fostered data consistency, and reduced dimensionality.

3.4.3.5 DATA SAMPLING

The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to address the class imbalance between the minority and majority classes in the dataset. Specifically, there were 237 instances where employees left the organisation, compared to 1233 instances where employees stayed. Such a significant disparity between observations is considered imbalanced and can adversely affect model performance. SMOTE oversampled the minority class (Attrition_Yes) to balance the dataset. This approach was chosen based on the success

of SMOTE in mitigating the challenges associated with class-imbalanced data, as highlighted by Li et al. (2021).

3.4.3.6 DATA SPLIT

The research dataset was split into training and testing sets, with 75% of the data allocated for training and 25% for testing. The training data was utilised for model building, while the testing data served as out-of-sample data to evaluate model accuracy and predictability. To ensure consistency in the train-test split process, a random state of 42 was set, which is the hyperparameter that controls the randomness in shuffling the data, facilitating reproducibility across different runs. In a similar experiment, Alkhateeb (2023) employed a random state of 0 and split the data into training and testing sets at a ratio of 67% and 33%, respectively.

3.4.4 MODELLING

Machine learning algorithms are used with the prepared data to predict employee attrition during this phase. The machine learning models employed include Logistic Regression, Random Forest Classifier, Support Vector Machine, and Extra Trees Classifier. These models are described below in the following sections:

3.4.4.1 LOGISTIC REGRESSION

According to Yahia et al (2021), Logistic Regression uses logistic functions and statistical techniques to predict categorical variables. The logistic Regression model is expressed in equation below.

$$logit_(p) = \ln \frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Equation 2: Logistic Regression

Where p is the probability of attrition (yes/no). Attrition represents y, the coefficients are $\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$

Equation 3: coefficients

A regression coefficient greater than 1 signifies positive influence on attrition, otherwise a negative influence on attrition.

3.4.4.2 RANDOM FOREST CLASSIFIER

Random Forest performs a classification task using multiple decision trees to vote on a class that an input variable belongs to (Molina 2021). Random Forest utilizes the result of multiple decision trees, and it is typically described in the diagram below, where the features can be Age, MonthlyIncome etc.

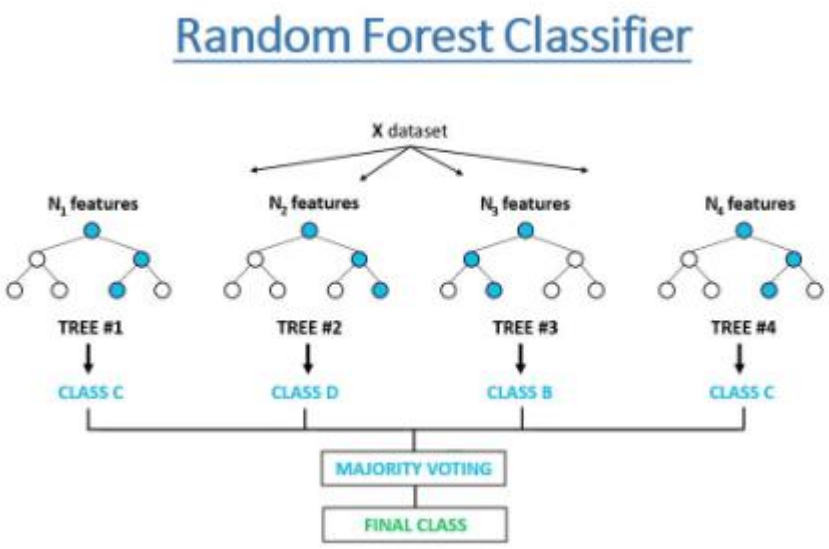

Figure 4: Random Forest – Prashant (2020)

3.4.4.3 SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) model, as described by Dong (2021), is a supervised machine learning approach that constructs an optimal decision boundary to segregate input data in an n-dimensional feature space into distinct target classes. This decision boundary, termed a hyperplane, effectively divides the feature space into two disjoint subsets. Raza et al. (2022) further elaborates that the SVM model's prowess lies in its capability to delineate a hyperplane, enabling the segregation of employee data into separate classes based on various features, thereby distinguishing between employees likely to remain with the company and those likely to depart.

3.4.4.4 EXTRA TREE CLASSIFIER

As outlined by Ossai and Wickramasinghe (2022), the Extra Trees Classifier (ETC) expands upon ensemble learning methodologies by employing a bagged ensemble of decision trees. The Extra Trees Classifier leverages multiple decision trees to forecast employee attrition by randomly selecting input variables to improve overall predictive performance and mitigate

overfitting. The primary goal of the Extra Trees Classifier is to optimise the ensemble of trees while minimising entropy, a standard criterion used in decision trees to partition data at each node, thereby diminishing uncertainty in the prediction process.

3.4.5 EVALUATION

In this phase, O'hara et al. (2023) emphasise that evaluating a model, entails considering several factors, including model performance, relevance to the business problem, simplicity, explainability, and cost-effectiveness. Adhering to these considerations will aid in recommending the most suitable model.

A commonly used and efficient method to evaluate the performance of models in classification problems is through a confusion matrix, as explained by Pasquarella (2023). This matrix provides a concise overview of predictions by tallying true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). Below is a diagram illustrating the structure of the confusion matrix and its components.

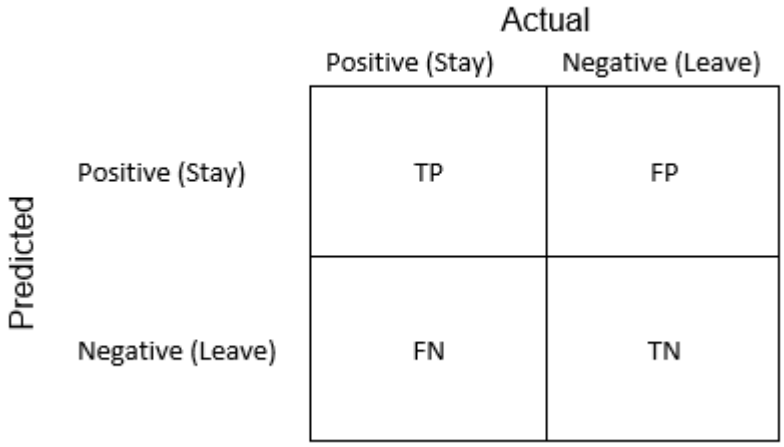

Figure 5: Confusion Matrix

Where:

TP stands for True Positive depicts that both the predicted outcome and actual result is true and that employee will stay.

TN is True Negative where both predicted outcome and actual value is false, and that employee will leave.

FP, which is False Positive, the employee left but prediction says that employee will stay.

FN is False Negative. Employee stays but prediction shows that employee will leave.

Evaluation using the matrix will be considered as follows:

3.4.5.1 ACCURACY

Accuracy is calculated as the total number of correct predictions out of the total number of employees observed. Accuracy is used here to measure how accurate the model performs in predicting that an employee leaves or stays.

Accuracy = TP + TN/ (TP+FP+TN+FN)

3.4.5.2 PRECISION

Precision measures the proportion of correctly predicted attrition cases. High precision means fewer false positives, indicating that when the model predicts attrition, it's usually correct.

Precision = TP / TP + FP

3.4.5.3 RECALL

Recall measures the proportion of correctly predicted attrition cases among all actual attrition cases. In other words, High recall indicates that the model is good at capturing actual attrition cases, minimizing false negatives.

Recall = TP / TP + FN

3.4.5.4 F1- SCORE

F1 Score is known for harmonizing and providing balance between the mean of precision and recall, offering a balanced assessment of the model's performance.

F1 Score = 2 x Precision x Recall / Precision +Recall

3.4.5.5 K-FOLD CROSS-VALIDATION

Pal and Patel (2020) proposed using k-fold cross-validation to determine the most reliable classifier from multiple models. This method involves assessing different data subsets to evaluate the variation in results. Therefore, k-fold cross-validation validates the models' performance across various employee dataset subsets. This approach helps to reduce the risk of overfitting and provides a more accurate estimate of the model's ability to generalize. The dataset is split into five folds using the cross_val_score function from Sci-kit Learn. Each fold is used once for testing, while the model trains on the other four folds. This process repeats five times, once for each fold, and the average of the performance scores derives the final performance estimate of the model.

3.4.6 DEPLOYMENT

The deployment phase is the last phase of the Cross-Industry Standard Process for Data Mining. At this stage, the selected model is applied in real life. This work did not explore the deployment of the models, thereby constituting part of its limitation.

CHAPTER FOUR: RESULTS AND FINDINGS

4.1 INTRODUCTION

This chapter offers an interpretation and analysis of the results derived from the study, serving as a crucial component in comprehending the research outcomes. As previously outlined, the CRISP-DM methodology guides this research endeavour. Therefore, this chapter will meticulously analyse and interpret the comprehension of the dataset, data preparation, modelling techniques, and evaluation of the models employed in this study.

4.2 DATA UNDERSTANDING

The dataset utilised in this research is structured and organised in a tabular format, making it suitable for supervised machine learning techniques. It consists of 1470 rows and 35 columns. Among these columns, one is the dependent or output variable (attrition), while the remaining thirty-four are independent or input variables. The dataset encompasses qualitative (categorical) and quantitative (numerical) variables. The process of understanding the data commenced with an in-depth study of the dataset and an examination of the meaning and significance of each variable. This exploration resulted in a summary, depicted in the figure below, which outlines the interpretation of each variable within the dataset.

DATASET VARIABLES		
S/N	Column	Description
1	Age	Age of employee
2	Attrition	employee has left the company or not
3	BusinessTravel	Frequency of business travel
4	DailyRate	The daily rate of pay for the employee
5	Department	The department in which the employee works.
6	DistanceFromHome	Distance from the employee's home to the workplace.
7	Education	Employee's level of education
8	EducationField	Field of education of the employee.
9	EmployeeCount	Number of employees
10	EmployeeNumber	The employee number meant for identification
11	EnvironmentSatisfaction	Satisfaction level with the work environment
12	Gender	Gender of the employee
13	HourlyRate	Employee hourly rate of pay
14	JobInvolvement	Level of job involvement
15	JobLevel	Level of the employee's job
16	JobRole	Job title of the employee
17	JobSatisfaction	Satisfaction level with the job
18	MaritalStatus	Marital status of the employee
19	MonthlyIncome	The monthly income of the employee
20	MonthlyRate	The monthly rate of pay for the employee.
21	NumCompaniesWorked	Number of companies the employee has worked for.
22	Over18	employee is over 18 years old or not
23	OverTime	Indicates whether the employee works overtime or not
24	PercentSalaryHike	The percentage increase in salary
25	PerformanceRating	Employee's performance rating
26	RelationshipSatisfaction	Satisfaction level with relationships at work
27	StandardHours	Standard number of working hours per week
28	StockOptionLevel	The stock option open to employees
29	TotalWorkingYears	Total number of years the employee has been working.
30	TrainingTimesLastYear	Number of training times attended by the employee in the last year.
31	WorkLifeBalance	Satisfaction with work-life balance
32	YearsAtCompany	Number of years the employee has been with the company.
33	YearsInCurrentRole	Number of years the employee has been in the current role.
34	YearsSinceLastPromotion	Number of years since the employee's last promotion.
35	YearsWithCurrManager	Number of years the employee has been with the current manager.

Table 2: Description of dataset variables

Statistical metrics for variables were analysed, especially for numerical ones. These metrics include the minimum, maximum, count, mean, standard deviation, and the 25th, 50th, and 75th percentiles. The values obtained are summarised in the table below.

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.0	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.0	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.0	7.0	14.00	29.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.0	66.0	83.75	100.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.0	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.0	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.0	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.0	14.0	18.00	25.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.0	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.0	3.0	3.00	6.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.0	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.0	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.0	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.0	3.0	7.00	17.0

Table 3: Descriptive statistics of numerical variables

The count of 1470 for each variable indicates no missing values in the affected columns. However, for some variables such as DailyRate, MonthlyIncome, and MonthlyRate, there is a significant difference between the standard deviation and the mean, suggesting a noticeable variation in employee pay rate, likely due to differences in job roles and levels. Normalisation will address this variation and achieve a balanced scale for the data.

4.2.1 UNIVARIATE ANALYSIS

This analysis was performed with histograms, bar plot and boxplot, to understand the data by visualizing the distribution of both numerical and categorical variables. The output of some of the analysis is shown below.

1. **Age:**

The "Age" variable represents the age of employees and is a numerical variable. The age range falls between 18 and 60, with a mean age of 36.92. The standard deviation is 9.14, indicating that ages deviate from the mean by approximately 9 years on average. The histogram below illustrates a normal distribution curve, indicating that most employees' ages cluster around the mean age.

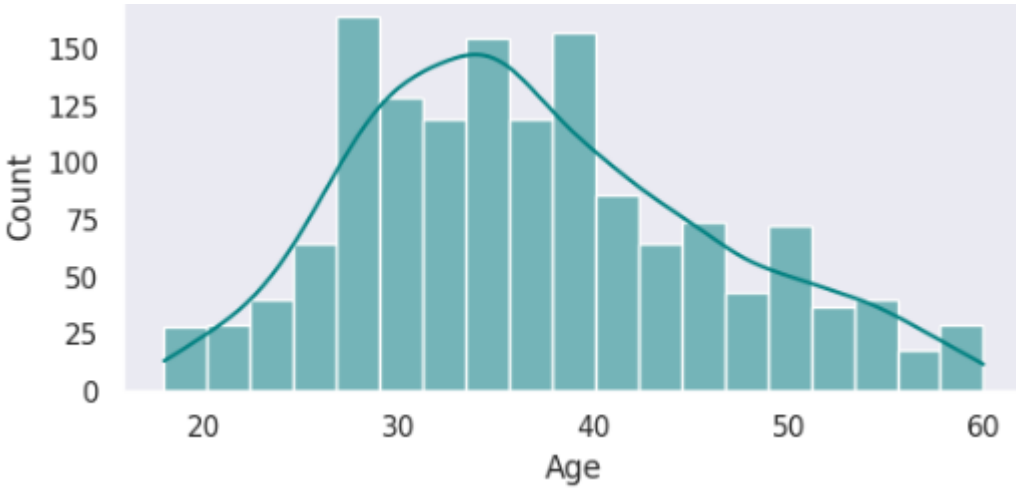

Figure 6: Histogram of Age

2. **DailyRate**:

The "DailyRate" variable represents the daily earnings of employees and is a numerical variable. With a mean DailyRate of 802.49, employees earn approximately 802.49 per day on average. The standard deviation of 403.51 indicates the spread or dispersion of DailyRate values around its mean. The range of DailyRate spans from 102 to 1499. The histogram below illustrates the distribution of this variable, showing that employees' earnings vary, with some earnings below and others above the mean DailyRate of 802.49.

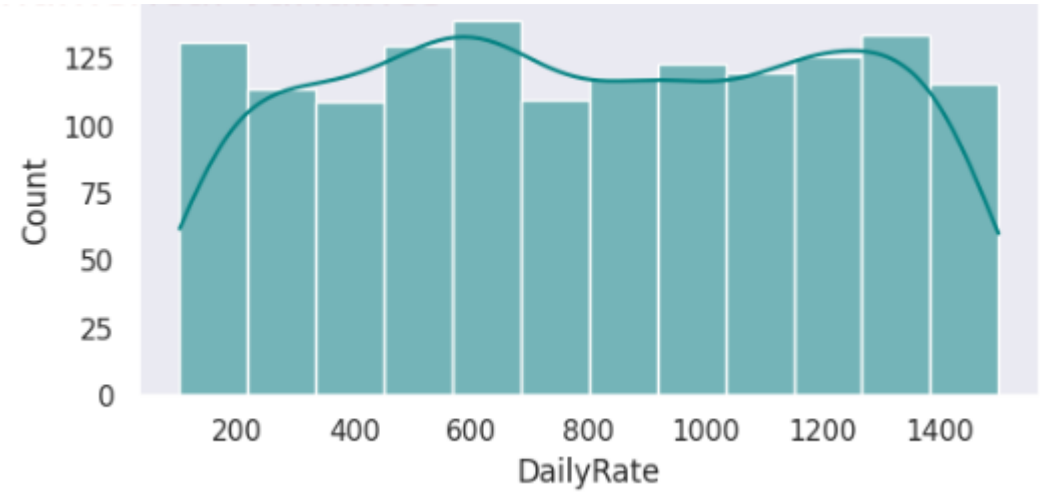

Figure 7: Histogram of DailyRate

3. **DistanceFromHome:**

DistanceFromHome which is also a numerical variable represents the average distance that employees live from their workplace. The table below and the summary statistics, indicates that the mean DistanceFromHome is 9.19. The standard deviation of 8.11 suggests that distances from home vary around the mean by approximately 8.11. Maximum (29 miles): The maximum distance from home is 29 while the minimum distance is 1. The distribution of DistanceFromHome as shown below is left skewed meaning that the distance travelled by most of the employees is below the mean distance of 9.19, hence this variable would be normalized during data preparation.

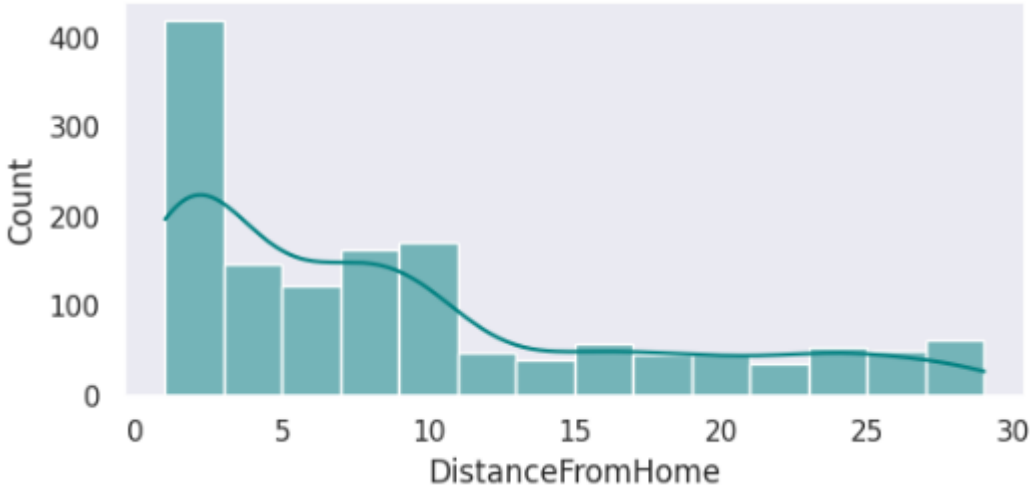

Figure 8: Histogram of DistanceFromHome

4. **MonthlyIncome:**

MonthlyIncome, a numerical variable, represents the monthly income of an employee. It ranges from 1009 to 19,999, indicating variations in earnings based on factors such as job level, experience, and role. However, the distribution of the variable is left-skewed, suggesting that most of the employees earn lower incomes compared to the higher earners. Normalisation will be applied to address this skewness in the distribution.

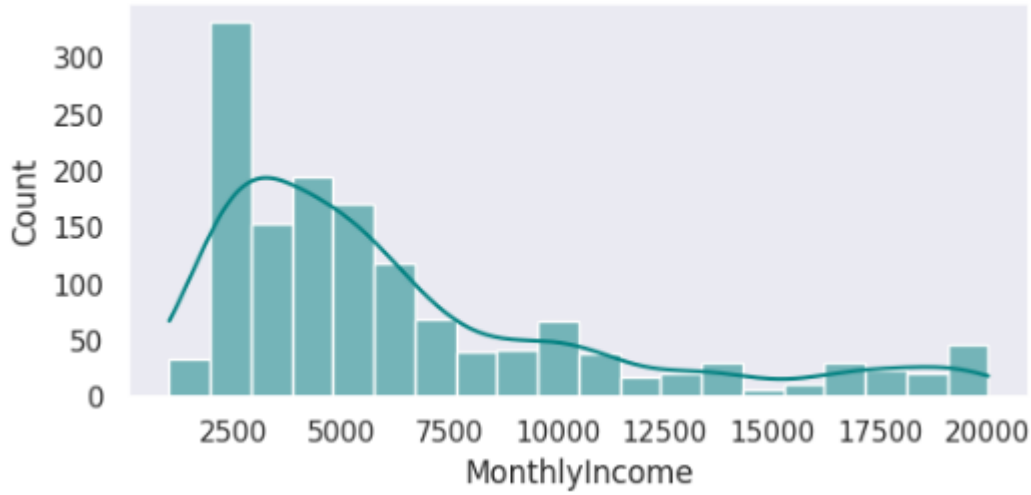

Figure 9: Histogram of MonthlyIncome

5. **Attrition:**

The "Attrition" variable is categorical because its attributes are not numerical; it has unique values "Yes" and "No". Attrition is the target, output, or dependent variable. The analysis shows that 234 employees have left the company, while 1,233 are still employed. These numbers represent 16.1% and 83.9% of the workforce, respectively.

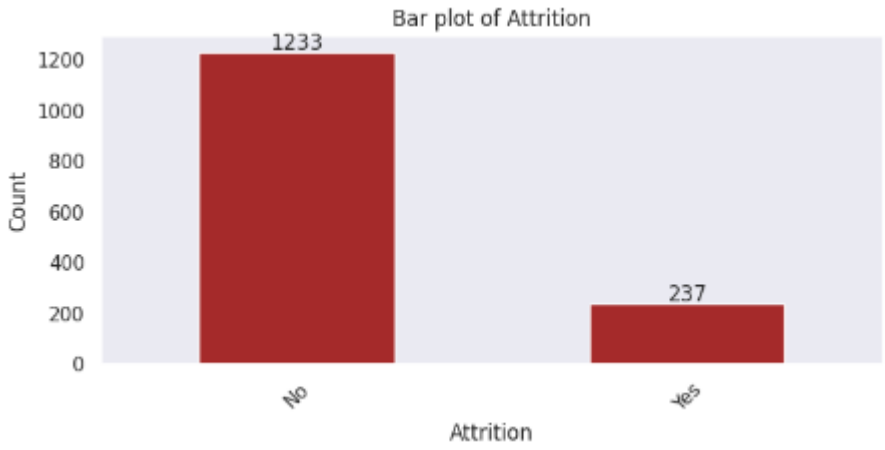

Figure 10: ***Bar plot of Attrition***

6. **Department:**

This is a categorical variable indicating the department where an employee works. There are three unique values: Research & Development, Sales, and Human Resources, with 961, 446, and 63 employees working respectively in these departments.

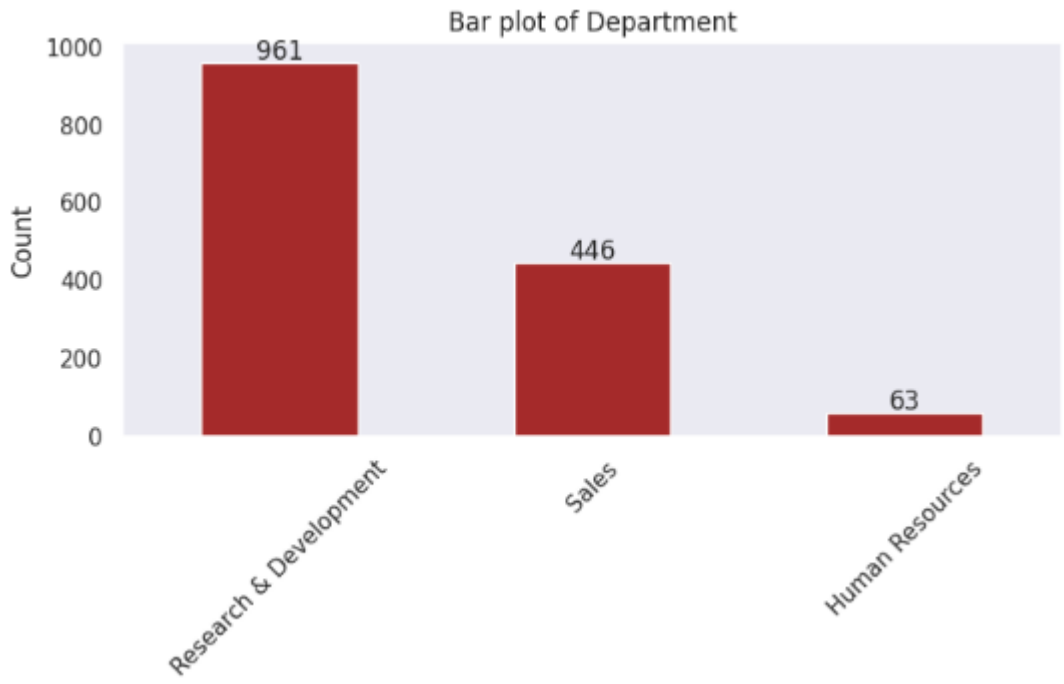

Figure 11: ***Bar Plot of Department***

7. **Education:**

This variable is equally a categorical variable. Although it was a numerical variable, it has been transformed into a categorical variable to achieve data consistency. The variable has five unique values: Below College, College, Bachelors, Masters, and Doctorate with 170, 282, 572, 398, and 48 employees, respectively, having various educational qualifications.

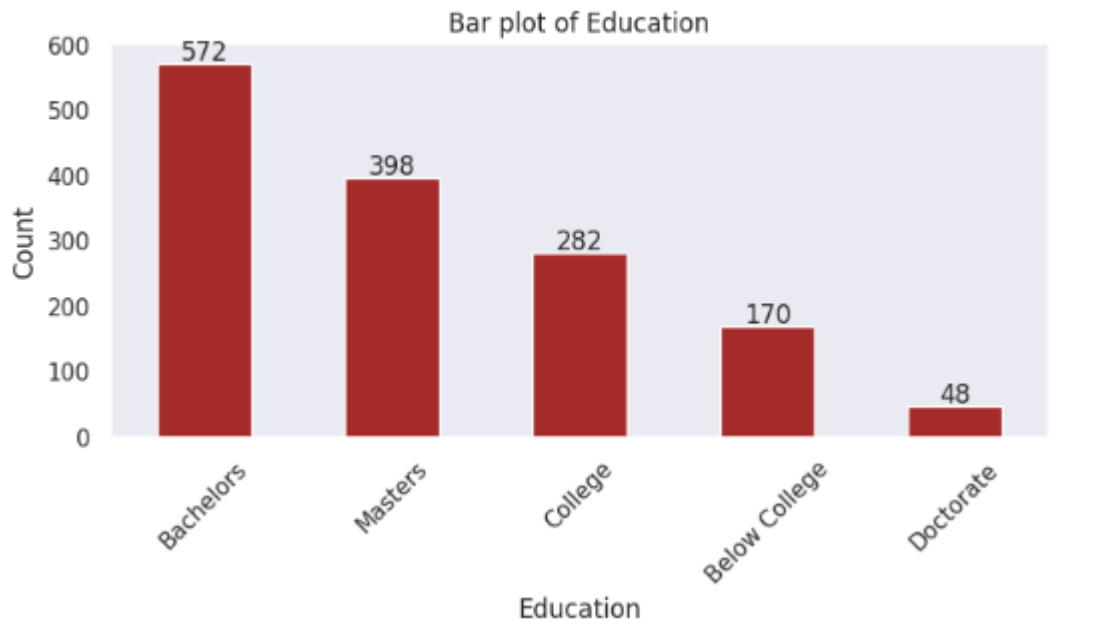

Figure 12: ***Bar Plot of Education***

4.2.2 BIVARIATE ANALYSIS

The analysis involved examining the relationship between the target variable, "Attrition," and other variables to understand how attrition rates vary across different factors. Below is the Visualizations illustrating the distribution.

1. **JobInvolvement vs Attrition**

In the provided figure, it is observed that out of 868 employees with high Job Involvement, 125 of them left the company, while 743 of them remained. Similarly, 71 employees with medium Job Involvement left the company, while 304 remained. Additionally, 13 employees with very high Job Involvement left the company, compared to 131 who stayed. Conversely, 28 employees with low Job Involvement are no longer with the company, while 55 of them stayed. Overall, most of the employees exhibit high Job Involvement, indicating positive engagement levels within the company.

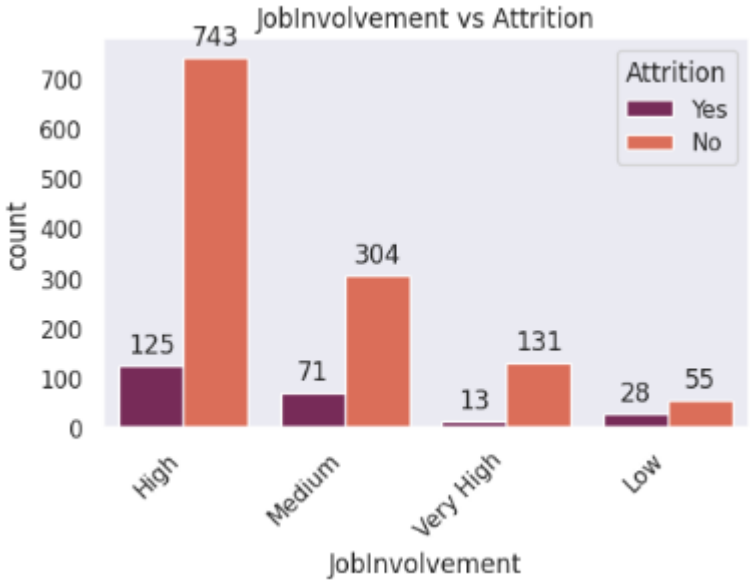

Figure 13 Bar plot Gender vs Attrition

2. WorkLifeBalance vs Attrition

It has been observed that a considerable number of employees are dissatisfied with their work-life balance. Although this discontentment has not led to a significant number of them leaving the company, it could pose a potential attrition risk if not resolved. Specifically, 766 employees are still with the company but are not satisfied with their work-life balance, out of which 127 have left. On the other hand, 286 employees are still working for the company and are very satisfied with their work-life balance, and 58 of them have left. Similarly, 126 employees are still with the company and are satisfied with their work-life balance, while 27 of them have left. Furthermore, 55 employees who are highly satisfied with their work-life balance are still working for the company, compared to 25 who have left. In conclusion, addressing work-life balance issues can help reduce attrition risk and improve overall employee satisfaction and retention.

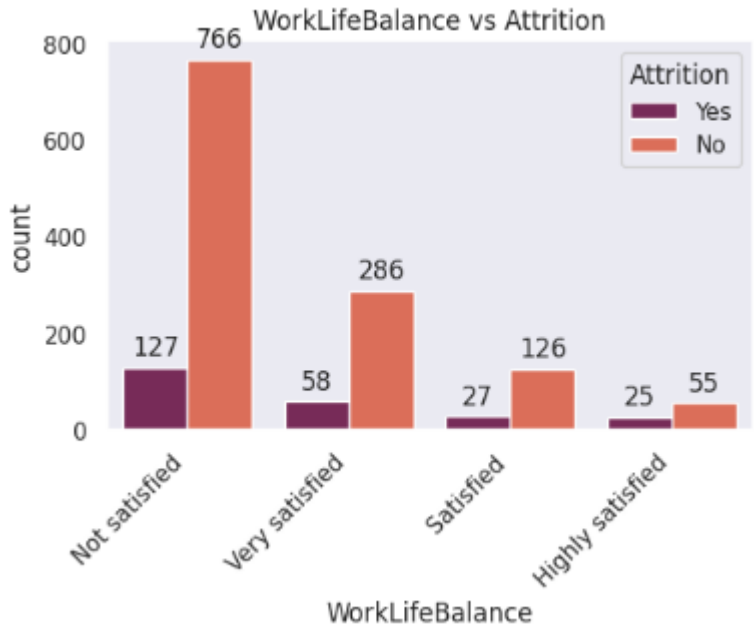

Figure 14: Bar plot WorkLifeBalance vs Attrition

3. PercentageSalaryHike vs Attrition

In the figure below, employees with a lower percentage salary hike, which is lower than 20%, are more in number. Considering the proportion of them that left and those that stayed, compared to that of the employees with higher percentage salary hikes, this factor does not pose an attrition risk.

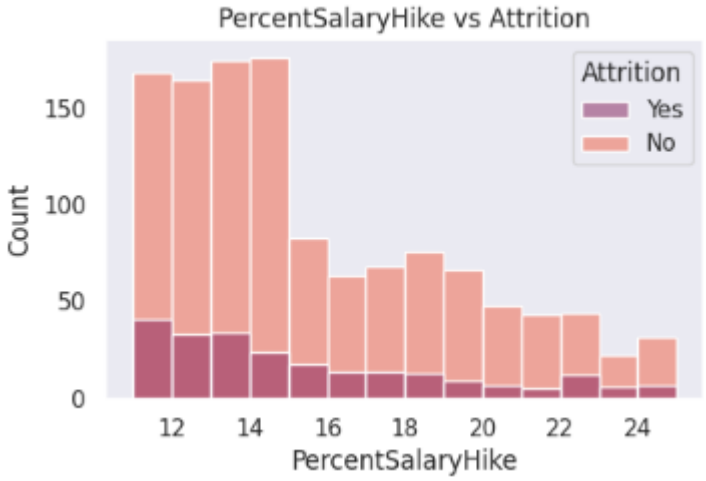

Figure 15: Histogram of PercentageSalaryHike vs Attrition

4.2.3 MULTIVARIATE ANALYSIS

This analysis is intended to evaluate the relationship between three or more variables, as shown below. The analysis is useful in comparing the distribution of categorical and numerical variables.

1. **Attrition vs Gender and YearsSinceLastPromotion**

The figure below demonstrates that among male employees, an equal number of those who were promoted in the last two years either left or stayed with the company. This suggests that the years since the last promotion does not significantly influence the decision of male employees to leave or remain in the company. Conversely, among female employees, those who have not been promoted for a longer period were more likely to stay with the company compared to those who were promoted more recently. Notably, female employees with an extended duration since their last promotion did not leave the company. This observation indicates that the impact of the years since the last promotion on attrition differs between male and female employees.

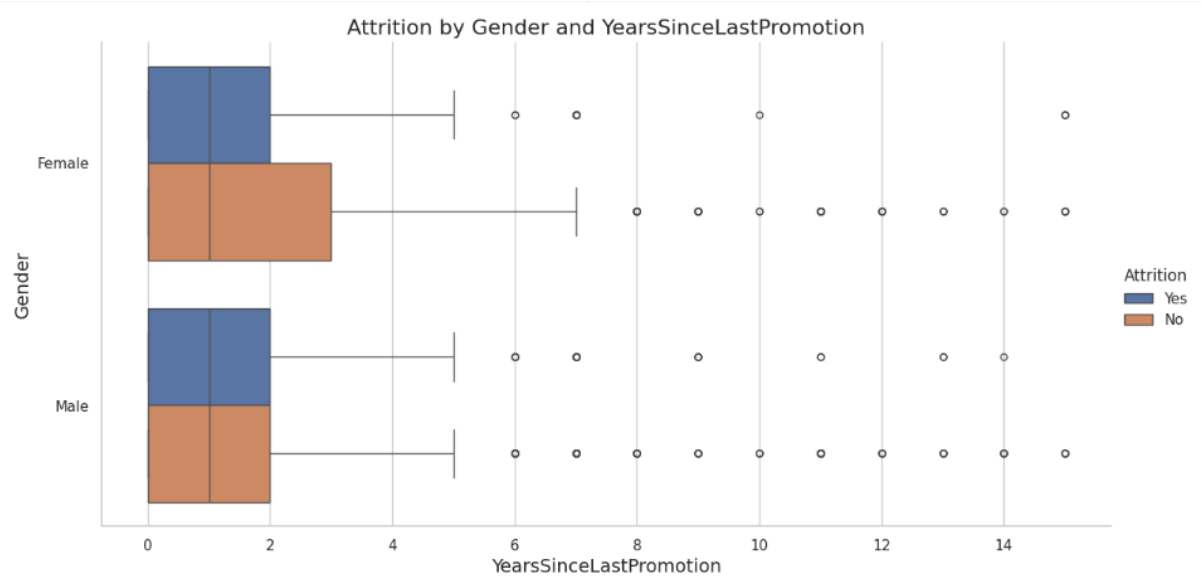

Figure 16 Boxplot of Attrition vs Gender and YearsSinceLastPromotion

2. **Education vs Gender and PerformanceRating**

The figure below illustrates that among female employees, 101 individuals with a college education, 44 with education below college, 132 with a master's degree, and 18 with a doctorate all received a high-performance rating. Similarly, among male employees, 133 individuals with a college education, 96 with education below college, 209 with a master's

degree, 291 with a bachelor's degree, and 21 with a doctorate were also rated high in performance.

Furthermore, among female employees, 16 individuals with a college education and below, 22 with a master's degree, 36 with a bachelor's degree, and 4 with a doctorate, all received an excellent performance rating. Similarly, among male employees, 32 individuals with a college education, 14 with education below college, 35 with a master's degree, 46 with a bachelor's degree, and 5 with a doctorate were also rated excellent in performance.

In summary, fewer male and female employees received excellent performance ratings than those who received high ratings.

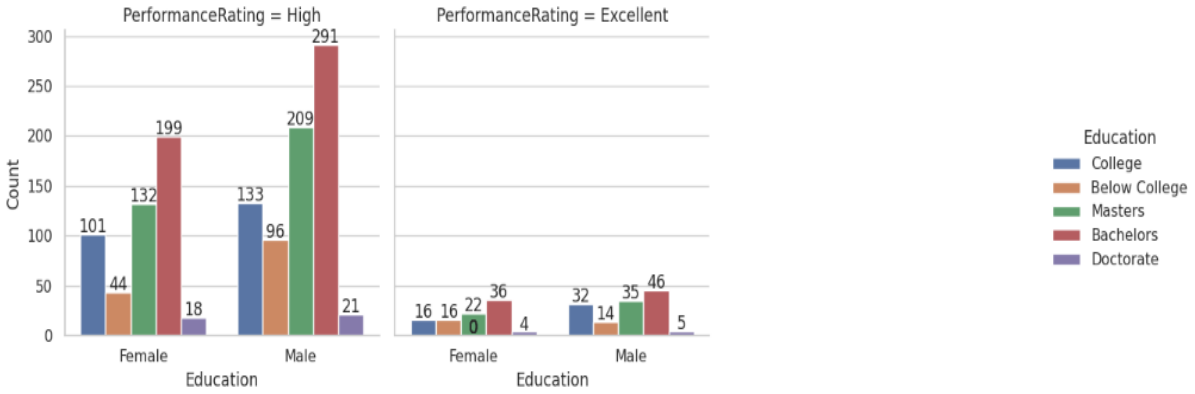

Figure 17: Bar plot of Education vs Gender and PerformanceRating

3. MonthlyIncome vs Age and Attrition

The scatterplot below indicates that most of the employees below the age of 50, who work overtime and earn below 15,000, have left the company. Conversely, most employees below the age of 50 who do not work overtime and earn below 15,000 did not leave the company, contrasting with the previous scenario. This suggests that monthly income and the age of employees predominantly influence their decision to stay or leave, rather than just whether they work overtime.

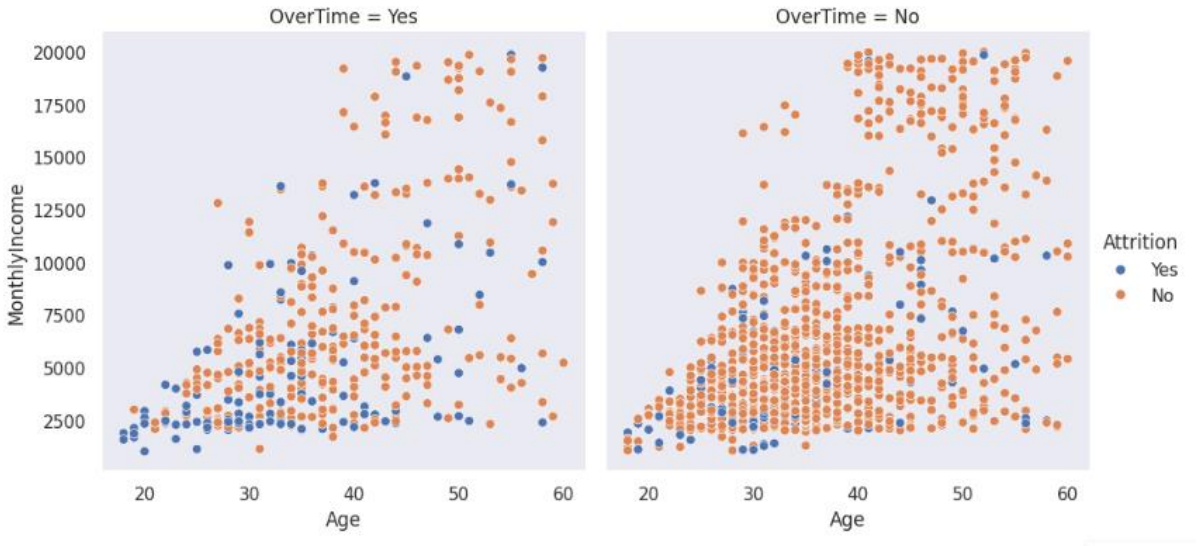

Figure 18: Scatter plot of Monthly income and Age vs Attrition

4. Correlation Analysis

Similarly, Spearman correlation was employed to assess the correlation among variables, aiding in identifying variables lacking correlation with others or themselves. This process assisted in addressing multicollinearity among variables. For instance, it was observed that EmployeeCount and StandardHours lack correlation with other variables or themselves, leading to their removal as they were deemed irrelevant for this study. Additionally, it was noted that YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager exhibit high positive collinearity. However, further examination will be conducted using the ANOVA F-test with Sci-kit Learn's SelectKBest to select the variables for modelling and determine which ones to drop. The correlation matrix heatmap below illustrates the outcomes of these correlations.

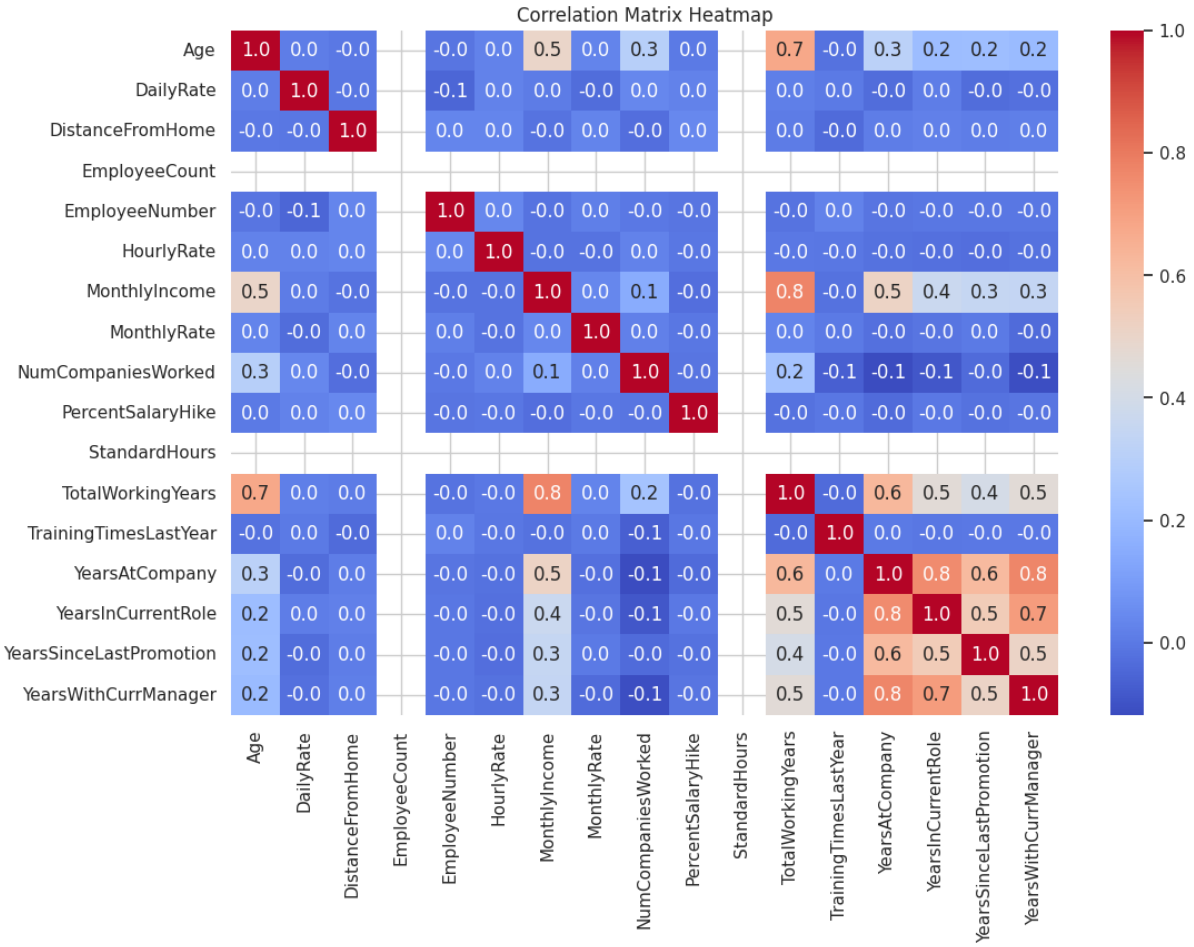

Figure 19: Correlation Matrix Heatmap

5. Analysis of outlier

It is necessary to analyse outliers (data points that are distinct from other data points) in the dataset to understand their nature and assess how they may affect preprocessing and overall model performance. Utilising box plots, it was discovered that outliers are present in several variables, including MonthlyIncome, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager, as depicted below.

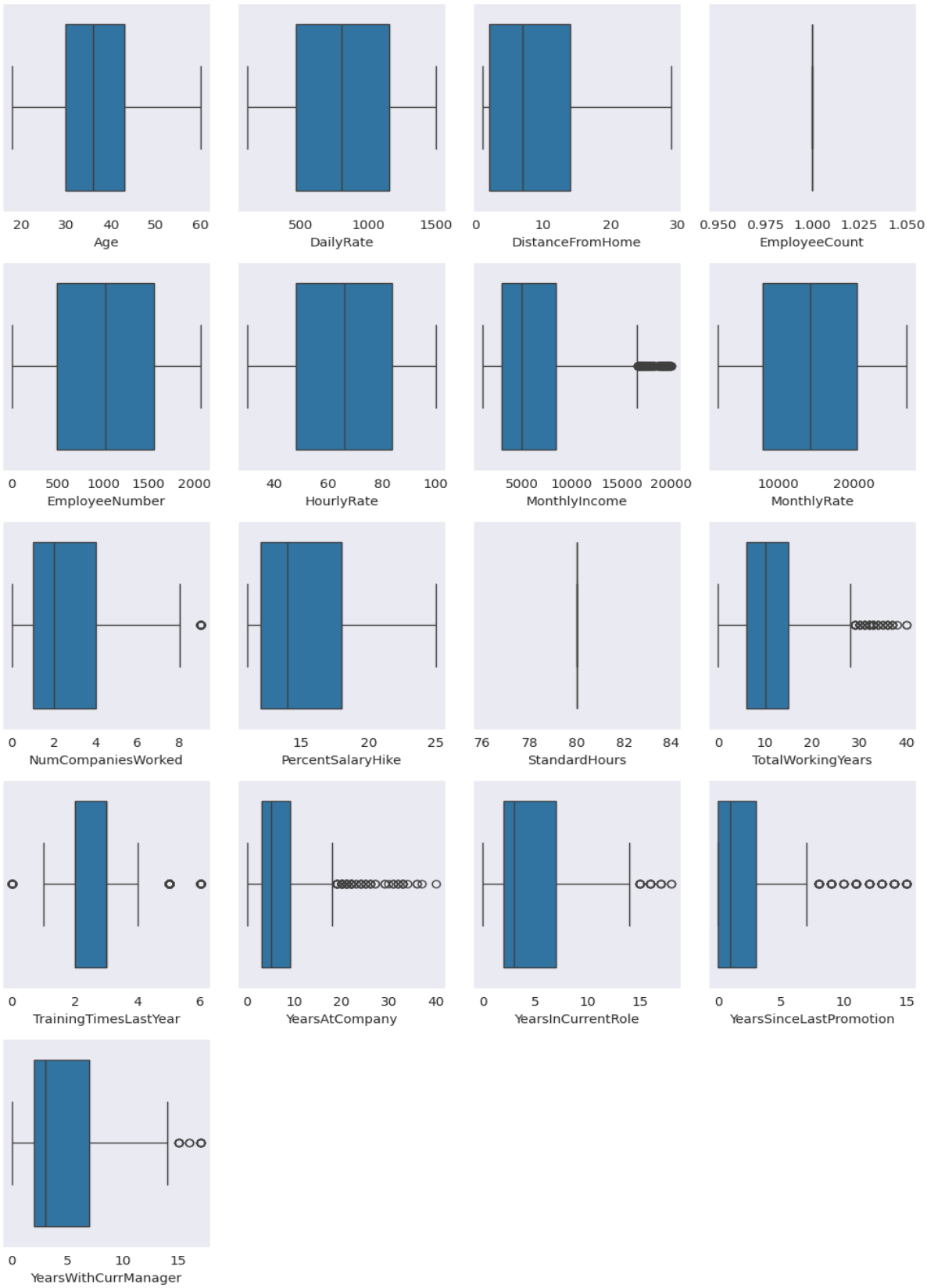

Figure 20: Box plot for numerical variables

The outliers were further scrutinized using scatterplots to verify that they were not erroneous. It was noted that both classes of attrition were evident at the outlier data points, affirming their validity and suggesting that the affected variables should be retained. Specifically, examining MonthlyIncome, the scatter plot below demonstrates that employees earning above 17,500 were both those who left the company and those who stayed, confirming that the outliers are not attributable to errors or data quality issues.

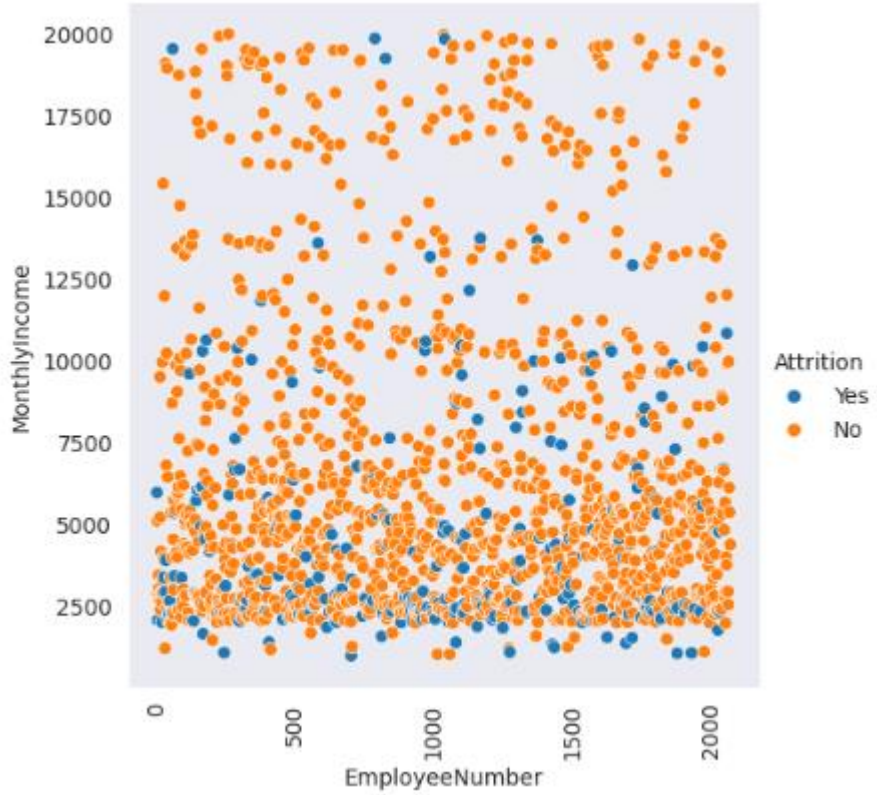

Figure 21: Scatter plot for Attrition vs MonthlyIncome and EmployeeNumber

4.3 DATA PREPARATION

In preparing the data for modelling, various preprocessing steps were undertaken. These activities encompassed handling inconsistent data, encoding categorical variables, normalisation of numerical features, feature selection, data sampling, and data splitting. As emphasised by Selvaraj and Nagarajan (2022), data preprocessing is paramount for achieving greater accuracy in modelling. It's worth noting that some preprocessing steps, such as

addressing missing or duplicate values, were not necessary for this dataset as it was devoid of such issues. Therefore, they were not described here.

4.3.1 HANDLING INCONSISTENT DATA

Upon reviewing the dataset, it became apparent that certain variables, by their inherent nature, should be categorical but were represented as numerical. These variables include WorkLifeBalance, StockOptionLevel, RelationshipSatisfaction, PerformanceRating, JobSatisfaction, JobInvolvement, EnvironmentSatisfaction, Education, and JobLevel. Notably, the minimum and maximum values of these variables range from 0 to 5, indicating the supposed weight of their unique values.

To address this inconsistency, a Python mapping function was employed to transform these numerical variables into their appropriate categorical data types. Alkhateeb (2023) also utilized a similar approach to ensure data consistency by reverting numerical variables to their original categorical representations. However, it's noteworthy that some other researchers retained these variables in their numerical form in their models.

For instance, the Education variable initially appeared as a numerical variable, with a range of values from 1 to 5. Yet, upon closer examination, it became evident that Education should denote distinct levels of education or the educational status of an employee, rather than numerical values. Hence, Education was transformed into categorical variables with unique values indicative of educational levels or attainment, as elaborated below.

```
The unique values in 'Education' variable are:

College
Below College
Masters
Bachelors
Doctorate
```

Figure 22 Unique values in Education variable

4.3.2 NORMALIZATION

Data normalisation was deemed essential, aligning with the perspective of Mansor et al. (2021), who emphasised its role in mitigating the dependency on the choice of measurement units for variables. Upon visualising the distribution of numerical values, certain variables

exhibited skewness, while others, such as Age, displayed a normal distribution. However, variables like MonthlyIncome and DistanceFromHome were left-skewed.

To address this, the researcher utilised Yeo-Johnson's Power Transformer function in Python to transform these non-normally distributed variables into Gaussian-like or normal distributions. Additionally, the Sci-kit Learn StandardScaler() function was used to scale the data uniformly. While Wadikar (2020) employed the MixMaxScaler() function for standardisation, the power transformer function was not utilised to normalise skewed variables. Conversely, Alkhateeb (2023) adopted both the PowerTransformer (method='yeo-Johnson') function and the StandardScaler() function, like this research.

4.3.3 ENCODING

Three encoding techniques were employed: one-hot encoding, label encoding, and ordinal encoding. One-hot encoding transformed variables such as Gender and OverTime. This technique increases the number of variables by splitting them into their unique values, creating binary variables for each category. Label encoding was applied to the target variable (Attrition), Department, EducationField, JobRole, and MaritalStatus. Label encoding assigns a unique numerical label to each category within a variable. Ordinal encoding transformed variables such as EnvironmentSatisfaction, Education, JobInvolvement, BusinessTravel, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, JobLevel, StockOptionLevel, and PerformanceRating. This technique encodes categorical features as ordinal, preserving the inherent order or hierarchy among the categories.

The rationale behind the selection of each type of encoding for the corresponding variables aligns with the principles outlined by the Scikit-learn machine learning library. Ordinal encoding is suitable for variables where there is a clear order or hierarchy among the categories. One-hot encoding is ideal for variables with a small number of unique values, as it assigns a separate feature for each category. Label encoding is typically used for labelling target variables and for variables with a higher number of unique values.

4.3.4 DATA SAMPLING

The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to tackle the class imbalance observed in the target variable, attrition. Initially, the dataset exhibited an imbalanced distribution, with 237 instances of attrition (represented by "Yes") and 1233 instances of non-attrition (represented by "No"). This significant disparity between the two classes could affect the model's ability to predict both outcomes accurately.

To address this issue and improve model performance, SMOTE was applied to oversample the minority class (Attrition_Yes) by generating synthetic samples.

Before applying SMOTE, the proportion of Attrition_Yes and Attrition_No classes in the dataset was visualised to highlight the class imbalance. This step provided a clear understanding of the distribution of the target variable and informed the decision to employ SMOTE to rectify the imbalance.

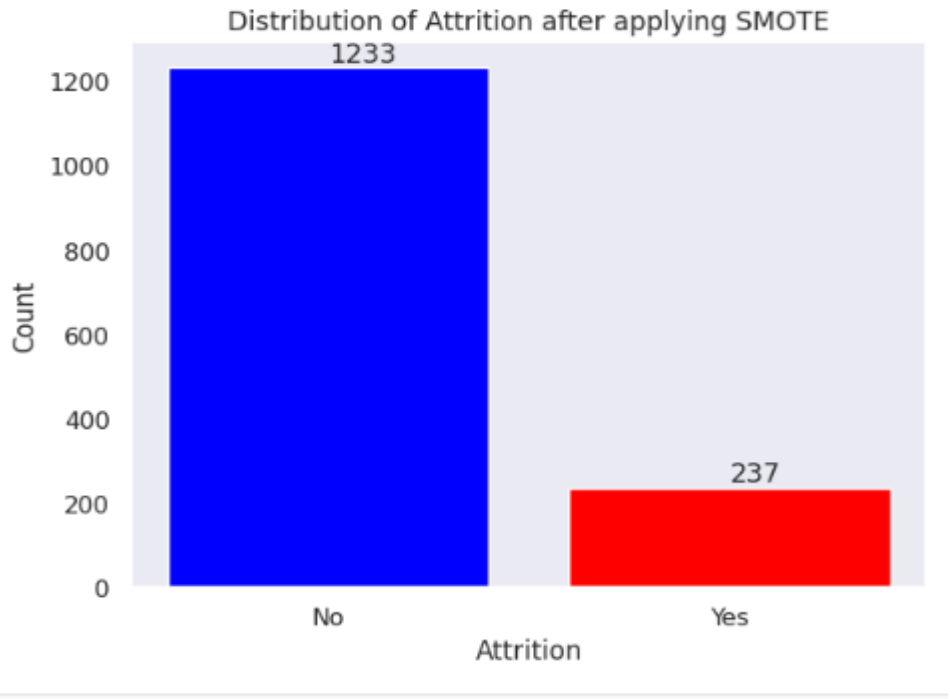

Figure 23: Distribution of Attrition before applying SMOTE

After applying SMOTE, the attrition classes were harmonized to 1233. The distribution after the harmonization is depicted in the following figure.

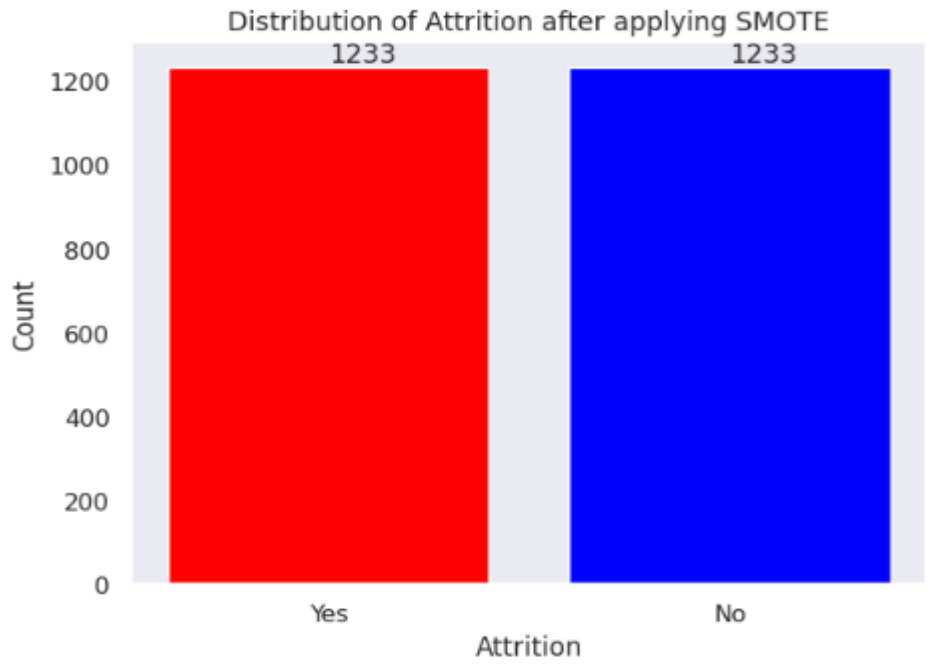

Figure 24: Distribution of Attrition after applying SMOTE

4.3.5 FEATURE SELECTION

Leveraging the ANOVA F-test in combination with the Sci-kit Learn SelectKBest() function to identify significant numerical variables, this approach aimed to rank the numerical variables based on their importance in predicting attrition.

The results of this analysis are depicted in the figure below, showcasing the scores assigned to each numerical variable. These scores provide insights into the relevance and predictive power of each variable regarding attrition. The analysis reveals that the six most important numerical variables for predicting attrition are YearsInCurrentRole, TotalWorkingYears, MonthlyIncome, YearsWithCurrManager, Age, and YearsAtCompany. These variables exhibit the highest significance and predictive power in determining attrition likelihood.

	Feature	Score
0	YearsInCurrentRole	40.864694
1	TotalWorkingYears	39.130242
2	MonthlyIncome	32.721340
3	YearsWithCurrManager	31.894762
4	Age	31.104330
5	YearsAtCompany	23.703454
6	NumCompaniesWorked	4.681376
7	DistanceFromHome	4.380866
8	TrainingTimesLastYear	3.629371
9	YearsSinceLastPromotion	2.865066
10	DailyRate	2.157048
11	MonthlyRate	0.133992
12	PercentSalaryHike	0.051197
13	HourlyRate	0.038903

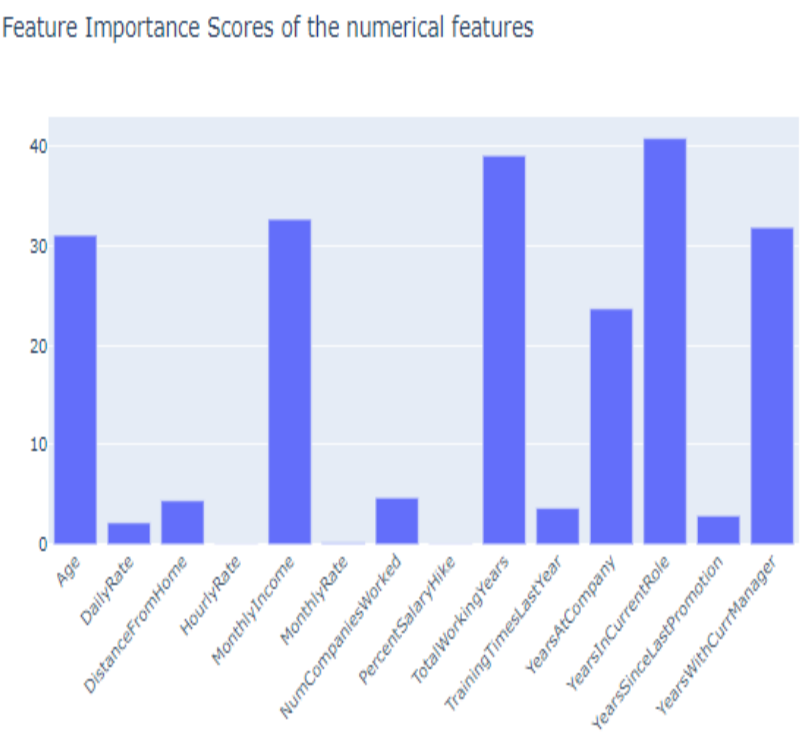

Figure 25: Feature Important scores of numerical variables

Furthermore, the Chi-Squared test and the Sci-kit Learn SelectKBest() were employed to identify the topmost important categorical variables. The results, illustrated below, highlight OverTime_Yes, JobLevel, StockOptionLevel, OverTime_No, and Marital Status as the top five significant categorical features contributing to attrition prediction.

	Feature	Score
0	OverTime_Yes	63.845067
1	JobLevel	58.268745
2	StockOptionLevel	25.268826
3	OverTime_No	25.198812
4	MaritalStatus	18.745657
5	JobRole	9.004448
6	JobSatisfaction	6.692900
7	Education	3.108840
8	WorkLifeBalance	2.442267
9	RelationshipSatisfaction	1.695992
10	Department	1.329297
11	EnvironmentSatisfaction	0.867333
12	EducationField	0.834912
13	Gender_Female	0.765130
14	Gender_Male	0.510087
15	JobInvolvement	0.149139
16	PerformanceRating	0.001886
17	BusinessTravel	0.000002

Table 4: Feature importance scores of categorical variables

The feature importance and selection process were reinforced by conducting a test for correlation between the input and target variables for modelling purposes. As depicted below, the significant features identified through the previously described methods emerged as the most important features in this correlation analysis. This observation suggests a strong indication of not just correlation but also potential causation between these features and the target variable. This finding partially aligns with Repaso et al. (2022) and Pasquarella (2023), which emphasised the importance of marital status and total working years as predictors of attrition.

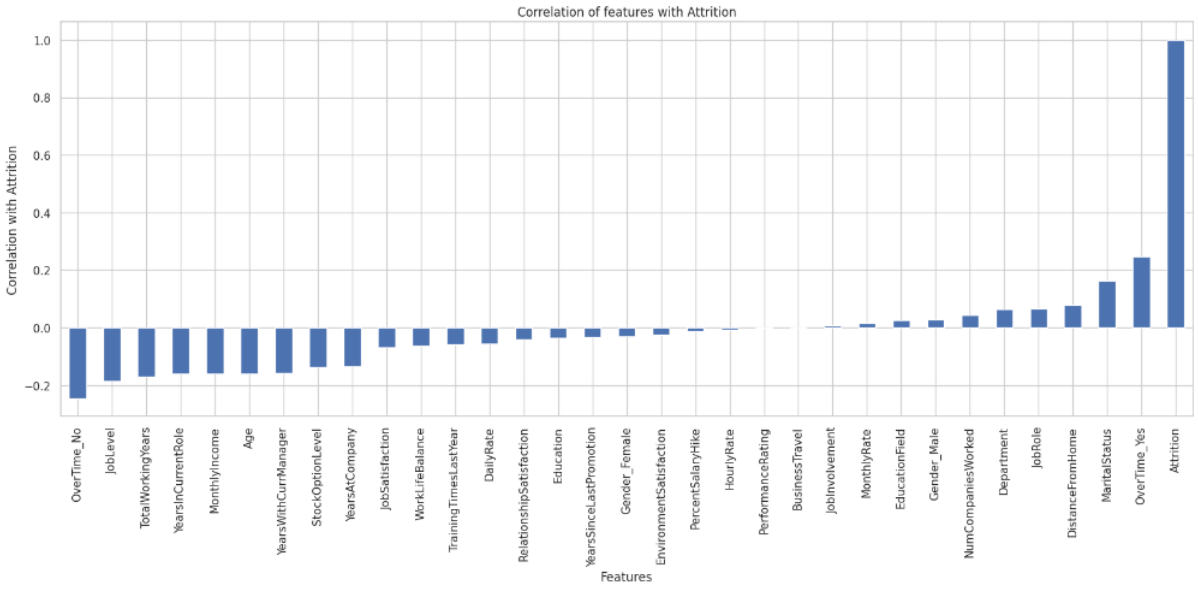

Figure 26: Correlation of input features with Attrition

Feature selection was also influenced by a test for multicollinearity among numerical variables, as indicated by the correlation matrix heatmap described earlier. Variables such as YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager exhibited multicollinearity, which guided the decision-making process. Based on the above analysis, variables such as Over18, EmployeeNumber, EmployeeCount, and StandardHours were dropped to improve model performance. These variables are considered irrelevant for model building, as they are unlikely to influence an employee's decision to stay or leave their job. For example, EmployeeNumber, which is synonymous with staff number, is not likely to impact attrition decisions.

However, no other variable was dropped. Dropping additional variables could result in the loss of important information. Nevertheless, if more variables were to be dropped,

YearsSinceLastPromotion would be a suitable candidate due to its low feature importance score and high multicollinearity with other variables such as YearsAtCompany, YearsInCurrentRole, and YearsWithCurrManager. Similarly, variables like BusinessTravel, PerformanceRating, PercentSalaryHike, and HourlyRate could be considered for removal due to their low feature importance scores, being among the least important variables in both categorical and numerical features.

4.4 MODEL RESULTS AND EVALUATION

The research explored four machine learning models: Random Forest Classifier, Logistic Regression, Support Vector Machine, and Extra Tree Classifier. The primary focus was on their efficacy in accurately predicting the minority and majority classes of the target variable, Attrition (represented by Attrition_Yes and Attrition_No, corresponding to 1 and 0).

These models were built on a 75/25 data split, with all variables included except for Over18, EmployeeNumber, EmployeeCount, and StandardHours, which were deemed irrelevant for modelling purposes. Below are the results of the models:

4.4.1 RANDOM FOREST CLASSIFIER

The random forest model achieved an accuracy of 93.35%, indicating the overall correctness of the model's predictions on the test dataset. For the "Attrition_No" class (represented as '0'), the precision is 0.89, indicating that out of all instances predicted as "Attrition_No" class, 89% were correctly classified. For the "Attrition_Yes" class (represented as '1'), the precision is 0.98, indicating that out of all instances predicted as "Attrition_Yes" class, 98% were correctly classified. Overall, the achieved precision score indicates fewer false positive predictions.

For the "Attrition_No" class, the recall is 0.98, indicating that the model correctly identified 98% of all actual "Attrition_No" class instances. For the "Attrition_Yes" class, the recall is 0.88, indicating that the model correctly identified 88% of all actual "Attrition_Yes" class instances. These recall values indicate fewer false negative predictions.

The F1 score for the "Attrition_No" class is 0.94, and for the "Attrition_Yes" class, it's 0.93, indicating that the model performed well on both precision and recall.

The performance of the model in predicting both classes of attrition is demonstrated further below.

Test Accuracy: 93.35%

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.98	0.94	311
1	0.98	0.88	0.93	306
accuracy			0.93	617
macro avg	0.94	0.93	0.93	617
weighted avg	0.94	0.93	0.93	617

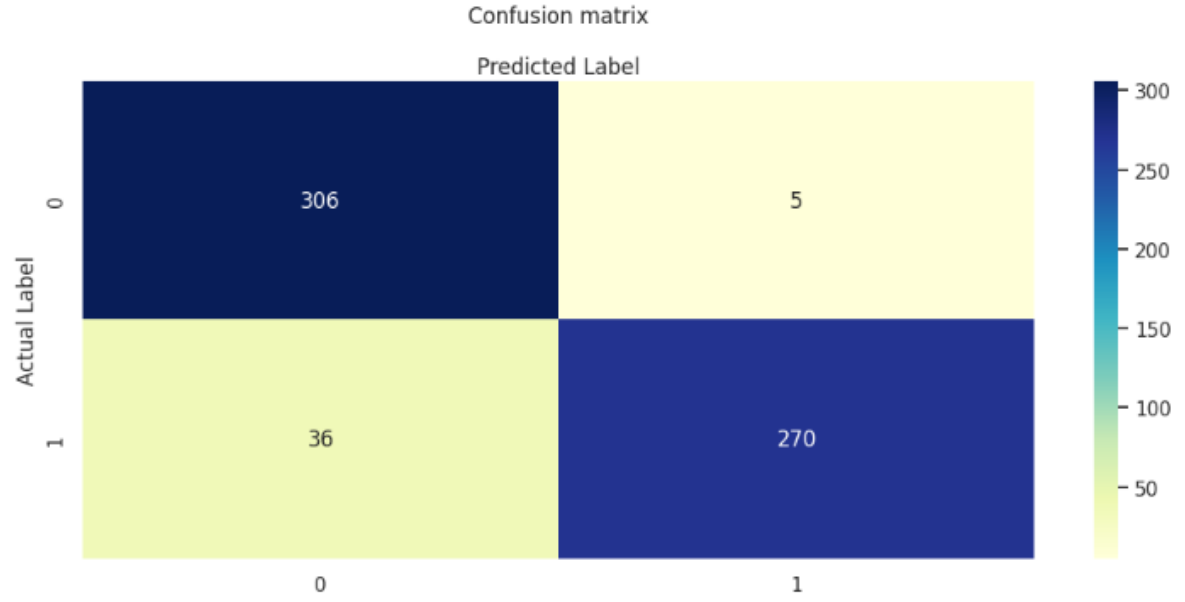

Figure 27: Random Forest Classifier model result

4.5.2 LOGISTIC REGRESSION MODEL

The logistic regression model achieved a test accuracy of 86.55%, indicating the ability of the model to correctly make predictions on the test dataset. For the 'Attrition_No' class (represented as '0'), the precision is 0.83, indicating that out of all instances predicted as 'Attrition_No', 83% were correctly classified. For the 'Attrition_Yes' class (represented as '1'), the precision is 0.91, indicating that out of all instances predicted as 'Attrition_Yes', 91% were correctly classified. Generally, the achieved precision scores indicate fewer false positive predictions.

Similarly, the recall for the 'Attrition_No' class is 0.92, indicating that the model correctly identified 92% of all actual 'Attrition_No' class instances. For the 'Attrition_Yes' class, the recall is 0.81, indicating that the model correctly identified 81% of all actual 'Attrition_Yes' class instances.

The F1 score for the 'Attrition_No' class is 0.87, and that for the 'Attrition_Yes' class is 0.86, indicating that the model performed reasonably well on both precision and recall. The weighted average F1-score is 0.87, suggesting a good balance between precision and recall across both classes.

The model demonstrates satisfactory performance in predicting both classes of attrition as depicted below.

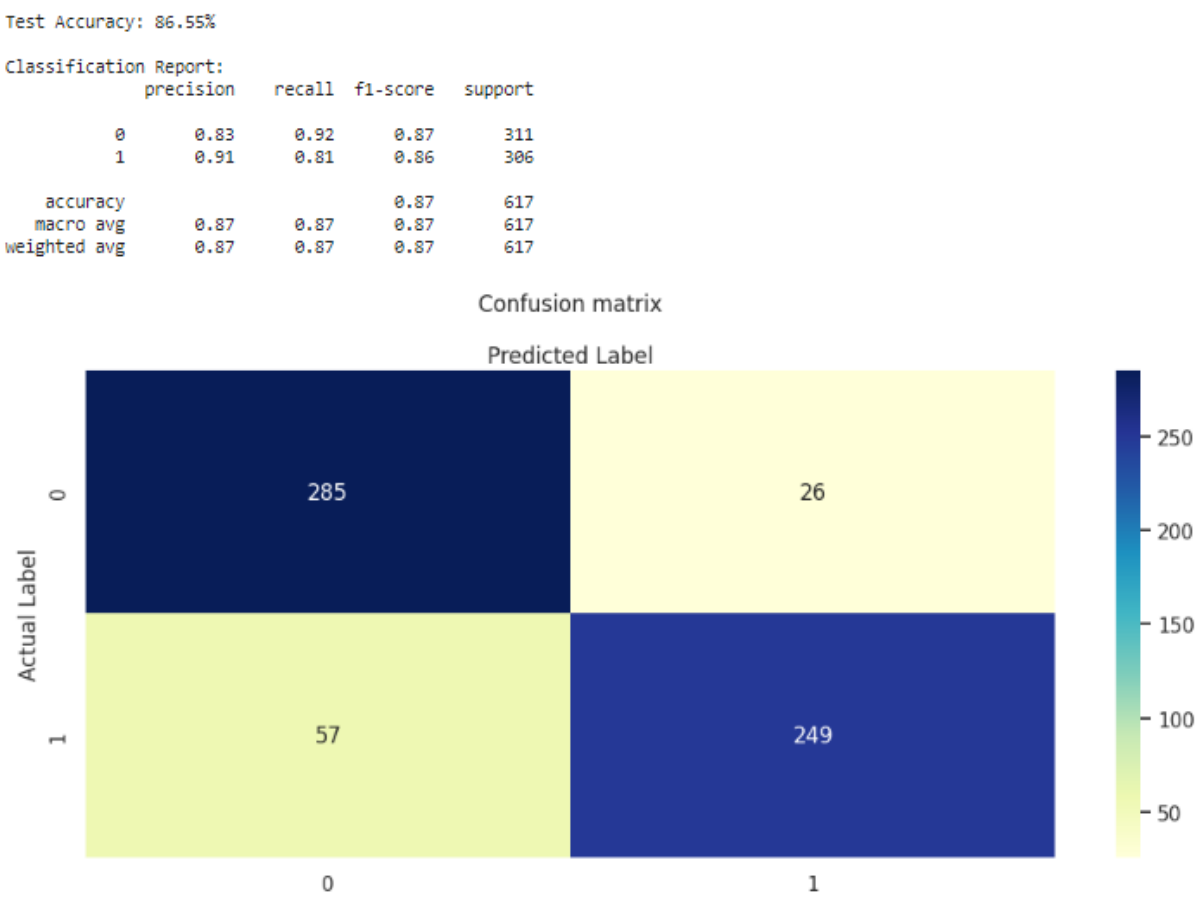

Figure 28: Logistic Regression model result

4.5.3 SUPPORT VECTOR MACHINE MODEL

The classification report for Support Vector Machine model reveals that the model performs well in distinguishing between instances of attrition and non-attrition, achieving an accuracy of 91.25%. When the model predicts an instance as 'Attrition_No', it is correct 88% of the time, with a recall rate of 96%, suggesting it effectively identifies the majority of actual 'Attrition_No' cases. Conversely, when predicting 'Attrition_Yes', it demonstrates a higher precision of 96% but a slightly lower recall of 86%. In summary, the model shows strong precision and recall

scores for both classes at 92% for Attrition_No and 91% for Attrition_Yes, suggesting its capability in correctly classifying instances of attrition and non-attrition, with a notable emphasis on minimizing false positives for 'Attrition_No'.

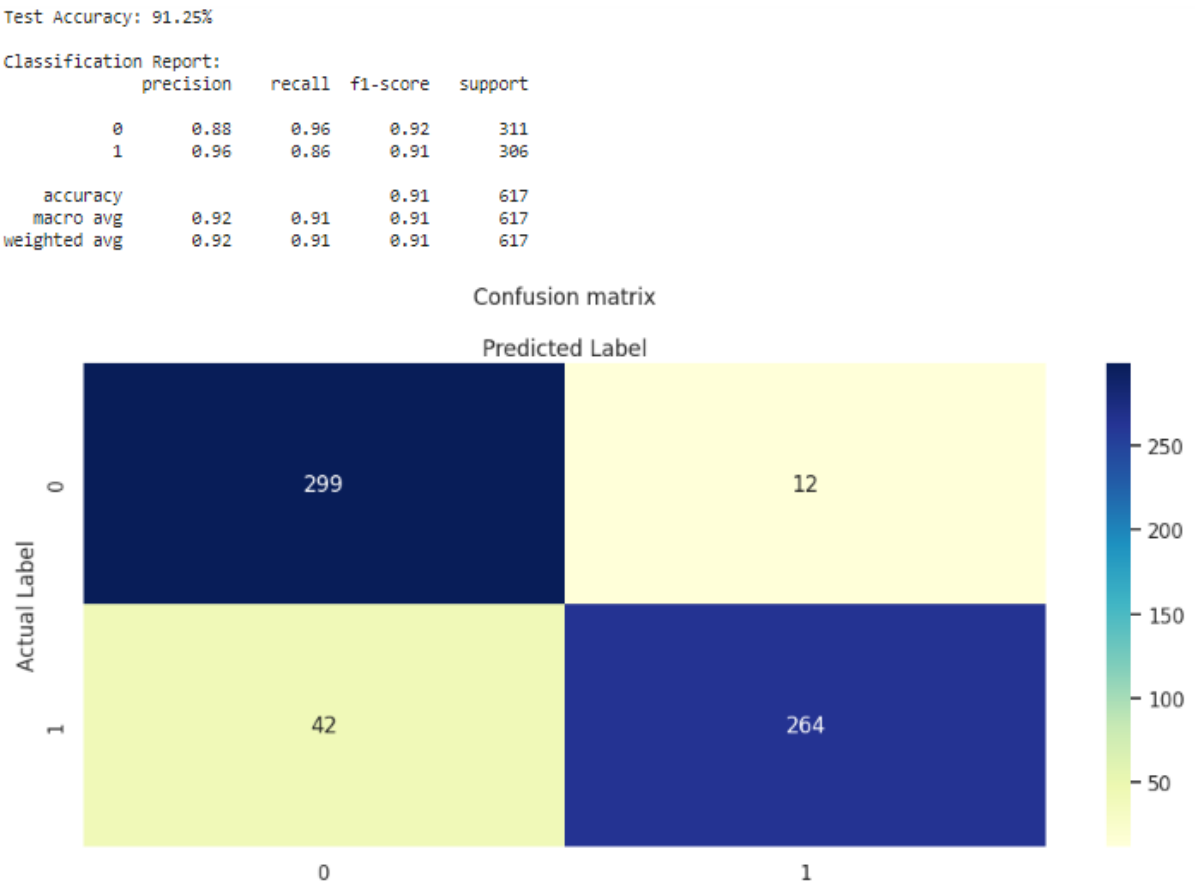

Figure 29: Support Vector Machine model result

4.5.4 EXTRA TREE CLASSIFIER MODEL

Among the models utilized, the Extra Trees Classifier demonstrated the highest test accuracy of 95.14%, indicating strong performance in accurately predicting employee attrition. With precision scores of 0.93 for 'Attrition_No' and 0.97 for 'Attrition_Yes', the model exhibits few false positive predictions, particularly for instances of 'Attrition_Yes'. Additionally, with recall values of 0.97 for 'Attrition_No' and 0.93 for 'Attrition_Yes', the model effectively captures a high proportion of actual instances for both classes.

The balanced F1-scores for both classes, standing at 0.95, further underscore the model's strong balance between precision and recall.

Test Accuracy: 95.14%

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.97	0.95	311
1	0.97	0.93	0.95	306
accuracy			0.95	617
macro avg	0.95	0.95	0.95	617
weighted avg	0.95	0.95	0.95	617

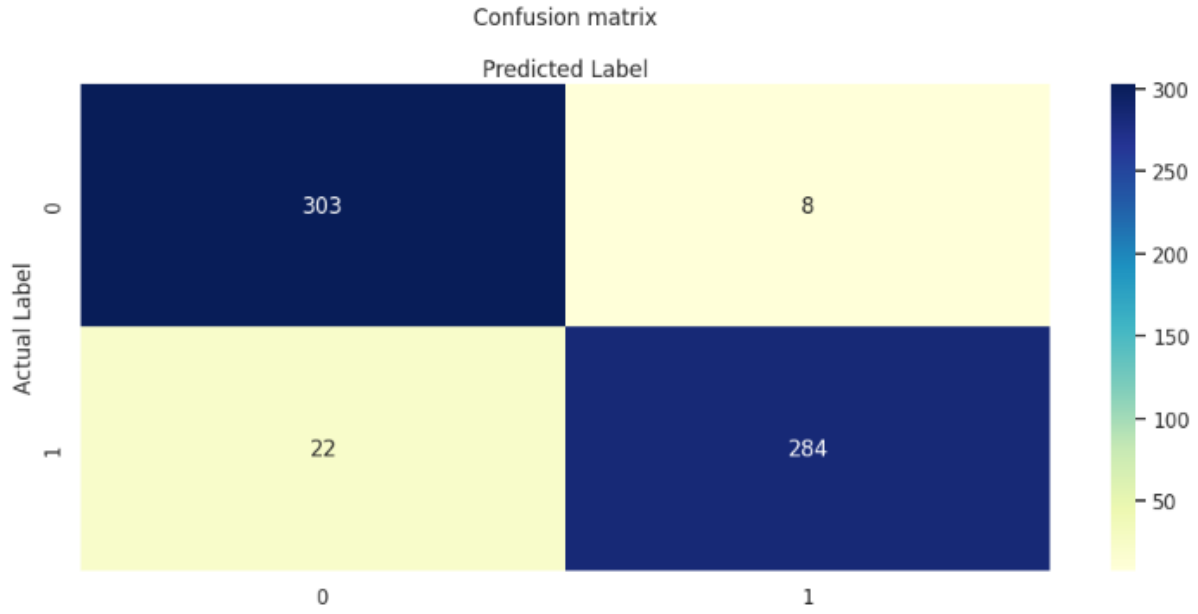

Figure 30: Extra Tree Classifier model result

In comparing the performance of the various models utilized, it is evident that the Extra Tree Classifier outperformed the other models, achieving an accuracy score of 95.14%. Following closely is the Random Forest Classifier with an accuracy score of 93.35%. The Support Vector Machine model achieved an accuracy score of 91.25%, while the Logistic Regression Classifier yielded an accuracy score of 86.55%. These results, depicted in the figure below, highlight the respective performance of each model in accurately predicting employee attrition. This observation is consistent with previous research by Pasquarella (2023), Ali et al. (2022), and Raza et al. (2022), reaffirming Extra Tree Classifier as the top-performing model.

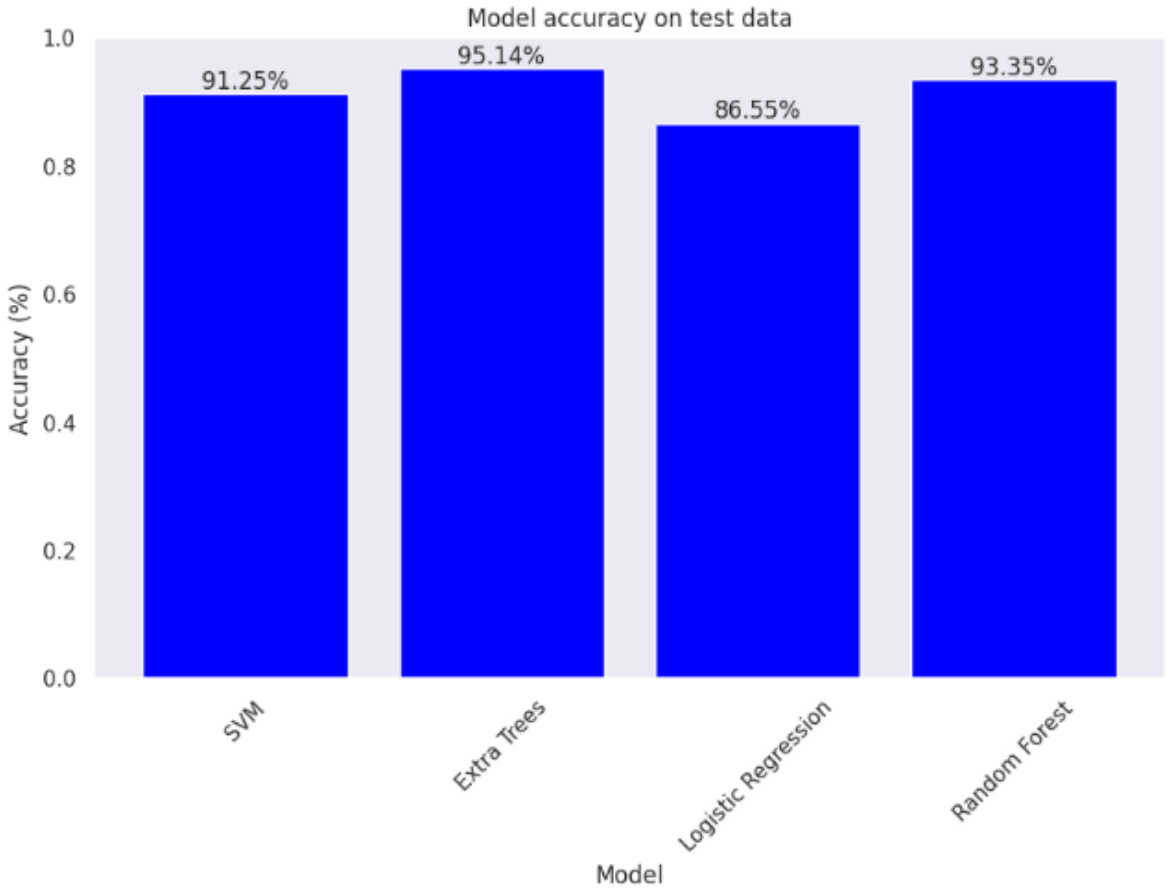

Figure 31: Model accuracy on test data

4.5.5 K-FOLD CROSS-VALIDATION

In addition to the previously applied evaluation methods (accuracy, precision, recall, and F1-score), this study also assessed the performance of the models using k-fold cross-validation. This method ensures that the models generalise well to unseen data. The mean accuracy scores obtained through k-fold cross-validation for each model are as follows: SVM (Support Vector Machine): 90.91%; Extra Tree Classifier: 93.13%; Logistic Regression: 86.69%. These scores closely align with the corresponding predicted accuracy scores of 91.25%, 95.14%, and 86.55% for SVM, Extra Tree Classifier, and Logistic Regression, respectively. This consistency validates the robustness of the model's performance.

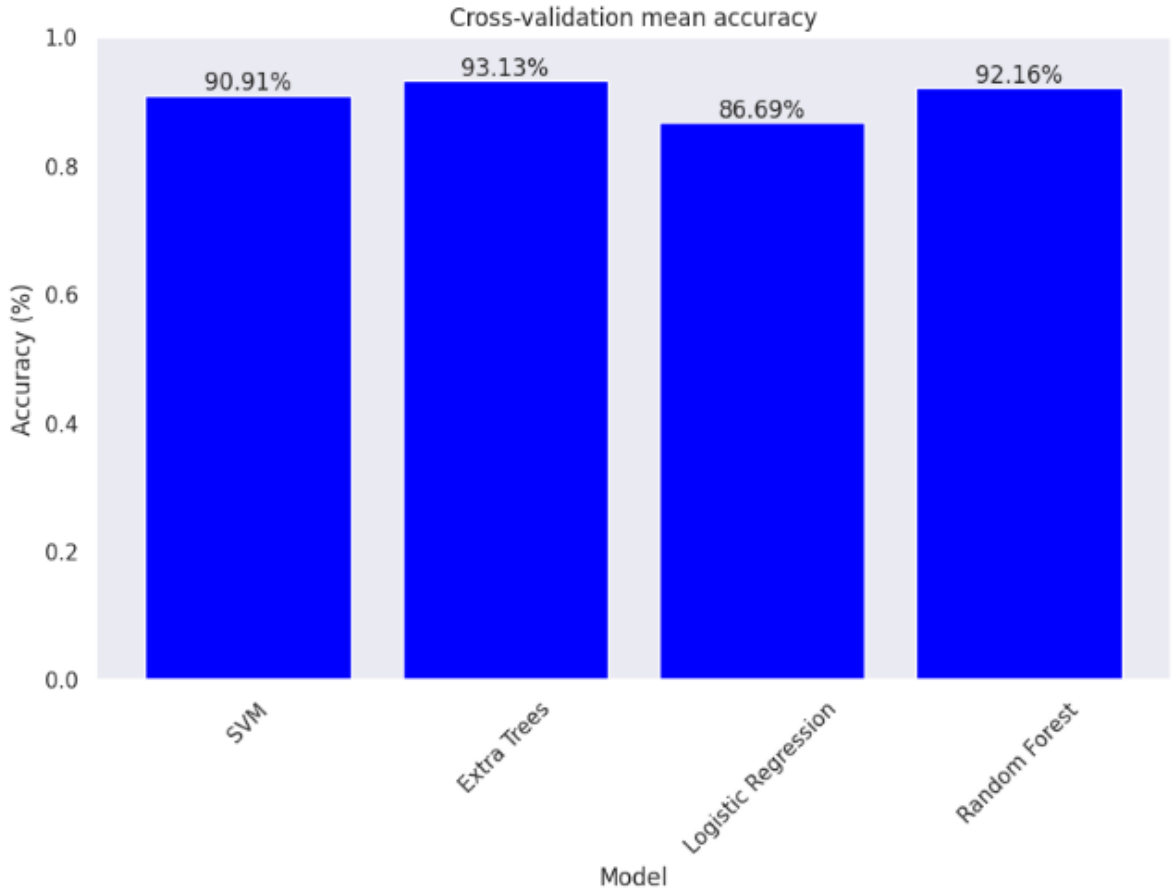

Figure 32 K-fold cross-validation mean accuracy score

4.5.6 EVALUATION OF MODEL WITH IMBALANCED DATA

The models were built using both the original dataset and the dataset augmented with the Synthetic Minority Over-sampling Technique (SMOTE). A comparison of the accuracy scores illustrates that before the application of SMOTE, the accuracy scores were lower, indicating the impact of class imbalance on model performance. The introduction of SMOTE addressed these issues associated with class imbalance, such as biased models favouring the majority class or insufficient generalisation to the minority class. This enhancement is evident in the higher accuracy scores observed after the implementation of SMOTE.

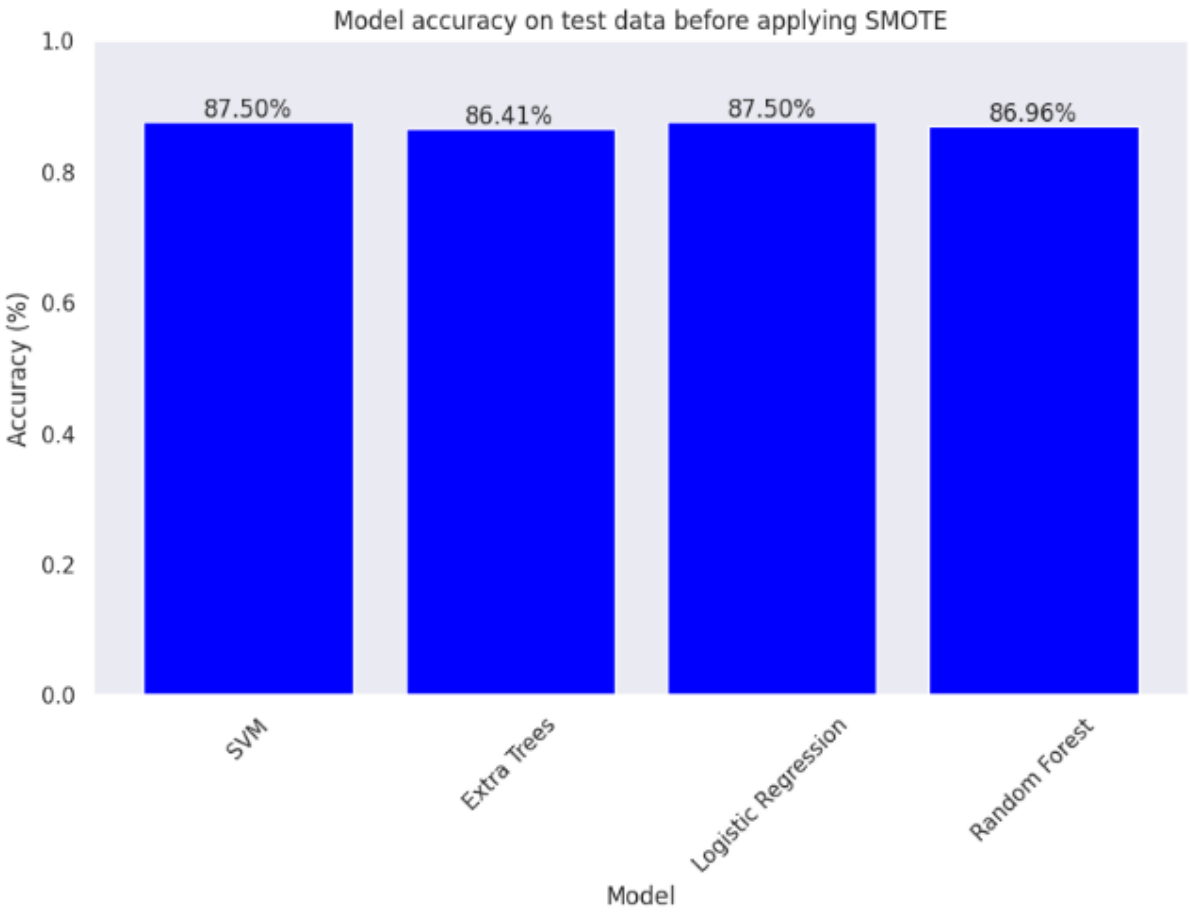

Figure 33: Model accuracy on test data before applying SMOTE

CHAPTER FIVE – CONCLUSION

5.1 DISCUSSION OF FINDINGS

This research successfully achieved the objectives outlined in Chapter One, which included developing machine learning models for predicting employee attrition, evaluating these models to recommend the best-performing one, and analysing data variables to determine their importance as predictors of attrition. With the CRISP-DM research design, the study generated the results discussed below.

5.1.1 PERFORMANCE OF MACHINE LEARNING

Machine learning has demonstrated promising capabilities in predicting employee attrition. Leveraging various models such as the Random Forest Classifier, Logistic Regression, Support Vector Machine, and Extra Tree Classifier, notable accuracies ranging from 86.55% to 95.14% were achieved. These findings underscore the potential of machine learning in providing valuable insights for proactive attrition management, which aligns with the literature reviewed.

The Extra Tree Classifier consistently outperformed other models, exhibiting superior performance across accuracy, precision, recall, and F1-score metrics.

Furthermore, the performance of the machine learning models was validated using k-fold cross-validation, confirming that the models effectively predicted employee attrition.

5.1.2 IDENTIFICATION OF KEY ATTRITION PREDICTORS AND RISKS

The analysis identified predictors of employee attrition, including Overtime, JobLevel, MaritalStatus, YearsInCurrRole, and TotalWorkingYears. Additionally, the analysis revealed that most employees are dissatisfied with their work-life balance. While this has not led to a noticeable number of departures from the company, it could pose a potential attrition risk if not addressed promptly. So, addressing work-life balance issues can help mitigate attrition risk and improve overall employee experience and retention.

5.2 CONTRIBUTIONS OF RESEARCH

This study is a significant contribution to the field of workforce management. It demonstrates the effectiveness of machine learning in predicting employee attrition and identifying the factors that influence it. The research has improved data handling techniques and model accuracy, which will help advance predictive analytics in workforce management. The study also highlights the importance of proactive retention strategies based on the identified attrition predictors and risks, enabling stakeholders to retain valuable employees effectively.

5.3 REFLECTION

Embarking on this research journey proved to be enlightening and transformative. Applying the CRISP-DM framework and delving into predictive analytics honed the researcher's ability to decipher complex datasets, effectively bridging theoretical knowledge with practical application. This experiential learning reinforced academic concepts and sharpened analytical and problem-solving skills, marking a significant personal and professional growth milestone.

Furthermore, the analysis underscored the challenge of imbalanced data in machine learning. The findings highlight the importance of addressing class imbalance to enhance model performance and achieve more accurate predictions. This insight is a valuable lesson for future research endeavours and underscores the importance of data preprocessing techniques like SMOTE.

5.4 LIMITATIONS OF RESEARCH

Despite its contributions, the study encountered limitations that constrained its broader impact. Firstly, the restricted deployment of the proposed model to real-life scenarios limited the practical applicability of the findings, hindering their potential to inform decision-making in actual organisational settings. Additionally, the absence of contemporary workforce data variables such as work-from-home dynamics and external factors like national economic indicators limited the depth of understanding of attrition drivers. Integrating these elements could offer a more comprehensive perspective on employee attrition in modern workplaces, enhancing the relevance and applicability of the research findings.

CHAPTER SIX: RECOMMENDATIONS

6.1 SUMMARY OF FINDINGS

The dataset and methodology used in this study proved valuable for predicting employee attrition and successfully met the research objectives. Similarly, the models' performance underscores the effectiveness of machine learning in predicting attrition, with the Extra Tree classifier notably outperforming other models. Furthermore, the research identified predictors of attrition and the risks that require proactive management.

However, certain limitations were encountered, such as the inability to deploy models to real-life projects and the absence of contemporary workplace factors affecting attrition in the dataset.

Based on the results obtained, the following recommendations are suggested.

6.2 RECOMMENDATION TO HUMAN RESOURCE MANAGERS

With this significant milestone achieved, workplace stakeholders should embrace machine learning technologies for proactive attrition prediction. By leveraging insights from predictive models, stakeholders can develop targeted retention strategies to mitigate attrition's adverse effects on organisational performance. Proactively addressing attrition risks can foster a stable and productive workforce environment, contributing to long-term organisational success and employee satisfaction.

6.3 RECOMMENDATIONS FOR FUTURE RESEARCH

Based on the insights and findings, the following recommendations can further advance research and practical applications in predicting employee attrition.

Firstly, future endeavours should prioritise practical model implementation in organisational settings because validating the efficacy of developed models in actual workplace scenarios will provide valuable insights into their practical application.

Secondly, future research on attrition should incorporate contemporary workplace variables such as work-from-home dynamics and economic indicators, as these factors could enhance a more comprehensive understanding of attrition drivers in modern organisational contexts.

In summary, the outlined findings and recommendations offer valuable guidance for academia and industry, providing actionable insights to advance the understanding and management of employee attrition.

REFERENCES

Alkhatieeb A. (2023) *EDA and classification(93.5)IBM Employee Attrition*. https://www.kaggle.com/code/ahmadialkhatib/eda-and-classification-93-5-ibm-employee-attrition Accessed.

Alduayj, S.S. and Rajpoot, K. (2018) Predicting Employee Attrition using Machine Learning. *ResearchGate* .

Ali, M. (2022) *Supervised Machine learning*. https://www.datacamp.com/blog/supervised-machine-learning Accessed.

Alshiddy, M.S. and Aljaber, B.N. (2023) Employee Attrition Prediction using Nested Ensemble Learning Techniques. *International Journal of Advanced Computer Science and Applications* 14 (7), .

Bennett, B. (2020) *Machine Learning and Biodata: Predicting Early Employee Flight During Preselection.* https://www.proquest.com/docview/2442576747/EA10FA949CB34852PQ/12?accountid=17193&sourcetype=Dissertations%20&%20Theses Accessed 9 January 2024.

BLS (2023) *Job Openings and Labor Turnover – December 2023*. https://www.bls.gov/news.release/pdf/jolts.pdf

Bryman, A. and Bell, E. (2015) *Business research methods*. Oxford University Press, USA.

Chemuturi, V. and Chemuturi, M. (2022) *Managing of people at work*. .

Coursera (2023) *3 types of machine learning you should know*. https://www.coursera.org/articles/types-of-machine-learning Accessed 25 March 2024.

Crail, C. (2023) 15 Effective employee retention Strategies in 2024. *Forbes Advisor* 13 July.

CIPD (2024) Employee turnover and retention. https://www.cipd.org/uk/knowledge/factsheets/turnover-retention-factsheet/ Accessed 25 March 2024.

CIPD (2024) Why staff turnover data matters. https://community.cipd.co.uk/cipd-blogs/b/cipd_voice_on/posts/why-staff-turnover-data-matters Accessed 25 March 2024.

D'Allessandro, R. (2023) *Employee Attrition and How to Minimize it - Qualtrics*. https://www.qualtrics.com/experience-management/employee/employee-attrition/ Accessed 21 October 2023.

De Smet, A., Dowling, B., Hancock, B. and Schaninger, B. (2022) The Great Attrition is making hiring harder. Are you searching the right talent pools? https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-great-attrition-is-making-hiring-harder-are-you-searching-the-right-talent-pools Accessed.

Deberry, K. (2021) Pulse of the American Worker Survey: Is this working? https://news.prudential.com/presskits/pulse-american-worker-survey-is-this-working.htm Accessed 9 August 2023.

Dong, S. (2021) Multi class SVM algorithm with active learning for network traffic classification. Expert Systems With Applications 176, 114885.

Fleckenstein, M. and Fellows, L. (2018) *Modern Data Strategy*. *Springer eBooks* .

Galvin, G. (2023) Nearly 1 in 5 health care workers have quit their jobs during the pandemic. *Morning Consult Pro* 29 June.

Gartner (2023) Definition of Attrition - Gartner Human Resources Glossary. https://www.gartner.com/en/human-resources/glossary/attrition#:~:text=Attrition%20rate%20is%20the%20rate,a%20given%20period%20of%20time.

GfG (2023) *Types of machine learning*. https://www.geeksforgeeks.org/types-of-machine-learning/ Accessed 25 March 2024.

Guerranti, F. and Dimitri, G.M. (2022) A comparison of machine learning approaches for predicting employee attrition. *Applied Sciences* 13 (1), 267.

Hayes, A. (2023) *What is attrition in business? Meaning, types, and benefits*. https://www.investopedia.com/terms/a/attrition.asp#:~:text=The%20term%20attrition%20refers%20to,human%20resources%20(HR)%20professionals Accessed 15 January 2024.

Hobson, K. (2023) Five reasons Employees are your company's no. 1 asset. *Forbes* 12 September.

Holtom, B. (2019) *Better ways to predict who's going to quit*. https://hbr.org/2019/08/better-ways-to-predict-whos-going-to-quit Accessed.

Hotz, N. (2023) *What is CRISP DM? - Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-2/ Accessed.

IBM (n.d.) *What is machine learning (ML)?* | *IBM*. https://www.ibm.com/topics/machine-learning Accessed 25 March 2024.

Jones, M. (2023) *What is Attrition Rate and How to Calculate It*. https://www.callcentrehelper.com/attrition-89611.htm Accessed 25 September 2023.

Kaggle (2017) *IBM HR Analytics Employee Attrition & Performance*. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset Accessed.

Kumar, P. (2023) Uncovering the top reasons behind employee attrition: effects and strategies to combat them. https://kredily.com/reasons-behind-employee-attrition/ Accessed 25 March 2024.

Li, J., Zhu, Q., Wu, Q. and Zhu, F. (2021) A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences* 565, 438–455.

Lucas S. (2023) *Employee Attrition: Meaning, Impact & Attrition Rate Calculation*. https://www.aihr.com/blog/employee-attrition/#Reasons Accessed.

Mansor, N., Sani, N.S. and Aliff, M. (2021) Machine learning for predicting employee attrition. *International Journal of Advanced Computer Science and Applications* 12 (11), .

Martínez-Plumed F., Contreras-Ochando L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J. and Flach, P.A. (2021) CRISP-DM Twenty years Later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* 33 (8), 3048–3061.

Masese O. (2016) Employee Attrition Management by Engagement. E-ISSN No : 2454-9916 | Volume : 2 | Issue : 12 | Dec 2016

Miranda D. (2023) Best employee benefits in 2024. *Forbes Advisor* 6 February.

Mizar B. (2018) Using predictive analytics in employee retention. *FM Magazine* 1 December.

Molina E. (2022) A practical guide to implementing a random forest classifier in Python. *Medium* 3 March.

Moss R. and Moss, R. (2023) *One in five workers to quit in 2024*. https://www.personneltoday.com/hr/uk-attrition-2024-culture-amp/ Accessed.

Najafi-Zangeneh S., Shams-Gharneh, N., Arjomandi-Nezhad, A. and Zolfani, S.H. (2021) An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. *Mathematics* 9 (11), 1226.

O'Hara, R, Haylon, L., Boyle, D. Strategic Finance; Montvale Vol. 104, Iss. 8, (Feb 2023): 38-45. A Data Analytics Mindset With Crisp-Dm. https://www.proquest.com/docview/2770097738?sourcetype=Scholarly%20Journals

Ossai, C.I. and Wickramasinghe, N. (2022) GLCM and statistical features extraction technique with Extra-Tree Classifier in Macular Oedema risk diagnosis. *Biomedical Signal Processing and Control* 73, 103471.

Pasquarella A. (2023) A Machine Learning Approach to Predicting Federal STEM Workforce Attrition. https://www.proquest.com/docview/2803170339/214AF14961BF409BPQ/4?%20Theses&accountid=17193&sourcetype=Dissertations%20 Accessed 2 December 2023.

Pellegrini, V. (2023) *Preparing for the turnover crisis*. https://www.microsoft.com/en-us/worklab/preparing-for-the-turnover-crisis Accessed 8 November 2023.

Prashant (2020) *Random Forest Classifier tutorial*. https://www.kaggle.com/code/prashant111/random-forest-classifier-tutorial Accessed.

PWC (2024) *Workforce survey: Almost 20% of UK workers expect to quit in the next 12 months*. https://www.pwc.co.uk/press-room/press-releases/pwc-workforce-survey-20221.html Accessed 8 February 2024.

Raman, R., Bhattacharya, S. and Pramod, D. (2018) Predict employee attrition by using predictive analytics. *Benchmarking: An International Journal* 26 (1), 2–18.

Rane, S.B. and Narvel, Y.A.M. (2022), "Data-driven decision making with Blockchain-IoT integrated architecture: a project resource management agility perspective of industry 4.0", International Journal of System Assurance Engineering and Management, Vol. 13 No. 2, pp. 1005-1023.

Raza, A., Munir, K., Almutairi, M., Younas, F. and Fareed, M.M.S. (2022b) Predicting employee attrition using machine learning approaches. *Applied Sciences* 12 (13), 6424.

Repaso, J.A.A., Capariño, E.T., Hermogenes, M.G.G. and Perez, J.G. (2022) Determining factors resulting to employee attrition using data mining techniques. *International Journal of Education and Management Engineering* 12 (3), 22–29.

Roberthalf (2023) *14 Effective employee retention strategies*. https://www.roberthalf.com/us/en/insights/management-tips/effective-employee-retention-strategies Accessed.

Robinson A. (2022) *Employee Attrition: Definition, causes & Examples*. https://teambuilding.com/blog/employee-attrition Accessed.

Salunkhe, T.P. (2018) *Improving employee retention by predicting employee attrition using machine learning techniques*. https://esource.dbs.ie/items/05b4c876-f02a-4deb-b6bd-e4c11827eebe Accessed.

Saunders, M.N.K., Thornhill, A. and Lewis, P. (2019) *Research methods for business students*. .

Selvaraj, R. and Nagarajan, S.K. (2021) Land Cover Change Detection from Remotely Sensed IoT Data for Assessment of Land Degradation: A Survey. *Journal of Information & Knowledge Management* 20 (Supp01), 2140011.

Sghir, N., Adadi, A. and Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: a decade systematic review. Education and Information Technologies, Vol. 28, pp. 8299-8333.

Shearer, D. (2000) The CRISP-DM Model: The new Blueprint for Data Mining. J. Data Wareh. 2000, 5, 13–18.

Shweta (2022) Employee Turnover Rate: Definition & Calculation. *Forbes Advisor* 12 October.

Singh K. and Singh R. (2019) A Study on Employee Attrition: Effects and Causes. https://www.ijresm.com/Vol.2_2019/Vol2_Iss8_August19/IJRESM_V2_I8_48.pdf

Siraj, M. (2022) Analyzing ANOVA F-Test and sequential feature selection for intrusion detection systems. *International Journal of Advances in Soft Computing and Its Applications* 14 (2), 186–194.

Spain, E. and Groysberg, B. (2016) *Making exit interviews count*. https://hbr.org/2016/04/making-exit-interviews-count. Accessed 4 September 2023.

Statista (2023) *Employee attrition of professional services organizations worldwide 2013-2022*. https://www.statista.com/statistics/933710/professional-services-worldwide-employee-attrition/#:~:text=The%20employee%20attrition%20rate%20of,rate%20of%20almost%2014%20percent. Accessed 25 March 2024.

Stomps, J., Wilson, P.P.H., Dayman, K., Willis, M.J., Ghawaly, J. and Archer, D.E. (2023) SNM radiation signature Classification using different Semi-Supervised Machine Learning Models. *Journal of Nuclear Engineering* 4 (3), 448–466.

Turits, M. (2021) The delicate art of the exit interview. *BBC* 2 December.

Twin, A. (2024) *What is data mining? How it works, benefits, techniques, and examples*. https://www.investopedia.com/terms/d/datamining.asp Accessed.

Varma, D. and Dutta, P. (2021) Empowering human resource functions with data-driven decision-making in start-ups: a narrative inquiry approach. The International Journal of Organizational Analysis 31 (4), 945–958.

Wadikar, D. (2020) Customer Churn Prediction. *Customer Churn Prediction* .

Wallace, L. (2023) Five hidden costs of employee attrition. *Forbes* 21 March.

Wang, C.C., Li, Z., Chen, T., Wang, R. and Ju, Z. (2023) Research on the Application of Prompt Learning Pretrained Language Model in Machine Translation Task with Reinforcement Learning. *Electronics* 12 (16), 3391.

Wood, T. (2020) *Unsupervised learning*. https://deepai.org/machine-learning-glossary-and-terms/unsupervised-learning Accessed.

Workday (2021) *Why an exit interview won't help you reduce attrition*. https://blog.workday.com/en-us/2021/exit-interview.html#:~:text=Moving%20From%20an%20Exit%20Interview,employee%20turnover%20in%20your%20organization Accessed 16 February 2024.

Yahia, N.B., Hlel, J. and Colomo-Palacios, R. (2021) From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access* 9, 60447–60458.

Yang, S., Ravikumar, P. and Shi, T. (2020) IBM Employee Attrition Analysis. *ResearchGate* .

APPENDICES

1. PYTHON CODE

https://colab.research.google.com/drive/1nlp_vydMXn4mQ7_TqJObNBV8HfEkPXBw#scrollTo=xQ84Pb37PFd2

(Ctrl + click to follow the link to Python code)

2. RESEARCH PROPOSAL

CHUKWU GODSON UDE

Research Proposal

Topic:

PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING

Table of content

- **Scope/Rationale of the project**
- **Literature review**
- **Aims, Research objective and research questions.**
- **Methodology and data sources to be used:**
  i. Sample size and selection criteria
  ii. Data Collection
  iii. Data Analysis
- **Limitations and ethical consideration**
- **Conclusion**
- **Proposed Chapter Headings and Subheadings**

1. Introduction
2. Literature review
3. Methodology
4. Findings/analysis: case analysis and cross-case analysis
5. Discussion and conclusion

Bibliography

Appendices

Scope/Rationale of the project

Employees remain key stakeholders of any organization as they are responsible for fulfilling clients' expectations and fostering the general performance of organizations (Indeed, 2023). It is against this backdrop that employee attrition is a cause for concern for any entity. According to Gartner (2023), employee attrition refers to the voluntary or involuntary exit of employees from an organization for various reasons, ranging from termination, resignation, retirement, or death. Employee attrition, which also refers to the rate at which an organization loses employees over time, has been noted as a major issue facing the performance of organizations as it poses a gap that can disrupt operations and services. Such gaps come with loss of competitive advantage, service failures, and costs, including the cost of employing a new employee(s).

For this reason and the fact that one in four employees plans to look for a new job in a different company after COVID (Katherine, 2021), it is important that organizations manage talents, skills, and employees to avoid or reduce employee attrition. And by understanding the reasons behind employee attrition, employers can adopt measures to improve employee retention (Peters, 2023). Managing attrition in this modern time and pandemic era, requires proactiveness and data-driven decision-making on the part of employers, in contrast to the traditional exit interviews and continuous employee feedback that organizations conduct.

In order to advance the solution to this problem, researchers have previously proposed a solution using machine learning. And with the use of machine learning, computers are programmed with available data to learn patterns and make predictions using algorithms. This means that historical data about employee attrition is programmed into a computer to uncover patterns in such data and use them to predict whether an employee will leave or stay.

To add to this body of knowledge, this research work will propose a model that can predict employee attrition using machine learning algorithms. This work will also analyze using the dataset, important factors that influence an employee’s decision to stay or leave an organization. Also, this research work will highlight the importance of predicting employee attrition with machine learning over traditional exit interviews and continuous employee feedback. Similarly, the objective of this work is to assist organizations in retaining their desired employees by identifying and addressing important factors that cause attrition.

- **Traditional approaches to managing employee attrition**

Organizations use exit interviews and continuous employee feedback as means of managing employee attrition. However, these approaches have not significantly addressed the problems associated with employee attrition. According to Spain and Groysberg (2016), exit interviews too often fail to improve retention. In the same way, Workday (2021) is of the opinion that exit interviews are flawed in combating attrition. Workday revealed that their research suggests that regular employee feedback is more effective in handling employee attrition for the following reasons:

1. Issues are addressed as they arise.
2. Historical data is gathered, which is used to uncover the underlying reasons behind the exit.
3. The impact of the changes made to historical data can easily be measured.

Notably, exit interviews are conducted when an employee is leaving or has already left. At that point, it is hard to achieve retention. Continuous employee feedback, though better than an exit interview, cannot address all issues raised at once. This is because implementing resolutions to feedback is usually periodic. An employee may decide to leave if his concerns are yet to be addressed.

- **Machine learning in employee attrition prediction**

To understand the application of machine learning in employee attrition prediction, it is important to discuss machine learning.

Machine learning is a branch of artificial intelligence that uses data and algorithms to learn the way that humans do and gradually improves the accuracy of such learning (IBM 2021). In other words, machine learning involves the use of statistical methods in training algorithms to make predictions and uncover patterns. According to Zhou (2021), machine learning refers to a technique that learns from experience in order to improve a system's performance through computational methods. Here, experience can also be referred to as data. And machine learning develops algorithms that construct models from data in order to make predictions. Mohri et al. (2018) define machine learning as a computational method that uses experience to make accurate predictions. In this case, experience refers to past data available to the learner.

Machine learning can be supervised, unsupervised, and reinforcement learning. Supervised machine learning uses labelled input and output data. Supervised learning can take the form of classification or regression. Charles et al. (2022). In classification, the output variable, such as attrition, is categorized as leave or stay. In regression, the output variable is a continuous value such as price, speed, weight, etc.

Machine learning is unsupervised when input and output are not labelled, but instead labels are learned by algorithms. It is further classified as clustering algorithm and an association algorithm. Charles et al. (2022).

In reinforcement learning, optimal behaviour is learned to in order to gain maximum reward.

In other words, Reinforcement learning uses an algorithm that learns from outcomes and uses the result to decide the next action (Prateek 2023).

Applying machine learning to employee attrition prediction entails using historical employee data to train an algorithm that can learn patterns from the data and use them to predict the chances of an employee leaving or staying.

In this project, historical data generated by IBM is used (source: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset). It is this data that will be trained or fitted into a model or algorithm and then used to ascertain if an employee will stay or leave.

- **Machine learning models for predicting employee attrition**

There are various machine learning algorithms; however, this work will focus on the ones listed below.

1. **Logistic Regression**

Logistic regression is an approach in statistics that describes the relationship between an independent variable and a binary outcome classified as yes or no, good or bad, leave or stay, etc. (Das 2021).

According to Hosmer et al. (2013), logistic regression is the model used to show the relationship between a discrete dependent variable, say Y, and independent variables, say X1 to Xn.

In their opinion, Setiawan et al. (2020) described logistic regression as an integral part of data analysis that explains the correlation between an output variable (attrition) and input variables (employee age, marriage status, salary, etc. They used this model and found out that improving job satisfaction, working environment, and employee workload, salary, can reduce attrition rate to a large extent. However, they did not rank the features or variables in their order of importance in determining if an employee would leave or not. Also, they did not identify the usefulness of using machine learning over exit interviews which this work will also cover. This research will cover the gaps.

Sri et al. (2020) noted that logistic regression is used for predicting binary classes or two-class classifications. With the use of logistic regression, they found out that a lack of pay raises and openings for employee development influence employee attrition.

In the case of this project, logistic regression will be used to describe the relationship between the independent variables (the 32 attributes in the dataset) and the dependent variable, attrition or no attrition.

- **Random Forest**

Random Forest is a supervised machine learning algorithm that aggregates random trees (Pratt et al., 2021). It is popularly known as an ensemble learning algorithm because it builds multiple decision trees to generate a machine learning predictor, which is the Random Forest (Reinstein 2017). Random forest is a machine learning algorithm that builds a multitude of decision trees and makes the individual trees' mean predictions the output, Charles et al. (2022). Random Forest is one of the machine learning algorithms that uses multiple decision trees without overfitting.

Pratt et al. (2021), in their application of Random Forest to employee attrition prediction, stated that Random Forest is a group of random trees. They further stated that, though often more effective than basic decision trees, random forests have the drawback of being more challenging to interpret. They found out in their work that the most important features determining attrition are monthly income, age, daily rate, total working years, and monthly rate. However, they suggested that other models be used to confirm if they produce similar results. This work will use other models as suggested.

- **Decision Trees**

Due to their resemblance to human reasoning and simplicity, decision trees are frequently employed to build classification models (Kotsiantis, 2011).

Decision tree classifiers have proven to work well for multistage decision-making, such as employee attrition prediction. A difficult decision is usually broken down into multiple simpler judgements in a multistage decision-making process (Bhartiya et al., 2019).

A decision tree is a non-parametric supervised learning approach that can be used for both regression and classification applications. It has an internal root node, branches, internal nodes, and leaf nodes in a hierarchical tree structure. By using a greedy search to locate the ideal split points inside a tree, decision tree learning uses a divide and conquer tactic (IBM 2020).

- **Gradient Boosting (XGBoost)**

In terms of effective memory utilisation, high accuracy, and quick processing times, XGBoost is regarded as a superior algorithm. It handles all types of noise from enormous data sets and transforms the data into a ready-acceptable form for precision findings (Jain et al., 2018). They suggested XGBoost for the purpose of successfully enabling the organisation to take preventive action in due course and proposed a real-world application. They got a more accurate prediction using XGBoost for predicting employee attrition when compared to other models used.

Research Objective and Research Questions

The objective of this research is to assist organizations to manage their workforce by predicting employee attrition using machine learning. And to achieve that, it is important that answers are provided to the following research questions:

- What are the features from the dataset that are more important in predicting whether an employee will leave or stay?
- What is the machine learning model that works best for predicting employee attrition?
- Why will an organization prefer the use of machine learning to predict employee attrition over traditional exit interviews?

Methodology

To achieve the objective of this research work, the following steps must be taken:

The Python programming language is to be used for data collection, preparation, exploration, visualization, analysis, model application, and evaluation within the Jupyter Notebook environment.

The research design for this work is both descriptive and predictive, as Jain et al. (2018) suggested for their related research work. It is descriptive because the data being used is historic, and predictive because the outcome of the research is to predict employee attrition.

The following steps are to be taken:

Collection of data: This will involve importing the IBM HR dataset to the Python library, as sourced from Kaggle and followed by these steps:

1. **Data preprocessing** will involve checking for missing values and outliers, separating input and output variables, removing duplicates and unwanted features, splitting data for training and testing, etc. For instance, ‘Employeecount’ and ‘EmployeeNumber’ are not so relevant in the work and would be removed at this stage. It is noted that the data does not have a missing value. Also, data will be split into training and testing sets. To split the data, the train_test_split function is imported from Python's Sklearn package to split the dataset into 70% for training and 30% for testing.
2. **Data exploration**: This will involve checking for data types, column names, and calculating summary statistics like the standard deviation, mean, and median. For example, the input variable 'hourlyrate' is a numerical value. This is what data exploration here will find out.
3. **Model training and testing**: At this stage, the model for the machine learning algorithm will be introduced or imported. The algorithms that will be used in this project include Logistic Regression, Decision Tree, Random Forest and XGBoost. Each of them will be used to train a model for predicting employee attrition. Each model will also be tested using the test data for validation.

Model Evaluation: The models will be evaluated for accuracy, precision, and recall. All the

1. algorithms used will be compared to ascertain the one that performs best, as this will inform the one to be deployed for employee attrition prediction.

Sample Size and Selection Criteria

Dataset:

The dataset consists of historical data generated by IBM (source: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset). The data consists of 1470 attributes and includes 34 independent metrics and variables listed and explained below:

- Age: employee age.
- 'BusinessTravel': This shows if an employee has embarked on business travel or not.
- 'DailyRate': This shows the daily payment rate.
- Department: The department for each employee
- 'DistanceFromHome': The distance between home and place of work.
- Education: depicts the level of education of the employee.
- 'EducationField': Shows the course of study.
- Employee count: shows the number of employees with a particular attribute.
- 'EmployeeNumber': This is the staff number.
- 'EnvironmentSatisfaction': depicts the level of satisfaction with the work environment.
- Gender: this is the gender of the employee, male or female.
- 'HourlyRate': This shows the hourly payment rate.
- 'JobInvolvement': means how involved employees are with their job.
- 'JobLevel': This is otherwise the rank of an employee.
- 'Jobrole': this is the role or function of an employee.
- 'Jobsatisfaction': this shows whether an employee is satisfied or not with his job.
- 'MaritalStatus': This shows whether an employee is married or not.
- 'MonthlyIncome': This is the monthly income of an employee.
- 'MonthlyRate': This shows the monthly payment rate.
- 'NumCompaniesWorked': This is the number of companies where an employee worked.
- 'Over18': This indicates if an employee is over or below the age of 18.
- 'OverTime': shows if an employee works overtime or not.
- 'PercentSalaryHike': shows the percentage increase in the salary of an employee.

- ‘RelationshipSatisfaction’: Shows relationship satisfaction between employees and management.
- ‘StandardHours’: This shows the hours worked by an employee.
- ‘StockOptionLevel’: employees offered stock option level.
- ‘TotalWorkingYears’: This shows the total number of years that an employee has worked.
- ‘TrainingTimesLastYear’: number of times that an employee embarked on training the previous year.
- ‘WorkLifeBalance’: This shows employees' satisfaction with work-life balance.
- ‘YearsAtCompany’: depicts the number of years an employee has worked in the company.
- ‘YearsInCurrentRole’: the number of years an employee has been in a current role.
- ‘YearsSinceLastPromotion’: number of years since last promotion.
- ‘YearsWithCurrManager’: shows how long an employee has worked with his current manager.

The dependent variable in the dataset is ‘Attrition’.

And these two features, ‘EmployeeNumber’ and ‘EmployeeCount’ will be excluded from the attributes because, from a glance, I consider them not important in determining if an employee will leave or stay back in an organization.

Data Collection

As stated earlier, the data used in this research work is secondary data provided by IBM and available on Kaggle. The dataset suits the research objective because it comprises demographic information, performance metrics, training records, and other human resource data required to effectively predict employee attrition. The Python programming language is used to collect or upload the data to the Jupyter Notebook environment for analysis.

Data Analysis

Data analysis will be done in the following stages:

1. **Data preprocessing**: the HR data is checked and cleaned for missing values, outliers, duplicates, etc. A check on the data shows that there are no missing values. However, there is an imbalance regarding gender, as there are more males than females, and this may not give a balanced outcome.
2. **Exploratory Data Analysis**: summary statistics will be applied here to get instances like the mean, median, and standard deviation of the data.

- **Feature selection**: The Random Forest classifier used in the research has a function for selecting feature importance and ranking them in order of importance. The most important feature has more impact on prediction than other features.
- **Model selection and evaluation**: Four machine learning algorithms and models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—are to be applied individually in training and testing data. The highest performing would be recommended for use.

Limitations & ethical consideration

Limitations

A major limitation of the research is the inability to apply it to a real-life HR project. Applying it in real life will prove its efficiency, or otherwise.

Another limitation is the amount of data used. A larger dataset will be more efficient in the prediction than the one currently in use, because it is generally believed that a larger dataset performs better than a smaller dataset.

Also, dataset that features post COVID trend like work-from-home, should be used for future analysis.

Ethical considerations

Employee data is highly sensitive, and the need to abide by the provisions of the General Data Protection Regulation has made it difficult to use true employee data for this purpose. Consideration is also given to my responsibility for the outcome of deploying the machine learning model that I will recommend.

Conclusion

Employee attrition has been identified as a major issue that organizations must deal with. Unplanned employee attrition poses a great concern because when an employee leaves, a gap is created, which can reduce performance and cause disruptions and service failure.

It is on account of this that this project tries to use machine learning algorithms to predict employee attrition. This is believed to offer a better solution than the traditional exit interview because it is more preventive than prescriptive.

Proposed Chapter Headings and Sub-Headings

1. Introduction
2. Literature Review
3. Methodology
4. Findings/analysis: case analysis and cross-case analysis
5. Discussion and conclusion

Bibliography

Alduayj S. S. and Rajpoot K. (2018) *Predicting Employee Attrition using Machine Learning*. https://ieeexplore.ieee.org/document/8605976/authors#authors.

Bhartiya, N., Jannu, S., Shukla, P. and Chapaneri, R. (2019) Employee Attrition Prediction Using Classification Models. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) IEEE.

Brayfield, A.H. (1955) *Employee attitudes and employee performance*.

Charles, V., Emrouznejad, A., Gherman, T. and Cochran, J. (2022) Why Data Analytics is an Art. Significance 19 (6), Oxford University Press (OUP)42–45.

Chemuturi, M. and Chemuturi, V. (2019) Managing People at Work. River Publishers Management Sc.

Chung, D., Yun, J., Lee, J. and Jeon, Y. (2023) *Predictive model of employee attrition based on stacking ensemble learning*. Elsevier BVhttps://doi.org/10.1016/j.eswa.2022.119364.

Das, A. (2021) *Logistic Regression*. https://doi.org/10.1007/978-3-319-69909-7_1689-2.

Gartner (2023) *Definition of Attrition - Gartner Human Resources Glossary*. https://www.gartner.com/en/human-resources/glossary/attrition.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*. John Wiley & Sons.

Huettich, J. (2022) *The Ultimate Guide to Employee Attrition: What It Is, Causes, Risks & Prevention*. https://www.teamly.com/blog/employee-attrition/.

IBM (2023) *What is a Decision Tree | IBM*. https://www.ibm.com/topics/decision-trees

Indeed Editorial Team (2023) What Is Attrition? (Plus How To Calculate It in 5 Steps). https://www.indeed.com/career-advice/career-development/what-is-attrition

Jain R. and Nayyar A. (2018) *Predicting Employee Attrition using XGBoost Machine Learning Approach*. https://ieeexplore.ieee.org/document/8746940.

Jhaver M, Gupta Y. and Mishra A. K. (2019) *Employee Turnover Prediction System*. https://ieeexplore.ieee.org/abstract/document/9036180?casa_token=mtOz2DBgx6MAAAAA:Q47qOpkvxHlq8XS4vSLoFByDVTcWcPA-RTV0kyrF53puegOel-uiylYfrGTnFQ8p7p75wIZzm2Q.

Kaggle (2023) IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Katherine (2021) Pulse of the American Worker Survey: Is This Working? Prudential. https://news.prudential.com/presskits/pulse-american-worker-survey-is-this-working.htm

Kotsiantis, S.B. (2011) Decision trees: a recent overview. Artificial Intelligence Review 39 (4), Springer Science and Business Media LLC261–283.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018) Foundations of Machine Learning, second edition.

MIT Press.

Peters R. (2023) *CIPD | Employee Turnover & Retention | Factsheets*. https://www.cipd.org/uk/knowledge/factsheets/turnover-retention-factsheet/.

Prateek B. (2023) *Reinforcement learning*. https://www.geeksforgeeks.org/what-is-reinforcement-learning/.

Pratt, M., Boudhane, M. and Cakula, S. (2021) Employee Attrition Estimation Using Random Forest Algorithm. Baltic Journal of Modern Computing 9 (1), University of Latvia

Pratt, M., Boudhane, M. and Cakula, S. (2021) Employee Attrition Estimation Using Random Forest Algorithm. Baltic Journal of Modern Computing 9 (1), University of Latvia.
Regression. International Journal for Research in Applied Science & Engineering Technology (IJRASET). ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429. Volume 8 Issue V May 2020- Available at www.ijraset.com

Reinstein, I. (2017) *Random Forests®, Explained - KDnuggets*. https://www.kdnuggets.com/2017/10/random-forests-explained.html Accessed.

Setiawan, I., Suprihanto, S., Nugraha, A.C. and Hutahaean, J. (2020) HR analytics: Employee attrition analysis using logistic regression. *IOP Conference Series: Materials Science and Engineering* 830 (3), IOP Publishing032001.

Spain, E., Groysberg B (2016) *Making Exit Interviews Count*. https://hbr.org/2016/04/making-exit-interviews-count.

Sri R.P., Gopi K. M., Srinivasulu P, Ramya V., Bhaskar K (2020) Employee Attrition Prediction using Logistic Regression.

Workday Staff Writers (2021) *Why an Exit Interview Won't Help You Reduce Attrition*. https://blog.workday.com/en-us/2021/exit-interview.html#:~:text=Moving%20From%20an%20Exit%20Interview,employee%20turnover%20in%20your%20organization.

Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A., & Kulkarni, D. (2018). Employee attrition prediction. arXiv preprint arXiv:1806.10480. 10.48550/arXiv.1806.10480.

Zhou, Z.-H. (2021) Machine Learning. Springer Nature.