



**AN ANALYTICAL FRAMEWORK FOR EMPLOYEE
PROMOTION MODELING**

THEERAMET KAEWWISET

**DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING**

**SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY**

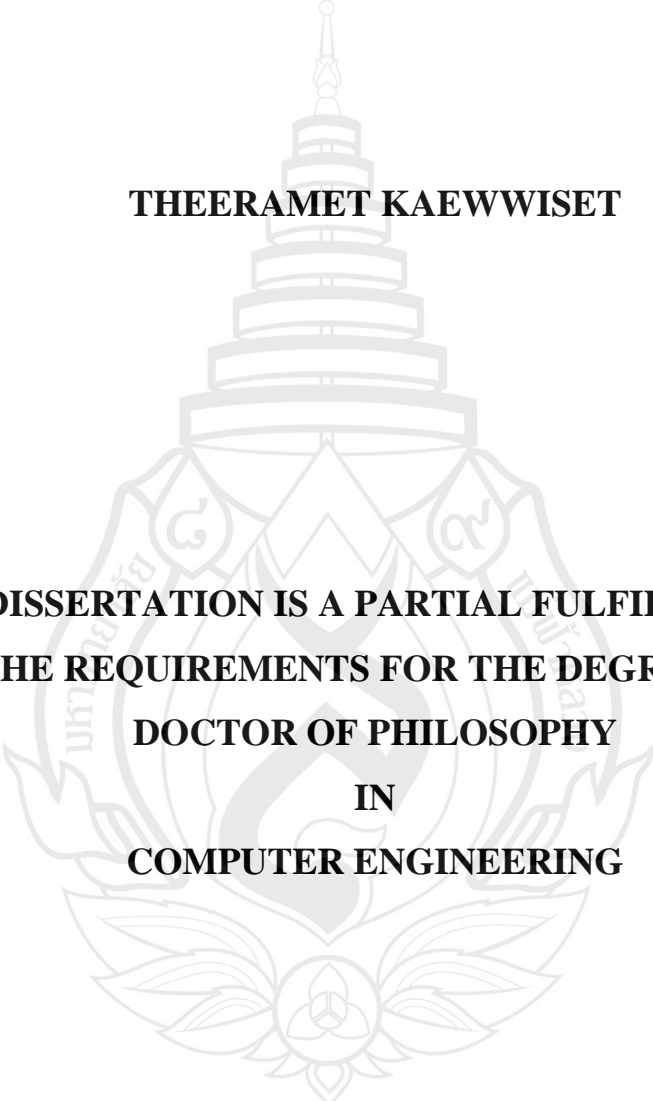
2025

©COPYRIGHT BY MAE FAH LUANG UNIVERSITY

**AN ANALYTICAL FRAMEWORK FOR EMPLOYEE
PROMOTION MODELING**

THEERAMET KAEWWISET

**THIS DISSERTATION IS A PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING**



**SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY**

2025

©COPYRIGHT BY MAE FAH LUANG UNIVERSITY



**DISSERTATION APPROVAL
MAE FAH LUANG UNIVERSITY
FOR**

DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING

Dissertation Title: An Analytical Framework for Employee Promotion Modeling

Author: Theeramet Kaewwiset

Examination Committee:

Associate Professor Adisorn Leelasantitham, Ph. D.	Chairperson
Associate Professor Punnarumol Temdee, Ph. D.	Member
Associate Professor Rounsang Chaisricharoen, Ph. D.	Member
Associate Professor Nattapol Aunsri, Ph. D.	Member
Assistant Professor Chayapol Kamyod, Ph. D.	Member

Advisor:

P. Temdee

.....Advisor
(Associate Professor Punnarumol Temdee, Ph. D.)

Dean:

N. Chondamrongkul

.....
(Assistant Professor Nacha Chondamrongkul, Ph. D.)

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation to my advisor, Associate Professor Punnarumol Temdee, Ph. D., for her invaluable guidance, continuous support, and insightful feedback throughout the course of this research. Her expertise, patience, and encouragement played a crucial role in shaping the direction of this thesis and in helping me overcome various academic challenges. Her dedication as a mentor has not only enhanced the quality of this work but also contributed significantly to my personal and professional development.

I am also grateful to Mae Fah Luang University for providing both academic and financial support during my studies, including a scholarship specifically granted for the completion of this doctoral dissertation. Special thanks are extended to the faculty members of the School of Information Technology and the staff of the Postgraduate Office, whose assistance and resources were vital to the successful completion of this thesis.

I wish to extend my sincere thanks to the members of my thesis examination committee: Associate Professor Adisorn Leelasantitham, Ph. D., Associate Professor Rungsan Chaisricharoen, Ph. D., Associate Professor Nattapol Aunsri, Ph. D., and Assistant Professor Chayapol Kamyod, Ph. D. Their critical evaluation, constructive comments, and thoughtful suggestions helped refine this research and broadened my academic perspective. I deeply value their time, expertise, and contributions to the improvement of this work.

Theeramet Kaewwiset

Dissertation Title	An Analytical Framework for Employee Promotion Modeling
Author	Theeramet Kaewwiset
Degree	Doctoral of Philosophy (Computer Engineering)
Advisor	Associate Professor Punnarumol Temdee, Ph. D.

ABSTRACT

Employee promotion represents a strategic function within human resource management, bearing significant implications for workforce motivation, organizational advancement, and career development. However, conventional promotional decisions often rely on subjective judgment, which can lead to inconsistencies and potential bias. Therefore, to address these challenges, this research introduces a comprehensive machine learning framework that incorporates both supervised and unsupervised learning techniques to enhance the identification of promotable employees. The proposed framework consists of three essential components. The first component is feature augmentation, which is performed through the construction of a novel engineered variable termed the Generated Promotion Feature (GPF), derived from performance-driven indicators such as key performance index (KPI) scores, award history, and average training performance. The second component is feature extraction, performed through Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce data dimensionality, identify critical structures, and improve computational efficiency while preserving informative patterns. The last component is the Synthetic Minority Oversampling Technique (SMOTE), employed to address class imbalance and enhance the model's ability to recognize underrepresented cases of promotion. In addition, two publicly available human resource datasets were utilized to validate the proposed methodology across six classification algorithms: Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, and Neural Network, as well as two clustering techniques, known as K-means and Fuzzy C-means. Experimental results demonstrate that, in classification tasks, the application of SMOTE significantly improves model performance across all algorithms, particularly in handling class imbalance and enhancing recall and F1-score.

In clustering tasks, the combination of GPF, PCA, and SMOTE yields the best results, producing more apparent cluster separations and greater consistency across different configurations. Among the dimensionality reduction methods, PCA outperforms t-SNE in both clustering quality and model stability. Additionally, the introduction of GPF, a domain-informed feature derived from high-correlation performance indicators, enhances model interpretability and discriminatory power. These findings suggest that the proposed framework offers a robust and generalizable approach for employee promotion modeling, adaptable to both supervised and unsupervised learning scenarios within diverse organizational contexts.

Keywords: Employee Promotion, Feature Engineering, SMOTE, Clustering, Classification

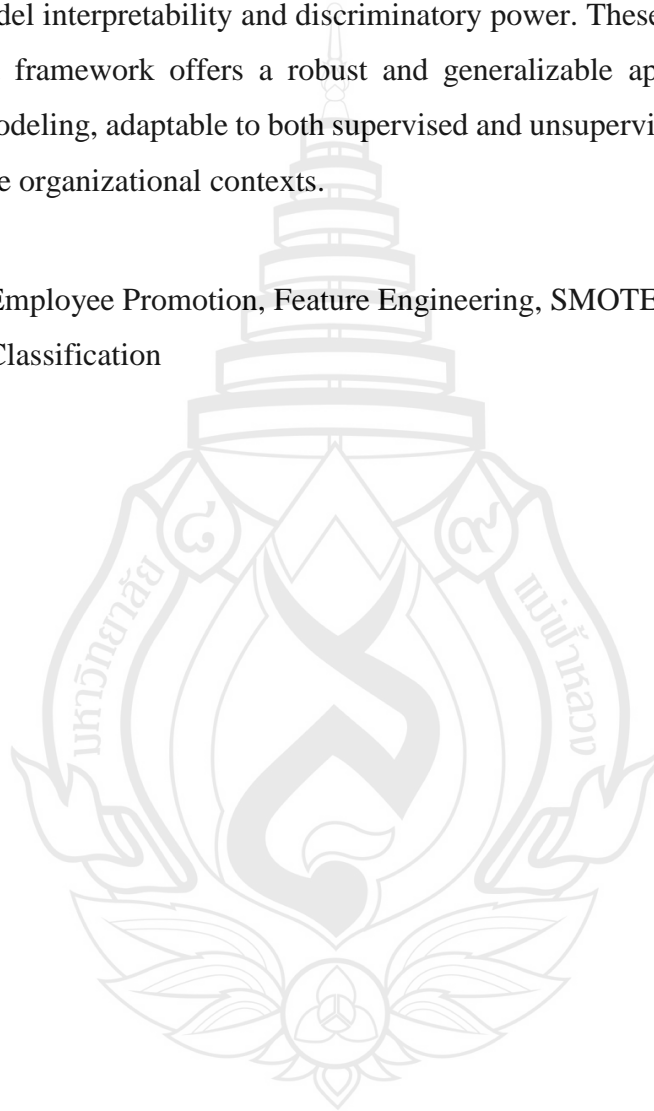


TABLE OF CONTENTS

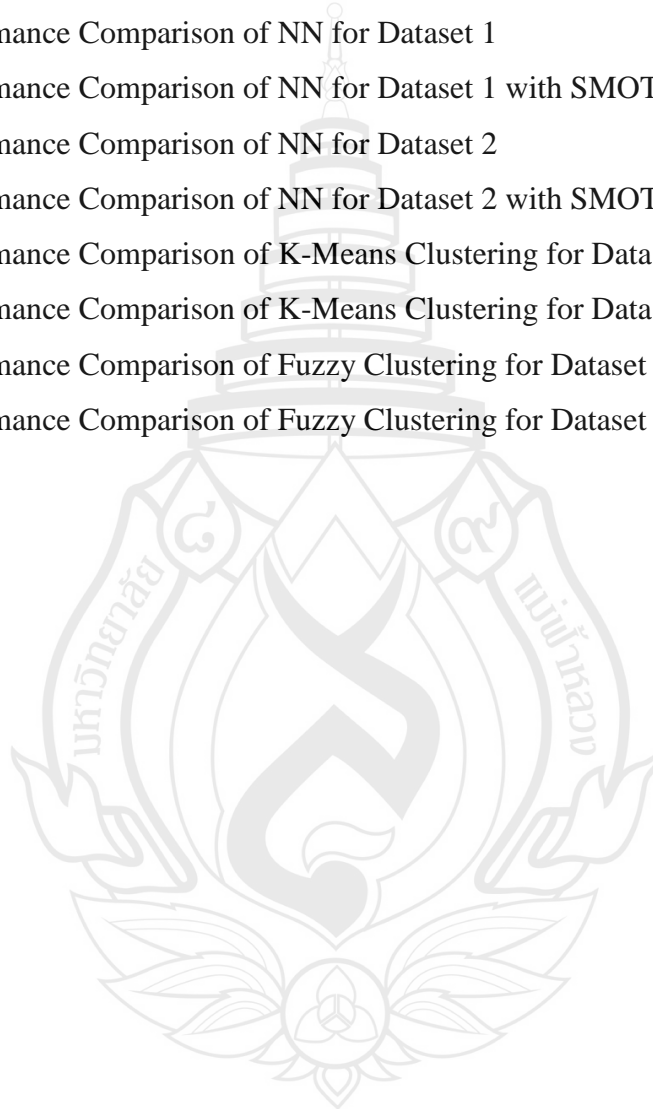
CHAPTER	Page
1 INTRODUCTION	1
1.1 Introduction	1
1.2 Objective	3
1.3 Scope of Work	3
1.4 Definition of Terms	4
2 LITERATURE REVIEW	5
2.1 Literature Review	5
2.2 Background Theory	15
2.3 The Proposed Work	23
3 METHODOLOGY	25
3.1 Method Overview	25
3.2 Data Collection	28
3.3 Data Preprocessing	31
3.4 Feature Engineering	32
3.5 Data Balancing	39
3.6 Model Construction	41
3.7 Result Evaluation	43
4 EXPERIMENTAL RESULTS	49
4.1 Classification Results	49
4.2 Clustering Results	71
5 DISCUSSION AND CONCLUSION	74
5.1 Classification	74
5.2 Clustering	76
5.3 Conclusion	77
5.4 Suggestion and Future Work	78
REFERENCES	80
CURRICULUM VITAE	91

LIST OF TABLES

Table	Page
3.1 Attribute Description of HR Analysis Case Study Dataset	28
3.2 Attribute Description of Data Science Staff Promotion Prediction Dataset	29
3.3 Examples of the Combined Data Sets (PCA and GPF) for Dataset1	38
3.4 Examples of the Combined Data Sets (T-SNE and GPF) for Dataset 1	38
3.5 Examples of the Combined Data Sets (PCA and GPF) for Dataset 2	39
3.6 Examples of the Combined Data Sets (t-SNE and GPF) for Dataset 2	39
3.7 Comparison of Selected Features of Dataset 1 before and after SMOTE	40
3.8 Comparison of Selected Features of Dataset 2 before and after SMOTE	41
4.1 Performance Comparison of Random Forest for Dataset 1	50
4.2 Performance Comparison of Random Forest for Dataset 1 with SMOTE	50
4.3 Performance Comparison of Random Forest for Dataset 2	51
4.4 Performance Comparison of Random Forest for Dataset 2 with SMOTE	52
4.5 Performance Comparison of Decision Tree for Dataset 1	53
4.6 Performance Comparison of Decision Tree for Dataset 1 with SMOTE	53
4.7 Performance Comparison of Decision Tree for Dataset 2	54
4.8 Performance Comparison of Decision Tree for Dataset 2 with SMOTE	55
4.9 Performance Comparison of SVM for Dataset 1	56
4.10 Performance Comparison of SVM for Dataset 1 with SMOTE	56
4.11 Performance Comparison of SVM for Dataset 2	57
4.12 Performance Comparison of SVM for Dataset 2 with SMOTE	58
4.13 Performance Comparison of KNN for Dataset 1	59
4.14 Performance Comparison of KNN for Dataset 1 with SMOTE	60
4.15 Performance Comparison of KNN for Dataset 2	61
4.16 Performance Comparison of KNN for Dataset 2 with SMOTE	62
4.17 Performance Comparison of LR for Dataset 1	63
4.18 Performance Comparison of LR for Dataset 1 with SMOTE	64

LIST OF TABLES

Table	Page
4.19 Performance Comparison of LR for Dataset 2	65
4.20 Performance Comparison of LR for Dataset 2 with SMOTE	66
4.21 Performance Comparison of NN for Dataset 1	67
4.22 Performance Comparison of NN for Dataset 1 with SMOTE	68
4.23 Performance Comparison of NN for Dataset 2	69
4.24 Performance Comparison of NN for Dataset 2 with SMOTE	70
4.25 Performance Comparison of K-Means Clustering for Dataset 1	71
4.26 Performance Comparison of K-Means Clustering for Dataset 2	71
4.27 Performance Comparison of Fuzzy Clustering for Dataset 1	72
4.28 Performance Comparison of Fuzzy Clustering for Dataset 2	72



LIST OF FIGURES

Figure	Page
3.1 Overall Methodology	25
3.2 Data Proportion of Dataset 1	30
3.3 Data Proportion of Dataset 2	30
3.4 Example Dataset Before Data Pre-processing Step	31
3.5 Example Dataset After Data Pre-processing Step	32
3.6 Conceptual Diagram of Proposed Framework	33
3.7 Performance-Oriented Feature Correlation Dataset 1	34
3.8 Performance-Oriented Feature Correlation Dataset 2	34
3.9 Twenty Percent of PCA Features that Most Variance with Promotion of Dataset 1	37
3.10 Twenty Percent of PCA Features that Most Variance with Promotion of Dataset 2	37
3.11 PCA Combined Dataset1 before and after SMOTE	40
3.12 PCA Combined Dataset2 before and after SMOTE	40

CHAPTER 1

INTRODUCTION

1.1 Introduction

Employee promotion represents a fundamental aspect of human resource management (HRM), as it directly affects workforce motivation, career progression, and overall organizational growth. Promotion decisions serve not only as a means of recognizing individual achievement and assigning higher levels of responsibility but also as a strategic tool to improve employee engagement and maintain top talent. Traditionally, promotion decisions have been based on subjective judgment by supervisors. This practice in large organizations may lead to particularly significant inconsistencies, bias, and inefficiencies.

The advent of data-driven decision-making, which involves machine learning (ML) techniques, has made them practical tools for analyzing employee data to support decision-making about employee promotions. ML-based approaches provide the analysis of large volumes of human resource data, patterns, and predictive models that support fair and scalable decision-making processes. Within the ML domain, promotion modeling can be approached through supervised learning, which performs classification based on outcomes (i.e., promoted or not promoted), and unsupervised learning applies clustering techniques to discover natural groupings without relying on predefined labels. However, both methodologies face key challenges, including imbalanced data distribution, noise, and high-dimensional features, all of which can compromise the accuracy and generalizability of these models.

To address these challenges, this study proposes a comprehensive framework that integrates feature extraction, feature augmentation, and methods for handling imbalances. Feature extraction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are employed to transform high-dimensional data into lower-dimensional spaces, thereby improving computational efficiency while retaining meaningful variance and local structure.

These transformations benefit both classification and clustering models by reducing complexity.

In addition, the study also presents a feature augmentation strategy through the construction of the Generated Promotion Feature (GPF), an engineered variable derived from highly correlated performance-oriented attributes such as KPI achievements, award records, and training scores. The inclusion of GPF aims to enhance the model's ability to distinguish between promotable and non-promotable employees by prioritizing relevant performance signals over less informative personal characteristics. This augmentation supports both supervised and unsupervised models by increasing the feature space with more representative indicators of promotion potential.

Furthermore, to address the imbalance issue where promotion cases are substantially more than non-promotion cases, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates synthetic instances for the minority class to create a balanced dataset that enables classification models and to learn equitably from both classes, as well as improving the quality of clustering by ensuring better representation of underrepresented groups.

Through the integration of these components, the proposed framework aims to construct strong and generalizable promotion models that are applicable across diverse datasets and learning paradigms. The resulting models not only enhance predictive accuracy and clustering quality but also provide valuable insights for HR practitioners in identifying, developing, and promoting high-potential employees. To evaluate the practical utility of the proposed framework, extensive experiments were conducted across both classification and clustering tasks. For classification, the integration of SMOTE with various classifiers demonstrated a significant performance improvement, particularly in handling imbalanced datasets by enhancing recall and F1-scores without significantly compromising precision. However, for clustering, the optimal results were observed when SMOTE was used in combination with the Generated Promotion Feature (GPF) and Principal Component Analysis (PCA). This combination yielded the most stable and well-separated clusters, reflecting an enhanced ability to distinguish between promotable and non-promotable employees. These findings suggest that while SMOTE alone is sufficient to improve classification models, effective clustering in HR

contexts benefits significantly from a multi-faceted preprocessing strategy involving both feature engineering and dimensionality reduction.

1.2 Objective

1.2.1 To create a novel feature engineering called the Generated Promotion Feature (GPF), derived from key performance indicators.

1.2.2 To create an analytical framework focusing on two primary feature categories: personal attributes and performance-oriented features for employee promotion modeling.

1.2.3 To evaluate the effectiveness of the proposed analytical frameworks with classification and clustering models.

1.3 Scope of Work

1.3.1 This study is conducted using two publicly available HR datasets to evaluate the performance and generalizability of the proposed classification and clustering models. While the evaluation focuses on these specific datasets, the findings intend to demonstrate the broader applicability of the models to various HR contexts, such as employee promotion, training needs analysis, and talent development planning.

1.3.2 The analysis emphasizes performance-oriented features within each dataset for the construction of the Generated Promotion Feature (GPF). This approach ensures that the engineering feature emphasizes data-driven relevance and practical interpretability.

1.4 Definition of Terms

1.4.1 Human Resource Management is a strategic approach to managing people in an organization, including recruitment, training, evaluation, compensation, and promotion to optimize organizational performance and employee development.

1.4.2 Employee Promotion is the advancement of an employee to a higher job position within an organization, generally accompanied by increasing responsibilities, higher compensation, and improved benefits.

1.4.3 Clustering is an unsupervised machine learning technique used to group similar data points based on feature similarity, without relying on predefined labels.

1.4.4 Generated Promotion Feature (GPF) is a newly constructed feature that synthesizes performance-based indicators such as KPI, training scores, and awards to represent an employee's potential for promotion accurately.

1.4.5 SMOTE (Synthetic Minority Oversampling Technique) is a technique to equalize class distribution in imbalanced datasets by creating synthetic instances of the minority class.

1.4.6 PCA (Principal Component Analysis) is a statistical technique for dimensionality reduction by changing features into a set of linearly uncorrelated components that capture the maximum variance in the data.

1.4.7 t-SNE (t-distributed Stochastic Neighbor Embedding) is a nonlinear dimensionality reduction technique, mainly used for data visualization, that maintains local similarities between data points.

1.4.8 KPI (Key Performance Indicator) is a quantifiable measure in evaluating the success of an employee to meet objectives relevant to their role.

1.4.9 Performance Feature is a type of feature in HR data that reflects an employee's work-related achievements, outcomes, or measurable behavior.

1.4.10 Personal Feature is a type of feature that records demographic or personal attributes of an employee, such as age, gender, marital status, or education level.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review

2.1.1 Human Resource Research with IT Aspect

According to a review of human resource literature, the variety of classifications of HR datasets is typically performed to satisfy different research objectives. For example, some studies have proposed the determinations of professionals (Asim et al., 2018) or experts (Hoon et al., 2015) within the organization (Kaewwiset et al., 2021). Some studies were related to talent (Stephanie & Sarno, 2019), competency (Guohao et al., 2019), or employee performance (Nedelcu et al., 2020). Some studies developed a model of employee engagement (Chen & Gong, 2013). Some works have proposed the determination of the position allocation (Ramdhani et al., 2016; Mathew et al., 2018) or job seeker classification (Hartanto et al., 2019) for finding new employees. Some works have proposed the determination of risk assessment or feature focusing on staff turnover (Tarusov & Mitrofanova, 2019; Wang et al., 2009), or predicting employee turnover (Juvitayapun, 2021). Overall, the reviewed literature demonstrates a growing emphasis on leveraging machine learning and data analytics in various HR domains, such as talent identification, competency analysis, position allocation, and turnover prediction. The proposed study built upon these foundations by introducing a unified, data-driven model designed explicitly for predicting promotions. By integrating feature engineering, dimensionality reduction, and data balancing, this research not only addresses key methodological gaps observed in prior studies but also advances HR analytics through a more interpretable and performance-oriented decision-making approach.

2.1.2 Performance Evaluation with IT Aspect

Many human resource studies in information technology fields provide the necessary performance evaluation in the human resource management process. The importance of human resource management lies in effectively managing human

resources within an organization to attract and retain high-quality manpower. Moreover, the quality and performance of practice can determine the company's fate. This research presents the goal of all human resource development theories referred to as “selecting the right people for the right positions” (Jing, 2009). The key to success in the company was the ability to manage employees' capabilities effectively. Matching the right jobs with excellent employees was a complex process and poses a challenge for managers (Huang & Jiang, 2011).

Several recent studies have highlighted the significant connection between performance metrics and outcomes related to promotion. Jamil et al. (2021) carried out a comprehensive study involving professional football teams in Europe, aiming to identify the key performance indicators (KPIs) that significantly influence promotion to elite leagues. Their analysis encompassed over 11,000 matches and observations concluded that specific technical actions, such as scoring from set plays and effective passing, correlate strongly with promotion success. In particular, performance elements such as goals from corners, penalty goals, and assists were identified as major contributors to increasing the odds of promotion. Although this research is rooted in sports, it emphasizes the general principle that measurable performance indicators can predict upward mobility, a concept relevant in various professional contexts, including human resources.

Similarly, the concluding report on Responsible Research and Innovation (RRI) indicators (Strand et al., 2015) suggested that promotion and evaluation processes should be based on measurable outcomes and evidence-based indicators. The report recommended using sets of indicators that span across all RRI dimensions while focusing on both process and performance. This supports the idea that data-driven frameworks for performance assessment are critical not just for fairness and transparency, but also for improving organizational development strategies.

These findings from various fields revealed a common understanding: that performance evaluation plays a crucial role in assessing eligibility for promotion. Nevertheless, while earlier research suggested that performance was a key factor, there was a lack of studies investigating how machine learning methods could systematically integrate performance data into promotion decisions. This study investigates filling that gap by introducing the Generated Promotion Feature (GPF) and incorporating it into

clustering and classification models, thereby offering a new dimension to performance-based promotion analytics.

2.1.3 Clustering in Human Resource Management

According to the literature review about clustering in human resource management, this research focuses on performance evaluation clustering. This research utilizes fuzzy data mining to cluster various employees' data, thereby improving the efficiency and effectiveness of human resource performance assessment to support managers' decisions. Fuzzy clustering is used to classify similar connections or objects into the same groups. Fuzzy cluster formulas are used to calculate the relationship between records. The experiment is separated into four processes. The first process is the data standardization process, designed to calculate the mean and standard deviation. The second process is finding the correlation coefficient by using a fuzzy similar matrix R . The third process is clustering analysis, which adopts the maximal tree method. The last process is the prediction and determination of instances. The results of clustering were divided into four clusters, where A refers to 'better', B refers to 'general', C refers to 'worse', and D refers to 'best' (Jing, 2009). This research employs a combination technique, comprising gene expression programming (GEP) and iterative self-organizing fuzzy clustering (fuzzy ISODATA clustering), to enhance the accuracy of clustering and the convergence speed in human resource management. Before implementing the clustering process, the input data was integrated, and any unnecessary data was removed. The data was then converted for implementation in the clustering process. After completing the clustering process, the results could be interpreted in sentence form, which explained the performance of each employee cluster (Huang & Jiang, 2011). The research focused on evaluating the quality of recruitment in college by using a novel grey clustering based on the standard triangular haptization weight function. The clustering results divided each attribute related to the quality of human resources into four groups: Excellent, Good, Medium, and Bad (Qian, 2013). Additionally, the SOM algorithm was applied to cluster and identify the correct character and existing problems in human resource management within the college. In college, human resources were divided into five groups: service and other, administration and teaching assistant, teaching, research, and teaching and research.

The results could indicate whether colleges have sufficient human resources in each cluster or a lack thereof (Huang, 2009).

Previous studies have focused on applying machine learning in human resource fields, with related works in this area. Many related studies focused on finding the optimal or expert position allocation, as well as job seekers and staff turnover. Furthermore, this research focuses on selecting the right people in the training and development process. It is based on employee promotions, identifying employees with strong work performance, and developing them for higher positions. In addition, the performance evaluation process features used in the evaluation are significant. Therefore, feature selection is a necessary process for selecting features and applying them with machine learning to evaluate employee performance. Most of the performance evaluations in related works classify performance by focusing on individual information. On the other hand, this research focuses on the performance evaluation by using a clustering process. In summary, previous studies investigated clustering techniques in HRM to segment employee characteristics, assess recruitment quality, and analyze workforce distribution using fuzzy clustering, SOM, and grey clustering approaches. However, most of these studies primarily rely on domain-specific or heuristic methods and lack integration with performance-oriented features using systematic feature engineering. This study builds upon previous studies by presenting a machine learning-driven clustering framework that integrates dimensionality reduction and a novel feature construction approach (GPF), thereby allowing for more precise and understandable performance-driven promotion clustering.

2.1.4 Employee Promotion

Employee promotion serves as a significant mechanism for identifying high-performing individuals and assigning them to positions with greater responsibility and authority. Promotion involved promoting employees to higher roles within the organization, which not only motivates staff but also supported stronger loyalty and increased productivity (Muhannad Ilwani et al., 2023; Alqahtani & Almaleh, 2022; Bagdadli et al., 2006). Additionally, effective promotion strategies have been related to higher levels of employee engagement, a key factor in organizational success. Ensuring fair and well-informed promotion decisions is crucial for enhancing the quality of future

leadership and management within an organization (Liu et al., 2019). Promotion decisions are related to several other HR functions, including compensation planning, performance evaluations, layoffs, and recruitment strategies. Poorly executed promotion practices or a lack of clear career pathways could lead to low organizational commitment, reduced job and career satisfaction, increased likelihood of turnover, absenteeism, and employee disengagement (Bagdadli et al., 2006). To address these challenges, HR professionals must be equipped with the skills and tools to develop and apply objective, data-driven promotion criteria that minimize the risk of bias (Gathungu et al., 2015). Considering professional categories and hierarchical structure was also necessary to reduce misclassification and ensure promotions were aligned with organizational needs (Dias da Silva & van der Klaauw, 2006). However, traditional promotion techniques often rely heavily on subjective evaluations by supervisors. These manual processes could be flawed due to human bias, favoritism, or incomplete performance assessments, which might unfairly hinder an employee's advancement opportunities (Muhannad Ilwani et al., 2023; Alqahtani & Almaleh, 2022). In larger organizations, the volume and complexity of HR data can be overwhelming, making it difficult for HR staff to extract meaningful insights. Therefore, leveraging advanced data analytics technologies became crucial to support evidence-based decision-making, formulate strategic talent development plans, and ensure that promotion decisions are both fair and effective (Huang, 2009). The present study addresses the limitations of traditional, subjective promotion systems by leveraging data-driven machine learning techniques to support more transparent and consistent promotional decisions. By incorporating structured performance metrics and techniques, such as the Generated Promotion Feature (GPF), this research introduces a systematic framework designed to minimize bias and enhance the fairness and accuracy of promotion assessments in complex HR settings.

2.1.5 Employee Promotion Model with Machine Learning Methods

Machine learning (ML) is a branch of artificial intelligence (AI) that allows computers to learn patterns from data and make decisions or predictions without requiring specific programming instructions. In the context of human resource management, ML has been widely adopted to analyze employee data and support various HR-related tasks (Huang, 2009; Silva & Krohling, 2018). Typically, machine

learning applications for employee promotion modeling can be categorized into two main categories: supervised learning and unsupervised learning. Supervised learning is commonly used for classification tasks, where models are trained on labeled data to predict outcomes, such as promotions. Several previous studies have applied supervised algorithms such as Logistic Regression (Muhannad Ilwani et al., 2023), Random Forest (Liu et al., 2019; Sahinbas, 2022), Decision Tree (Kaewwiset & Temdee, 2022), and Gradient Boosting (Alqahtani & Almaleh, 2022) to develop predictive models for employee promotions. On the other hand, unsupervised learning is frequently applied for clustering tasks, where the goal is to identify inherent groupings within the data without relying on predefined labels. Unsupervised models are particularly valuable in HR analytics for identifying hidden patterns and segmenting employees based on shared characteristics. Notably, many studies have applied Fuzzy Clustering (Liu, 2021; Ouyang & Ge, 2020; Dang et al., 2021; Sun et al., 2022; Wang, 2021) and K-means Clustering (Seyed Alireza Mousavian et al., 2021; Bruna Villa Todeschini et al., 2016; Liu et al., 2023; Sun & Li, 2019; Zhao, 2020; Sarker et al., 2018) for clustering HR-related data. Clustering approaches align closely with the objectives of promotion modeling, as they focus on grouping individuals based on similarity in characteristics or performance, regardless of their original roles or positions. This similarity-based grouping is beneficial for identifying employees who are promotable across different contexts. For these reasons, the present study applies an unsupervised learning approach to develop an employee promotion model. In particular, the study applies two widely used clustering algorithms: K-means clustering and Fuzzy clustering, both of which are well-suited to capture structural patterns within HR datasets. Moreover, the studies apply both supervised and unsupervised learning for employee promotion analysis. The present study contributes by integrating clustering techniques, specifically K-means and Fuzzy clustering, into a unified framework enriched with engineered performance features. This methodology improves the identification of employees eligible for promotion within unlabeled data, so extending current methods with greater interpretability and relevance across various HR contexts.

2.1.6 Features for Promotion Model

In the development of promotion prediction models, two major types of features are commonly employed: personal features and performance-oriented features.

Personal features refer to background information about an employee, including aspects such as age, gender, educational background, marital status, state of origin, recruitment channel, region, foreign education status, and department affiliation. These attributes provide demographic and contextual insights into each employee's profile. On the other hand, performance-oriented features focus on assessing an employee's effectiveness and contributions in the workplace. These generally include Key Performance Indicators (KPIs), performance ratings from previous years, training scores, awards received, and a composite performance score that reflects overall achievement. Various studies emphasize different types of features when constructing promotion models. For example, some research has relied on personal attributes such as age, gender, and education to investigate promotion patterns (Liu et al., 2019). Other studies have focused exclusively on performance-based indicators, particularly performance and potential metrics, to inform promotion decisions (Bagdadli et al., 2006). Additionally, several studies combined both personal and performance-oriented features to provide a more holistic view (Alqahtani & Almaleh, 2022; Long et al., 2018). In this study, both personal and performance-oriented features are incorporated into the model to create a comprehensive framework for clustering. However, the design of the model provides a significant emphasis on performance-oriented features, based on the rationale that work performance should carry more weight in promotion decisions than demographic or background characteristics. This prioritization aligns with promoting fairness and merit-based advancement within the organization. This study expands on previous research by integrating both personal and performance-oriented features, with a particular emphasis on performance metrics to support merit-based promotion. By emphasizing measurable outcomes such as KPIs, training scores, and awards, the suggested framework aligns with modern HR strategies to improve the fairness and objectivity in promotion decisions within data-driven settings.

2.1.7 Feature Engineering

Feature engineering is regarded as a fundamental step in the machine learning pipeline, with the primary objective of improving model performance by creating new, informative variables from the original dataset. These feature engineering techniques are integrated into learning algorithms to help reveal hidden patterns, thereby enhancing predictive accuracy and model reliability. Various studies have applied feature

engineering to both supervised and unsupervised learning tasks, demonstrating its impact across HR-related applications such as using data fusion for attendance monitoring (Wu & Shen, 2023), feature fusion for job matching (He et al., 2022), and data generation with GANs to enhance model accuracy (Hatanaka & Nishi, 2021).

Two notable techniques within feature engineering are feature extraction and feature augmentation, which gained significant attention in recent years due to their role in enhancing the manageability and quality of input data. Feature extraction focuses on reducing the dimensionality of datasets by selecting and transforming only the most relevant features. Among the most used techniques are Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), which represent linear and nonlinear dimensionality reduction approaches, respectively.

PCA is particularly effective in maintaining the overall variance of the data and transforming correlated variables into a set of uncorrelated components. This approach was applied in various HR-related contexts such as KPI weight analysis (Lai & Wei, 2007), vendor evaluations (Xu, 2010), training assessment (Sun & Zhao, 2011), and enterprise performance analysis (Qi & Sun, 2011). Additionally, PCA was often employed as a preprocessing step in classification and clustering models (Kaewwiset & Temdee, 2022), including applications that combine PCA with K-means clustering (Sun & Li, 2019). It was performed by examining covariance matrices to assess the significance of variables and eliminate redundancy, resulting in objective, data-driven weights (Peng et al., 2023; Guo & Yi, 2010).

In contrast, t-SNE was responsible for maintaining local relationships by maintaining data points during projection to lower dimensions. It is beneficial in exploratory analysis, visualization, and discovering patterns within high-dimensional datasets. Its ability to reveal subtle clustering structures has been successfully applied in large-sample HR contexts (Chan et al., 2018; Yu et al., 2017; Feng et al., 2020). In this research, both PCA and t-SNE are explored and compared to identify the more effective dimensionality reduction method for enhancing clustering performance in employee promotion modeling. Another important aspect of feature engineering was the feature augmentation, which involved expanding the feature space by generating new variables derived from existing ones (Petkov et al., 2012; Duan et al., 2018). This technique enhances model robustness by generating additional training data through

systematic transformations of current features. Feature creation, a subcategory of augmentation, focuses on developing new variables that show a significant relationship with the target outcome.

In this research, a novel feature named Generated Promotion Feature (GPF) is introduced based on the assumption that performance-related variables play a more significant role in promotion decisions than demographic characteristics. GPF is developed by combining the top performance-oriented features that show the strongest correlation with promotion status. This additional variable is added to the dataset after feature extraction, with the expectation that it will enhance the clustering model's capability to identify employees who are eligible for promotion.

Although previous studies have explored employee promotion using classification techniques or multi-criteria decision-making (e.g., fuzzy DEMATEL, AHP), they often relied on expert judgment or qualitative assessments, which can introduce subjectivity and lack reproducibility. In addition, while recent studies have integrated ML techniques, they primarily focused on direct prediction from raw features without addressing data imbalance or the interpretability of performance metrics (Alqahtani & Almaleh, 2022). This study extends previous work by introducing the Generated Promotion Feature (GPF), which is a derived feature based on top-correlated performance indicators that enhances both interpretability and clustering/classification performance. The GPF construction is informed by Pearson correlation, under the assumption that features highly correlated with promotion outcomes carry predictive importance. This is based on the filter method in feature selection theory, where statistical dependence between features and target variables is used as a selection criterion. Mathematically, GPF is a linear combination of binary decisions based on whether specific performance metrics meet defined thresholds. This connects to the idea of scoring functions in statistical learning, where a composite score is created to reflect hidden traits (e.g., qualification for promotion) from observable variables.

2.1.8 Imbalanced Data Management

An imbalanced dataset refers to a scenario where the number of instances across different classes is not evenly distributed—typically, one class (the majority class) has a considerably larger number of instances compared to another (the minority class). This imbalance often results in biased classification models that demonstrate high

overall accuracy but fail to predict instances from the underrepresented class accurately. In such cases, the model typically prioritizes the majority class and overlooks the minority class, resulting in inadequate performance in practical applications where identifying minority instances is crucial. To address this challenge, two main categories of techniques are commonly employed: oversampling and undersampling (Tallo & Musdholifah, 2018). Over-sampling increases the representation of the minority class by generating additional instances, thereby preserving the original dataset's structure. In contrast, under-sampling reduces the number of instances in the majority class, which can help balance the dataset but may also result in the loss of valuable information if not applied carefully. A commonly used oversampling approach is the Synthetic Minority Oversampling Technique (SMOTE), which aims to enhance class balance by creating synthetic instances for the minority class (Li & Zhou, 2019). SMOTE works by interpolating between existing instances of the minority class and their nearest neighbors. This process creates new, artificial data points that represent plausible but unseen examples within the minority class. By expanding the representation of this class, SMOTE enhances the model's ability to learn its characteristics, which in turn improves recall, precision, and the F1-score for predictions related to the minority class. SMOTE has been effectively applied in various fields, including employee promotion modeling (Liu et al., 2019) and clustering model construction (Xuan et al., 2013; Wang et al., 2020), demonstrating its adaptability and strength in managing imbalanced datasets.

In this study, SMOTE was selected as the primary strategy to address the class imbalance present in the two publicly available HR datasets. Since traditional accuracy metrics can be misleading when working with imbalanced data, it is crucial to focus on performance measures that more accurately reflect the model's capability on the minority class. SMOTE tackles class imbalance using a resampling-based approach, rather than cost-sensitive training. The technique adopted a K-nearest neighbors (KNN) algorithm to identify the nearest minority instances and interpolate new data points between them. The synthetic instances are then added to the dataset, thereby enlarging the minority class and enabling the classifier to generalize patterns from both classes better. As a result, classifiers trained on SMOTE-enhanced data typically achieve more

balanced performance, especially in recognizing instances from the underrepresented class (Tallo & Musdholifah, 2018).

2.2 Background Theory

In this study, the data preprocessing stage is given significant emphasis as a foundational component of the modeling pipeline. The theoretical background explored includes essential techniques in data preparation, particularly feature extraction methods such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), which are utilized to reduce dimensionality and reveal latent structures in high-dimensional HR datasets. Furthermore, this research addresses the issue of class imbalance, a common challenge in promotion prediction tasks, by applying the Synthetic Minority Oversampling Technique (SMOTE). These preprocessing strategies are employed prior to model construction, which involves both supervised learning (classification) and unsupervised learning (clustering), to improve model performance and generalizability.

2.2.1 Feature Extraction

To address the challenges posed by high-dimensional and complex HR datasets, this study incorporates feature extraction techniques as a crucial component of the data preprocessing pipeline. Specifically, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are employed to transform and reduce the dimensionality of the input space while maintaining meaningful structure. These two techniques represent fundamentally different approaches to dimensionality reduction: PCA is a linear projection method that emphasizes global variance, and t-SNE is a nonlinear method that focuses on preserving local neighborhood relationships.

The inclusion of both PCA and t-SNE in the experimental design enables a comprehensive evaluation of how linear versus nonlinear transformations impact the performance of clustering and classification models. By reducing redundancy and noise in the data, these techniques are expected to enhance model generalization, improve interpretability, and reveal the hidden patterns in employee performance data that are critical for promotion modeling.

2.2.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a feature extraction method using an unsupervised algorithm. It is used to reduce the dimensionality of the data. Linear algebra and statistics are utilized in PCA calculations to identify highly variant and highly correlated outputs, and to rearrange the features through a linear transformation, thereby creating new variables in a simple matrix (Syafrudin et al., 2020). The first feature of PCA is characterized by high variance and captures the most information about the dataset. The second feature is more informative and has more considerable variance than the third, and so on. The steps of PCA are as follows:

1. Normalized features by Standardize.
2. Covariance matrix calculation.
3. Finding the eigenvalues and eigenvectors for the covariance matrix.
4. Plot the vectors on the scaled data.

2.2.1.2 t-distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique, primarily designed to visualize high-dimensional data by mapping it into a lower-dimensional space (typically two or three dimensions). Created by van der Maaten and Hinton, t-SNE transforms the pairwise similarities of data points into joint probabilities and reduces the divergence between these probabilities in both high-dimensional and low-dimensional spaces. Unlike linear methods such as Principal Component Analysis (PCA), t-SNE effectively preserves local structures and captures intricate non-linear relationships among features, making it a powerful tool for exploratory data analysis and pattern discovery in complex datasets. In addition, t-SNE begins by calculating a conditional probability that a data point would select another data point as its neighbor within the high-dimensional space. It then strives to find a low-dimensional representation that retains these neighbor relationships by employing a student's t-distribution to assess similarity. This approach addresses the "crowding problem" and yields more meaningful visualizations resembling clusters.

t-SNE has been applied in various domains of human resource analytics. For example, it has been employed in clustering analysis on extensive HR datasets to maintain significant groupings and reveal hidden patterns (Chan et al., 2018; Feng et al., 2020; Yu et al., 2017). In these studies, t-SNE effectively revealed internal clusters

of employees or performance groups, which could be further used for promotion analysis, workforce segmentation, or anomaly detection.

2.2.2 Data Unbalancing

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002; Chawla, 2010) is a data preprocessing technique used to manage imbalanced data. Generally, the machine learning performance is challenged by the imbalanced data. The imbalanced data occurs because of an unequal distribution of classes in a dataset, leading to the incorrect choice of distribution when creating a model, as the majority class is more extensive than the minority class. The imbalanced data can be addressed in two ways. Firstly, it can be addressed by assigning distinct costs to training examples. Secondly, it can be addressed by resampling the original dataset.

SMOTE processes by creating artificial objects of a minority class. The oversampling of the minority class requires the use of k-nearest neighbors (Tarusov & Mitrofanova, 2019) at random. The identification of a sample from the nearest neighbor of the minority class is calculated using Euclidean distance, as demonstrated in Equation (2.1).

Euclidean distance equation

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \quad (2.1)$$

Where

1. $d(x, y)$ is Euclidean distance between one minority data to another minority data.
2. x and y is minority data.
3. n is the maximum number of attributes.

Then, sample data are generated between two minority data by using the linear interpolation formula.

Linear interpolation formula

$$z = x_n + \text{random}(0,1) \times (x_n - y_n) \quad (2.2)$$

Where

1. z is synthetic data.
2. x and y is minority data.
3. n is the maximum number of attributes.
4. $random(0,1)$ is a random number between 0 and 1.

2.2.3 Machine Learning Classification

To build reliable and interpretable models for predicting employee promotions, this research adopted six classification algorithms: Decision Tree, Random Forest, SVM, Logistic Regression, KNN, and Neural Network. These classifiers were intentionally chosen to encompass a wide variety of learning approaches. Additionally, each model has unique features, including rule-based learning (Decision Tree), ensemble methods (Random Forest), margin-based optimization (SVM), probabilistic models (Logistic Regression), instance-based learning (KNN), and deep learning strategies (Neural Network). With these varied algorithms, the study aims to comprehensively evaluate the effectiveness of the proposed preprocessing strategies and feature engineering techniques across different modeling philosophies, ensuring generalizability and robustness of the classification outcomes.

2.2.3.1 Decision Tree

A Decision Tree is a predictive model represented as a flowchart in a tree format, utilized for establishing decision rules and classifying subjects into several groups. The tree structure consists of a root node, branches, inner nodes, and leaf nodes. The root node is an attribute selection at the top node. The branch is an object that meets the node condition. The leaf node is a class or a prediction result of an object. Additionally, it can be presented in two types. The first one presents discrete values, known as a classification tree, and the second one presents continuous values, known as a regression tree. The Decision Tree is widely used in human resource management, such as information security risk analysis of human resources (Eminagaoglu & Eren, 2010), and VARK learning style analysis with physiological signals (Dutsinma & Temdee, 2020).

2.2.3.2 Random Forest

Random Forest is a widely used classifier that operates on the principle of decision trees by combining multiple trees instead of relying on a single tree for

classification. Each tree is constructed from the original training sample randomly and reversibly. The majority of the votes of the decision tree are selected as a Random Forest model. The random forest is used in many classification applications, such as learner and professional development (Asim et al., 2018; Guohao et al., 2019), skilled job position replacements (Mathew, Chacko, & Udhayakumar, 2018), and evaluating the risk of employee turnover (Tarusov & Mitrofanova, 2019).

2.2.3.3 Support Vector Machine

Support Vector Machine (SVM) is a classification method that assigns classes by separating each class through a decision boundary where the data points are separated by a line, called linear SVM, and a hyperplane, called non-linear SVM. Two sides of a hyperplane separate the dataset into two classes. When new input data is predicted for either of the two, the margin between the hyperplane and the support vector will be significant in reducing the error in the classification model. For example, replacement in skilled job positions (Mathew et al., 2018), text classification of competency and professional learning (Aottiwerch & Kokaew, 2018; Adnan et al., 2020), improving performance of learners (Guohao et al., 2019), feature selection for human resources management (Wang, Li, & Hu, 2009).

2.2.3.4 Logistic Regression

Logistic Regression is a widely used statistical model for binary classification tasks, which estimates the probability that a given input belongs to a specific category. It is particularly suitable for scenarios where the dependent variable is dichotomous (e.g., promoted vs. not promoted). The model operates by applying the logistic (sigmoid) function to a linear combination of input features, resulting in an output bounded between 0 and 1. The threshold (commonly 0.5) is then used to determine the final class label. Logistic Regression is valued for its simplicity, interpretability, and efficiency, making it a preferred baseline model in many human resource analytics applications.

In HRM contexts, Logistic Regression has been applied to analyze factors influencing promotion decisions (Ilwani et al., 2023), predict employee turnover, and examine job satisfaction levels. It has also been used in workforce planning and identifying high-potential employees based on performance indicators and demographic variables. For example, Ilwani et al. (2023) conducted a study using

Logistic Regression to assess promotion trends within organizational data, emphasizing the significance of using balanced datasets and performance-related features to enhance the accuracy of predictions. Though it has a linear framework, Logistic Regression continues to be an effective tool when paired with feature engineering and balancing methods like SMOTE to address class imbalance and improve generalizability.

2.2.3.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a non-parametric, instance-based learning algorithm used for both classification and regression tasks. In classification, the algorithm assigns a class label to a new data point based on the majority class among its k nearest neighbors in the feature space. The distance between points is typically measured using the Euclidean distance, although other metrics, such as Manhattan or Minkowski distance, may also be applied depending on the context. The selection of the value for k plays a crucial role in the model's effectiveness—a minimal k may result in overfitting, whereas a k that is too large may lead to underfitting.

KNN does not build an explicit model during training, which makes it computationally efficient for small datasets but potentially expensive during prediction for large datasets. Its simplicity and effectiveness make it suitable for HR analytics, particularly in tasks involving similarity-based reasoning, such as employee clustering or predicting the possibility of promotion.

In human resource contexts, KNN has been applied in various settings, including job seeker classification using Twitter data (Hartanto et al., 2019) and skill or competency matching based on behavioral profiles. Its ability to model decision boundaries based on historical similarity enables HR managers to make data-driven decisions without assuming any prior distribution of features. Although sensitive to data imbalance, KNN's performance can be significantly improved when combined with data preprocessing techniques such as SMOTE for balancing and PCA or t-SNE for dimensionality reduction.

2.2.3.6 Neural Network

Artificial Neural Network (ANN) is a computational model inspired by the structure and functioning of biological neural networks. It consists of interconnected layers of nodes, referred to as neurons, including an input layer, one or more hidden layers, and an output layer. Each neuron processes weighted input and passes the result

through an activation function, enabling the network to learn complex, non-linear relationships between input features and target outputs.

In classification tasks, artificial neural networks (ANNs) learn to differentiate between classes through a training method called backpropagation, where discrepancies between predicted and actual values are propagated backward through the network to modify the weights. The ability of ANN to model intricate decision boundaries makes it particularly useful in domains involving high-dimensional and noisy data, such as human resource analytics.

In human resource management, Neural Networks have been used for various tasks, including job recommendations through feature fusion (He et al., 2022), employee performance prediction (Liu et al., 2021), and enhancement of employee classification accuracy using deep learning (Muhammad et al., 2020). These studies have demonstrated the effectiveness of ANN in identifying hidden patterns within performance-related data, thus aiding decisions regarding employee growth and advancement.

Although Neural Networks have significant capabilities, they demand a large amount of training data and can be affected by factors such as class imbalance. As a result, they are frequently used alongside data augmentation or balancing strategies like SMOTE, as well as dimensionality reduction techniques such as PCA or t-SNE to enhance stability and understanding in real-world HR applications.

2.2.4 Machine Learning Clustering

To investigate unsupervised learning approaches for identifying promotable employees, this study employs clustering techniques that group individuals based on similarities in feature space without relying on labeled outcomes. Clustering is particularly useful in human resource analytics for discovering latent patterns, segmenting employee populations, and supporting data-driven decision-making where class labels may not be predefined or consistently available. Moreover, two prominent clustering methods are adopted: K-Means Clustering and Fuzzy Clustering. These techniques are selected to capture both hard and soft clustering paradigms. K-Means assigns each data point to a single cluster based on distance minimization, and Fuzzy Clustering allows partial membership in multiple clusters, better reflecting the complexity and overlap inherent in human characteristics. By evaluating both methods

on enhanced HR datasets improved through feature engineering and data balancing, this research aims to assess the effectiveness of clustering models in supporting promotion analysis and segmentation within organizational environments.

2.2.4.1 K-Means Clustering

K-Means is an unsupervised learning algorithm that divides a dataset into K unique, non-overlapping groups based on similarities in features. It aims to minimize the within-cluster variance by assigning each data point to the cluster with the nearest mean (centroid). The algorithm operates iteratively by initializing K centroids, assigning data points to the nearest centroid, and recalculating the centroids until convergence is achieved.

K-Means is particularly effective for discovering hidden structures in large datasets without prior labeling. Despite its simplicity and efficiency, it is influenced by the initial selection of centroids and the predetermined number of clusters (K). Therefore, techniques such as PCA are often used to reduce dimensionality before applying K-Means, which improves clustering performance and visualization.

In the field of human resource management, K-Means clustering has been extensively used for categorizing employee profiles, evaluating performance metrics, and supporting strategic decision-making. For example, Zhao (2020) applied K-Means to improve informatization in HR datasets, while Sun and Li (2019) used PCA in conjunction with K-Means to enhance talent clustering in enterprise contexts. Furthermore, Sarker et al. (2018) demonstrated the effectiveness of K-Means in identifying performance patterns that assist in promotion and training decisions. These studies confirm the utility of K-Means in extracting actionable insights from multidimensional HR data, making it a valuable tool in data-driven promotion modeling.

2.2.4.2 Fuzzy Clustering

Fuzzy Clustering is an unsupervised machine learning technique that allows each data point to belong to more than one cluster, with varying degrees of membership. Unlike hard clustering algorithms, K-Means, which assign each observation to a single cluster, fuzzy clustering provides a more flexible approach that reflects the uncertainty or overlap often present in real-world data. The most widely used fuzzy clustering method is the Fuzzy C-Means (FCM) algorithm, which optimizes a membership function to minimize intra-cluster variation while allowing partial membership across clusters.

The algorithm operates by iteratively updating the membership degrees and the cluster centroids until convergence is reached. This approach is particularly valuable in human resource contexts, where employee characteristics frequently encompass various roles, competencies, or performance levels. By assigning probabilistic membership to different groups, fuzzy clustering captures the nuanced relationships among HR variables more effectively than crisp classification.

Fuzzy clustering has been applied to various HR analytics tasks, including performance evaluation, workforce segmentation, and skill assessment. For example, Jing (2009) applied fuzzy data mining to categorize employee performance into different quality tiers, while Huang and Jiang (2011) developed a fuzzy ISODATA clustering framework using gene expression programming to enhance convergence speed in HR analysis. Additionally, Qian (2013) proposed a fuzzy-based assessment using triangular Whitenization weight functions to classify HR data into levels of excellence. These works demonstrate the suitability of fuzzy clustering in modeling ambiguous and overlapping characteristics within human resource datasets, providing a more realistic representation of employee profiles for tasks such as promotion modeling.

2.3 The Proposed Work

This study proposes a comprehensive and structured machine learning framework to support employee promotion analysis, addressing critical limitations in traditional promotion decision-making, including subjectivity, data imbalance, and high-dimensional feature spaces. The proposed approach incorporates both supervised and unsupervised learning techniques to enable predictive modeling and pattern discovery from HR datasets.

The core of this framework lies in three components. Firstly, the feature augmentation is applied through the development of a novel domain-informed variable called the Generated Promotion Feature (GPF). GPF is constructed by aggregating multiple high-correlation performance-related attributes such as key performance index (KPI) scores, award history, and average training results into a single interpretable

numerical value. This feature aims to provide greater clarity on an employee's promotability than individual metrics alone.

Secondly, feature extraction is conducted to reduce the dataset's dimensionality while preserving its meaningful structures. Both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are utilized to evaluate their effectiveness in improving model performance. PCA is expected to reveal global variance structures, while t-SNE is anticipated to enhance local data relationships.

Lastly, to mitigate the challenge of class imbalance inherent in promotion datasets, the Synthetic Minority Oversampling Technique (SMOTE) is employed. SMOTE generates synthetic data points for the underrepresented class (i.e., promoted employees) based on interpolation, thereby balancing the class distribution and enhancing the model's generalizability.

The enriched datasets—combinations of original features, GPF, PCA/t-SNE, and SMOTE—are used to train and evaluate both clustering models (K-means and Fuzzy C-means) and classification models (Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, and Neural Network). The performance of each model is assessed using appropriate metrics, including the Rand Index (RI), Mutual Information (MI), V-measure, and Fowlkes-Mallows Index (FMI) for clustering, as well as Accuracy, Precision, Recall, and F1-score for classification.

This proposed work is expected to enhance the fairness, interpretability, and predictive capabilities of promotion decision systems, enabling HR professionals to make more data-driven and transparent decisions across various organizational settings.

CHAPTER 3

METHODOLOGY

3.1 Method Overview

This research comprises six key methodological stages, as illustrated in Figure 3.1: Data Collection, Data Preprocessing, Feature Engineering, Feature Scaling, Model Construction, and Result Evaluation. Each stage is designed to systematically prepare, transform, and evaluate HR datasets to construct effective models for employee promotion analysis.

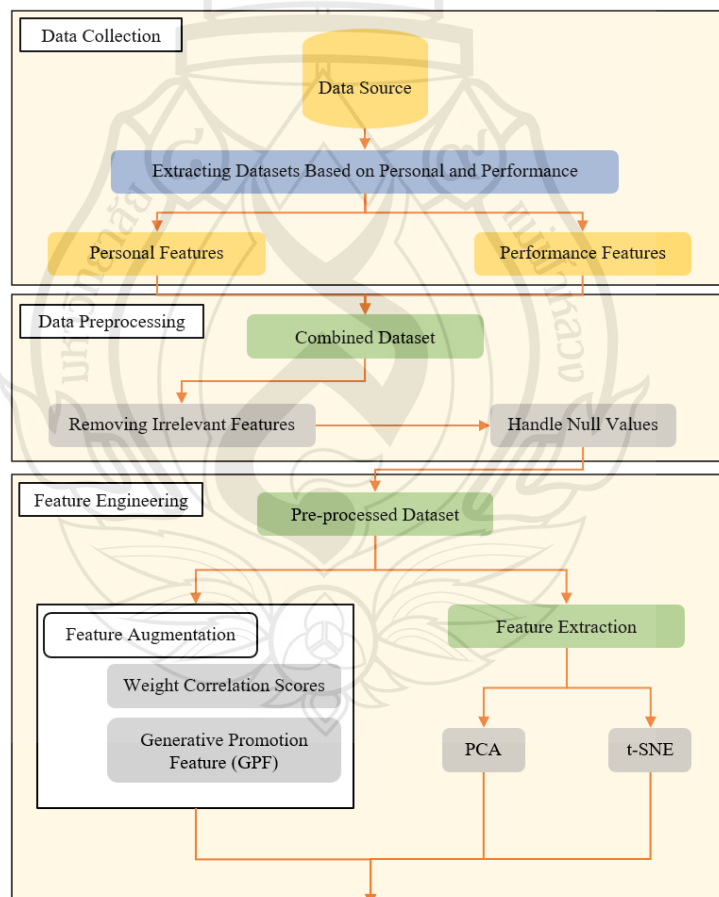


Figure 3.1 Overall Methodology

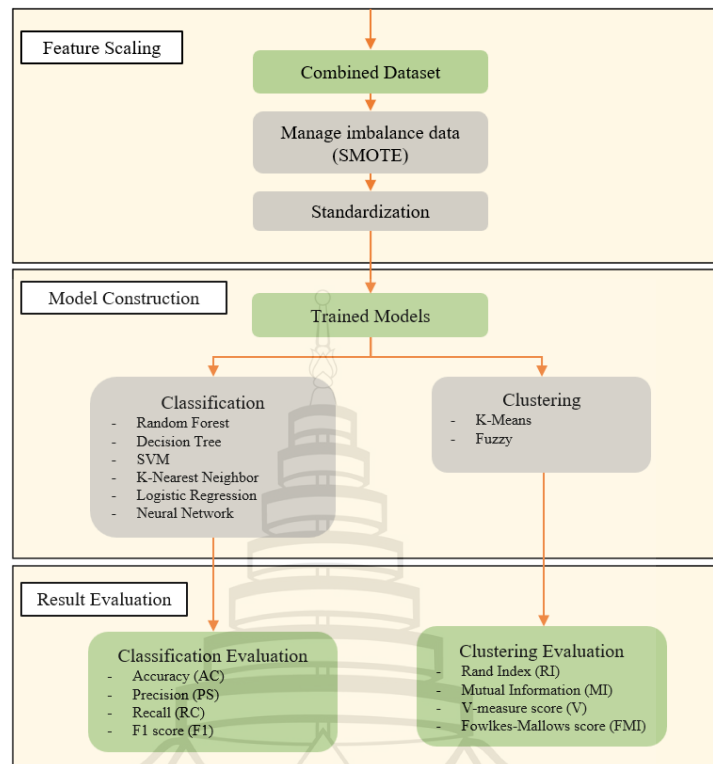


Figure 3.1 (continued)

The details of each stage are as follows:

The Data Collection Stage introduces the two publicly available HR datasets that were selected as the basis for evaluating the effectiveness of the proposed clustering approach for employee promotion analysis. The datasets are described in terms of the total number of records, the distribution of promotion and non-promotion classes, and a detailed explanation of all available fields. The features are categorized into two groups: personal features (e.g., age, education, region) and performance-oriented features (e.g., KPI scores, awards, training scores). The correlation between each feature and promotion status is also examined to understand their relevance.

During the data preprocessing stage, irrelevant fields, such as employee identification numbers, are removed due to their lack of analytical significance. Missing values in both datasets are handled using mode imputation, ensuring that the data is complete and ready for further processing.

The Feature Engineering phase stands out as the key aspect of the proposed methodology. Feature Engineering consists of two main techniques: feature augmentation and feature extraction. A novel feature, called the Generated Promotion Feature (GPF), is created using performance-oriented variables that exhibit a high correlation with promotion status. To determine which variables are most closely linked to promotion outcomes, a correlation heatmap was created. Features that exhibited high correlation with the promotion label were selected to construct a new feature set known as the Generated Promotion Feature (GPF). In addition, two dimensionality reduction techniques, which are Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), are applied to extract informative features that prepare the datasets for more effective model construction. These techniques helped simplify the data while maintaining its most informative patterns. The transformed data from the extraction process was then combined with GPF to form an enhanced dataset.

The Feature Scaling (Imbalanced Data Handling) addresses the severe imbalance present in HR promotion data, where the number of promoted employees is significantly lower than that of non-promoted ones. To handle this, Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic instances of the minority class, thereby balancing the dataset and reducing bias in model training.

Model Construction focuses on evaluating the suitability of various models on the enriched datasets. Both classification models (such as Random Forest, SVM, and Logistic Regression) and clustering models (such as K-means and Fuzzy Clustering) are developed and evaluated to identify the most effective method for predicting employee promotions and clustering.

The final stage (Result Evaluation) evaluates the performance of the constructed models. For classification models, standard metrics such as Accuracy, Precision, Recall, and F1-score are used. For clustering models, evaluation is conducted using clustering-specific metrics, including Rand Index (RI), Mutual Information (MI), V-measure, and Fowlkes–Mallows Index (FMI), to assess the quality and consistency of the clusters formed. The overall methodology is illustrated in Figure 3.1.

3.2 Data Collection

Two datasets from Kaggle, containing information on human resources and employee promotions, were analyzed in this study: Dataset 1, HR Analysis Case Study Dataset (Kumar, 2020), and Dataset 2, Data Science Staff Promotion Prediction (Sulaiman, 2019). These HR datasets were used to evaluate the generalization of the proposed features and the employee promotion clustering model.

3.2.1 HR Analysis Case Study Dataset

This dataset is an open data Human Resource Analysis Case Study dataset from Kaggle. The dataset comprises 54,808 records and 14 columns, as follows.

The dataset contains 12 inputs and 1 output for classification. One of them is eliminated because it is unnecessary for classification, which is the employee ID. The input data included department, region, education, gender, recruitment channel, number of trainings, age, previous year rating, length of service, KPI, awards, and average training score. The output data indicates the promotion status. The promotion status is a data field with two possible statuses, including 0 for non-promotion and 1 for promotion.

Table 3.1 Attribute Description of HR Analysis Case Study Dataset

Features	Category	Description
Employee ID	Personal	Employee ID ranges: 1–78298
Department	Personal	Company department: analytics, finance, HR, legal, operations, procurement, R&D, sales & marketing, technology
Region	Personal	Region ranges: region_1–region_34.
Education	Personal	Education level: bachelor, below secondary, master & above
Gender	Personal	Gender: male, female
Recruitment channel	Personal	Recruitment channel: referred, sourcing, other
No of trainings	Performance	Number of trainings: 1–9
Age	Personal	Age of employee: 20–60
Previous year rating	Performance	Previous year rating: 1–5
Length of service	Performance	Time spent by a worker: 1–34
KPI	Performance	KPI score: more than 80% (1), less than 80% (0)
Awards won	Performance	Award winning: received awards (1), did not receive awards (0)
Average training score	Performance	Average training score: 0–100

3.2.2 Data Science Staff Promotion Prediction Dataset

The Data Science Staff Promotion Prediction dataset is open data from Kaggle. The dataset has 38,312 records and 19 columns as follows.

The dataset contains 17 inputs and 1 output for classification, and one of them is eliminated because it is unnecessary for classification, which is the employee number. The input data includes division, qualification, gender, channel of recruitment, training attended, year of birth, last performance score, year of recruitment, targets met, previous award, training score average, state of origin, foreign schooled, marital status, past disciplinary action, previous intradepartmental movement, and no of previous employers. The output data indicates the promotion status. The promotion status includes two possible statuses: 0 for non-promotion and 1 for promotion.

Table 3.2 Attribute Description of Data Science Staff Promotion Prediction Dataset

Features	Category	Description
Employee No	Personal	Employee no ranges: YAK/S/00001–YAK/S/54761
Division	Personal	Company Division: business finance operations, etc.
Qualification	Personal	Qualification level: non-university education, first degree or HND, MSc, MBA, and PhD
Gender	Personal	Gender: male, female
Channel of recruitment	Personal	Recruitment channel: direct internal process, agency and others, referral and special candidates
Trainings attended	Performance	Training attended: 2–11
Year of birth	Personal	Year of birth: 1950–2001
Last performance score	Performance	Last performance score: 0, 2.5, 5, 7.5, 10, 12.5
Year of recruitment	Personal	Year of recruitment: 1982–2018
Targets met	Performance	KPI Target: meet the target (1), did not meet the target (0)
Previous award	Performance	Award winning: received awards (1), did not receive awards (0)
Training score average	Performance	Training score: 31–91
State of origin	Personal	State of origin: KADUNA, PLATEAU, BORNO, etc.
Foreign school	Personal	Foreign school: yes, no
Marital status	Personal	Marital status: single, married, not sure
Past disciplinary action	Performance	Past disciplinary action: yes, no
Previous Intradepartmental movement	Performance	Previous intradepartmental movement: yes, no
No of previous employers	Performance	No of previous employers: 0–5, and more than 5

Dataset 1 consists of 54,808 records and 13 features, with a promotion rate of 8.52% (4,668 records) and 91.48% (50,140 records) classified as not promoted. The proportion of promotion and non-promotion classes is illustrated in Figure 3.2. Dataset 2 contains 38,312 records and 18 features, with a promotion rate of 8.46% (3,241 records) and 91.54% (35,071 records) classified as not promoted. The proportions of both classes are depicted in Figure 3.3.

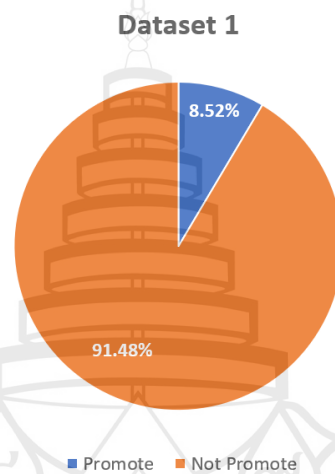


Figure 3.2 Data Proportion of Dataset 1

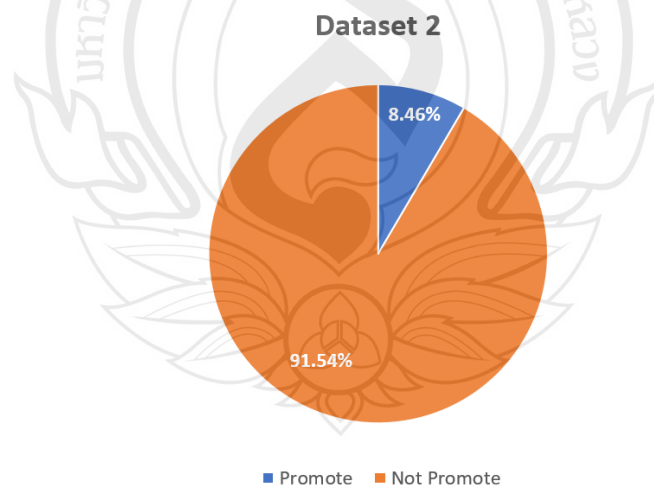


Figure 3.3 Data Proportion of Dataset 2

3.3 Data Preprocessing

As part of the data preprocessing process, employee identification numbers and employee codes were removed because they do not contribute meaningful insights for analysis or model prediction. These fields serve only as unique identifiers and have no direct relationship with the outcome of the promotion.

The next step involved handling missing data, which was represented as null values in the datasets. In Dataset 1, two features, which are education and previous year rating, contained missing values. Specifically, the education feature had 2,409 missing records, while the previous year's rating had 4,124 missing records. To handle this, the most frequent value (mode) in each feature was used to fill the missing data, ensuring consistency without introducing bias. In Dataset 2, the qualification feature was found to have 1,679 missing values, which were filled with the mode value. This approach helps preserve the integrity of the data while avoiding the complications that may arise from data deletion or arbitrary imputation. An example of a dataset before and after processing in the data pre-processing step, as shown in Figure 3.4 and Figure 3.5, respectively.

employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards_won?	avg_training_score	is_promoted
92	Sales & Marketing	region_7	Bachelor's	m	other	1	56	1	5	0	0	50	0
94	Technology	region_17	Bachelor's	m	sourcing	2	26	1	1	1	0	79	0
95	Technology	region_2	Bachelor's	f	other	1	23	1	2	0	0	82	0
96	Sales & Marketing	region_11	Bachelor's	m	sourcing	1	38	3	10	0	0	52	0
98	Technology	region_22	Master's & above	m	other	1	53	1	11	0	0	80	0
99	Analytics	region_28	Bachelor's	m	sourcing	1	30	4	3	1	0	88	0
100	Sales & Marketing	region_28	Master's & above	m	sourcing	2	34	1	6	0	0	50	0
101	Sales & Marketing	region_2	Bachelor's	m	other	2	23	3	2	0	0	49	0
102	Operations	region_7	Bachelor's	m	other	1	51	3	10	0	0	59	0
105	Technology	region_7	Below Secondary	m	other	2	26	3	3	0	0	81	0
106	Operations	region_22	Bachelor's	f	sourcing	1	32	1	7	0	0	56	0
107	Sales & Marketing	region_16	Master's & above	f	other	1	36	2	9	0	1	52	0
108	Analytics	region_15	Bachelor's	m	other	2	31	2	3	0	0	88	0
109	Technology	region_27	Bachelor's	m	sourcing	2	28	3	2	0	0	79	0
110	Operations	region_21	Bachelor's	m	sourcing	1	53	5	8	1	0	60	0
111	Sales & Marketing	region_19	Bachelor's	m	other	1	40	5	9	0	0	55	0
113	Analytics	region_22	Bachelor's	f	other	3	28	5	5	1	0	83	0
114	Technology	region_30	Bachelor's	f	other	1	31	3	5	0	0	81	0
115	Sales & Marketing	region_22	Bachelor's	m	sourcing	1	25	3	2	1	0	46	0
116	Technology	region_7	Master's & above	f	sourcing	1	35	5	10	0	0	83	1
119	Sales & Marketing	region_21	Bachelor's	m	other	1	44	1	7	0	0	46	0
120	Sales & Marketing	region_2	Bachelor's	m	sourcing	1	42	3	6	1	0	51	0
121	Sales & Marketing	region_2	Bachelor's	m	other	1	56	3	15	0	0	51	0
122	Analytics	region_32	Bachelor's	m	sourcing	1	32	5	6	1	0	84	0
124	Operations	region_22	Master's & above	f	sourcing	2	38	3	12	1	0	87	1
125	Operations	region_4	Master's & above	m	other	1	48	3	20	1	0	59	1
127	Technology	region_31	Master's & above	m	other	2	35	3	3	0	0	81	0
128	Procurement	region_28	Bachelor's	m	sourcing	3	29	3	4	0	0	68	0
129	R&D	region_2	Master's & above	m	sourcing	1	38	4	9	0	0	83	0
130	Analytics	region_34	Bachelor's	m	other	2	29	3	2	0	0	88	0
132	Sales & Marketing	region_28	Master's & above	m	other	1	45	2	14	0	0	49	0
133	Technology	region_15	Master's & above	m	other	1	46	3	6	0	0	77	0
134	Procurement	region_20	Bachelor's	m	other	1	34	5	8	1	0	87	1
135	Analytics	region_29	Bachelor's	m	other	1	26	3	2	1	0	84	0
136	Sales & Marketing	region_22	Bachelor's	m	other	3	28	1	2	0	0	50	0

Figure 3.4 Example Dataset Before the Data Pre-processing Step

department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met>80%	awards_won?	avg_training_score	GPF	is_promoted
Sales & Marketing	region_7	Bachelor's	m	other	1	56	1	5	0	0	50	0	0
Technology	region_17	Bachelor's	m	sourcing	2	26	3	1	1	0	79	25	0
Technology	region_2	Bachelor's	f	other	1	23	1	2	0	0	82	0	0
Sales & Marketing	region_11	Bachelor's	m	sourcing	1	38	3	10	0	0	52	0	0
Technology	region_22	Master's & above	m	other	1	53	1	11	0	0	80	0	0
Analytics	region_28	Bachelor's	m	sourcing	1	30	4	3	1	0	88	25	0
Sales & Marketing	region_28	Master's & above	m	sourcing	2	34	1	6	0	0	50	0	0
Sales & Marketing	region_2	Bachelor's	m	other	2	23	3	2	0	0	49	0	0
Operations	region_7	Bachelor's	m	other	1	51	3	10	0	0	59	0	0
Technology	region_7	Below Secondary	m	other	2	26	3	3	0	0	81	0	0
Operations	region_22	Bachelor's	f	sourcing	1	32	1	7	0	0	56	0	0
Sales & Marketing	region_16	Master's & above	f	other	1	36	2	9	0	1	52	25	0
Analytics	region_15	Bachelor's	m	other	2	31	2	3	0	0	88	0	0
Technology	region_27	Bachelor's	m	sourcing	2	28	3	2	0	0	79	0	0
Operations	region_21	Bachelor's	m	sourcing	1	53	5	8	1	0	60	50	0
Sales & Marketing	region_19	Bachelor's	m	other	1	40	5	9	0	0	55	25	0
Analytics	region_22	Bachelor's	f	other	3	28	5	5	1	0	83	50	0
Technology	region_30	Bachelor's	f	other	1	31	3	5	0	0	81	0	0
Sales & Marketing	region_22	Bachelor's	m	sourcing	1	25	3	2	1	0	46	25	0
Technology	region_7	Master's & above	f	sourcing	1	35	5	10	0	0	83	25	1
Sales & Marketing	region_21	Bachelor's	m	other	1	44	1	7	0	0	46	0	0
Sales & Marketing	region_2	Bachelor's	m	sourcing	1	42	3	6	1	0	51	25	0
Sales & Marketing	region_2	Bachelor's	m	other	1	56	3	15	0	0	51	0	0
Analytics	region_32	Bachelor's	m	sourcing	1	32	5	6	1	0	84	50	0
Operations	region_22	Master's & above	f	sourcing	2	38	3	12	1	0	87	25	1
Operations	region_4	Master's & above	m	other	1	48	3	20	1	0	59	25	1
Technology	region_31	Master's & above	m	other	2	35	3	3	0	0	81	0	0
Procurement	region_28	Bachelor's	m	sourcing	3	29	3	4	0	0	68	0	0
R&D	region_2	Master's & above	m	sourcing	1	38	4	9	0	0	83	0	0
Analytics	region_34	Bachelor's	m	other	2	29	3	2	0	0	88	0	0
Sales & Marketing	region_28	Master's & above	m	other	1	45	2	14	0	0	49	0	0
Technology	region_15	Master's & above	m	other	1	46	3	6	0	0	77	0	0
Procurement	region_20	Bachelor's	m	other	1	34	5	8	1	0	87	50	1
Analytics	region_29	Bachelor's	m	other	1	26	3	2	1	0	84	25	0
Sales & Marketing	region_22	Bachelor's	m	other	3	28	1	2	0	0	50	0	0

Figure 3.5 Example Dataset After the Data Pre-processing Step

3.4 Feature Engineering

This step enhances feature dimensions to prepare the dataset for improved modeling performance. The focus of this study is to use GPF. GPF is created based on the strong relationship between performance indicators and promotion outcomes. This method helps simplify the dataset while preserving the most meaningful information, as it combines feature extraction techniques such as PCA and t-SNE to transform the data structure, aiming to improve the overall performance of the models. The following sections provide a detailed explanation of how each process contributes to enriching and optimizing the data for analysis. The conceptual diagram is shown in Figure 3.6.

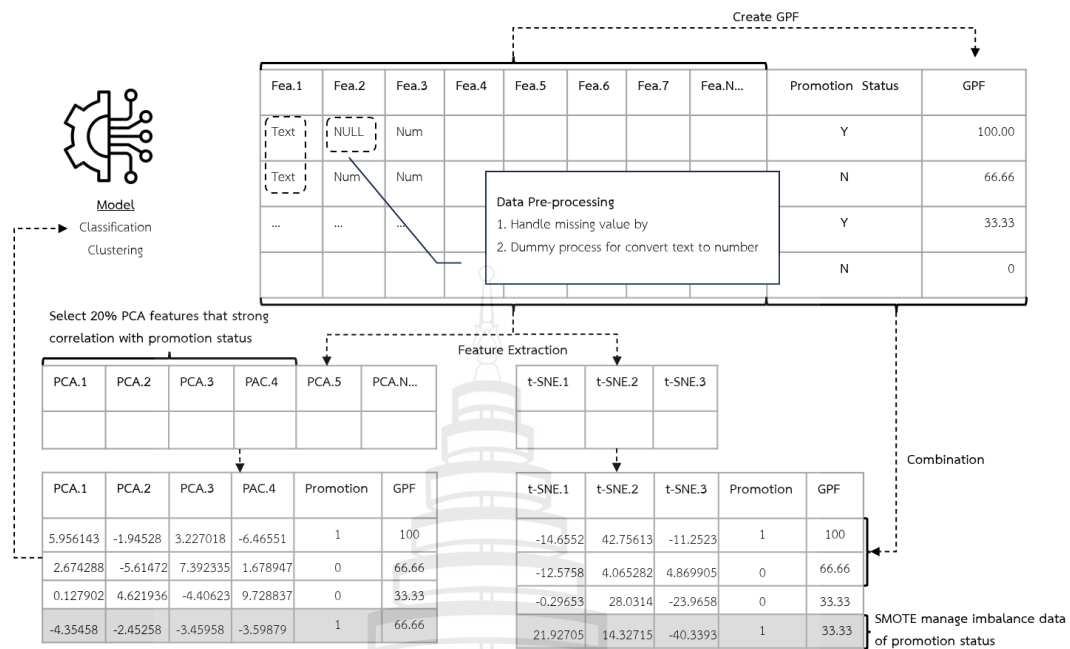


Figure 3.6 Conceptual Diagram of Proposed Framework

3.4.1 Feature Augmentation

The primary objective of this study is to develop a new feature, known as the Generated Promotion Feature (GPF), designed to enhance the effectiveness of models used in employee promotion analysis. The core idea behind GPF is to enrich the dataset with more meaningful, performance-based information that could support machine learning algorithms in both learning and prediction tasks. GPF is carefully designed using a combination of domain expertise and data-driven insights, with an emphasis on prioritizing performance-related attributes over personal characteristics when identifying employees with potential for promotion. To identify the most relevant performance indicators for constructing GPF, a correlation heatmap is generated to measure the strength of the relationship between each feature and the promotion status. The three features with the highest correlation scores limited to performance-oriented dimensions were selected to form the GPF.

In Dataset 1, the selected features include “KPI”, “Awards won”, and “Average training score”, as illustrated in Figure 3.7. In Dataset 2, the top features corresponding to the targets are “Targets met,” “Previous award,” and “Training score average,” as

shown in Figure 3.8. These features serve as the foundation for generating GPF, which is integrated into the dataset to support more insightful clustering analysis.

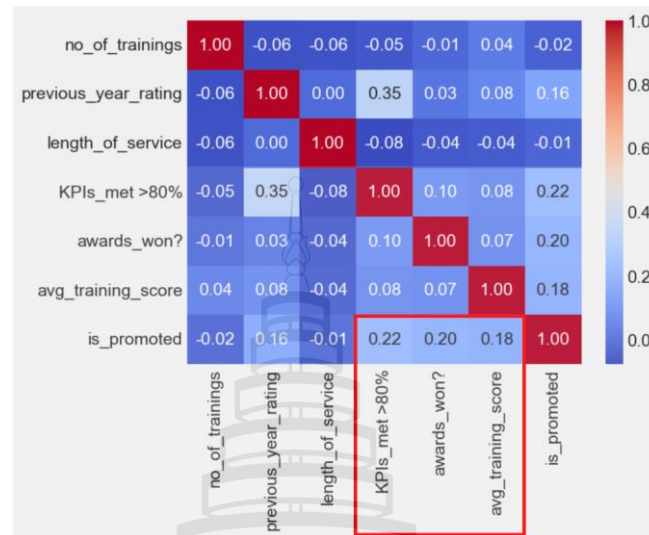


Figure 3.7 Performance-Oriented Feature Correlation Dataset 1

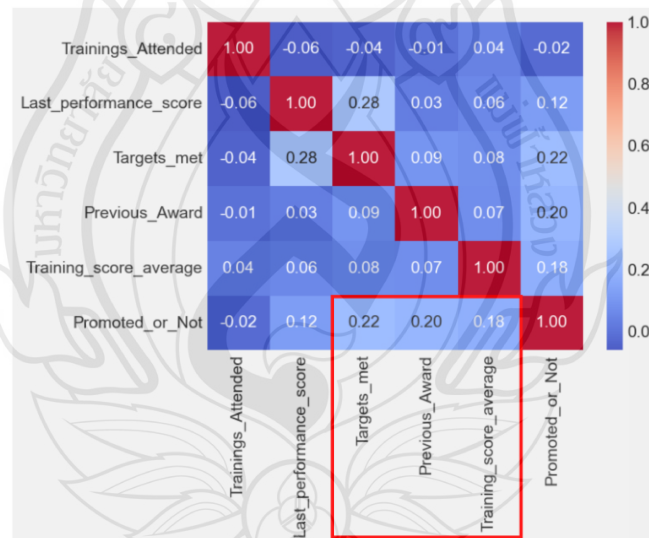


Figure 3.8 Performance-Oriented Feature Correlation Dataset 2

The construction of GPF is detailed in Algorithm 1, as follows.

Algorithm 1: GPF Creation

Input:

1. *dataset* = list of data records (each record is a dictionary of features)
2. *performance_features* = list of feature names selected based on high correlation
3. *target_values* = list of target values corresponding to *performance_features*

Output:

dataset_with_gpf = dataset augmented with GPF score for each record

Process:

Set *num_features* = length(*performance_features*)

Set *score_per_match* = 100 / *num_features*

For each record in *dataset*:

Set *gpf_score* = 0

For *i* from 0 to *num_features* - 1:

Set *feature_name* = *performance_features*[*i*]

Set *target_value* = *target_values*[*i*]

If *record*[*feature_name*] == *target_value*:

gpf_score = *gpf_score* + *score_per_match*

Set *record*['*gpf*'] = round(*gpf_score*, 2) #2 decimal places setting

Return *dataset_with_gpf*

Example Algorithm 1: GPF Creation of dataset1

This example illustrates the calculation of the GPF for dataset1, which utilizes the top three performance-related features with the highest correlation to promotion: KPI, Awards Won, and Average Training Score.

Input:

1. *dataset* = list of data records (each record is a dictionary of features)
2. *performance_features* = (KPI, Awards won, Average training score)
3. *target_values* = (KPI >= 80%, Awards won = 1, Average training score >= 90)

Output:

dataset_with_gpf = dataset augmented with GPF score for each record

Process:

```

Set num_features = length(KPI, Awards won, Average training score)
Set score_per_match = 100 / 3
For each record in dataset:
    Set gpf_score = 0
    For i from 0 to 3 - 1:
        Set feature_name = KPI
        Set target_value = (KPI >= 80%)
        If record[KPI = 85] compare with (target_value) is true:
            gpf_score = 0 + 33.33
    #New Loop
    Set feature_name = Awards won
    Set target_value = 1
    If record[Awards won = 1] == (target_value = 1):
        gpf_score = 33.33 + 33.33
    #New Loop
    Set feature_name = Average training score
    Set target_value = (Average training score >= 90)
    If record[Average training score = 80] compare with (target_value) is
false:
        gpf_score = 66.66 + 0
    Set record['gpf'] = round(66.66, 2)    #2 decimal places setting
Return dataset_with_gpf

```

3.4.2 Feature Extraction

In this study, dimension reduction techniques, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), were applied during the feature extraction phase. These methods were selected because they represent two complementary approaches; PCA is a linear technique, while t-SNE is a nonlinear method.

PCA operates by transforming the original variables into a new set of linearly uncorrelated components, known as principal components, which capture the directions of maximum variance in the data, which helps reduce the number of features while still

preserving essential patterns and relationships within the dataset (Liu et al., 2020; Pal & Sharma, 2020). In addition, 20% of the most variable PCA features with the strongest relationship to promotion status were selected for further analysis. As a result, features were chosen from Dataset 1 and Dataset 2, as shown in Figures 3.9 and 3.10, respectively.

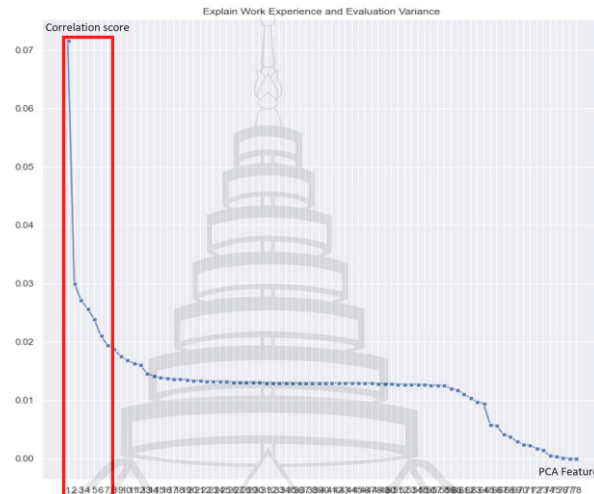


Figure 3.9 Twenty Percent of PCA Features That Contribute the Most Variance to The Promotion of Dataset 1

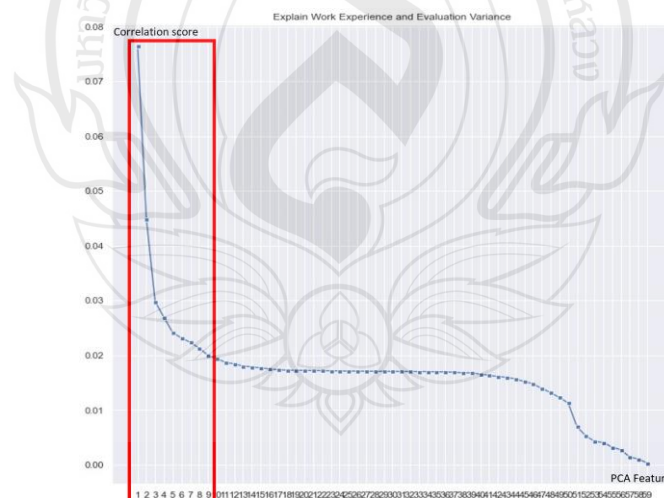


Figure 3.10 Twenty Percent of PCA Features That Contribute the Most Variance to The Promotion of Dataset2

In contrast to PCA, which is designed to preserve overall variance within the dataset, t-distributed Stochastic Neighbor Embedding (t-SNE) is more focused on capturing local relationships between data points in high-dimensional space. This technique works by comparing the probability distributions of data point similarities in both the original high-dimensional space and a lower-dimensional embedded space. t-SNE in visualizing complex, high-dimensional data by projecting it into two or three dimensions, where patterns and groupings become easier to interpret. Unlike linear methods, t-SNE uses a nonlinear approach that converts pairwise similarities into probabilities by applying Gaussian distributions in the high-dimensional space and t-distributions in the reduced space. The algorithm minimizes the Kullback–Leibler divergence between the two probability distributions using gradient descent, effectively maintaining the local structure of the data during the dimensionality reduction process. This makes t-SNE especially well-suited for tasks such as clustering, pattern recognition, and data exploration. In this study, the t-SNE was applied to reduce the original dataset into three dimensions, allowing for clearer clustering patterns and a better understanding of the underlying data structure.

After feature extraction and GPF construction, the data combination process was carried out. Two combination sets were created for each dataset. The combination of datasets is shown in Tables 3.3–3.6

Table 3.3 Examples of the Combined Data Sets (PCA and GPF) for Dataset 1

PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	GPF	Promotion
5.798384	1.635531	-2.77267	-3.113	11.88974	-7.24371	100	1
-5.02838	0.066464	-1.96308	2.155092	4.206181	-4.86832	66.66	1
-1.50148	4.037889	-3.04857	-4.61725	1.192841	-24.9768	33.33	0
2.183395	-7.05445	0.498713	2.519167	2.657033	-7.38592	0	0

Table 3.4 Examples of the Combined Data Sets (t-SNE and GPF) for Dataset 1

tSNE1	tSNE2	tSNE3	GPF	Promotion
22.52338	26.86526	16.61523	100	1
-18.422	-7.82231	-32.4508	66.66	0
-33.0769	8.412391	9.332352	33.33	1
12.17863	28.64398	-21.2067	0	1

Table 3.5 Examples of the Combined Data Sets (PCA and GPF) for Dataset 2

PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8	GPF	Promotion
277.8713	-153.4450	38.1881	4.4691	0.7346	-28.2834	8.8606	-14.0338	100	1
268.5994	-157.4280	42.7372	8.0960	0.4309	-24.8680	15.1353	-9.8614	66.66	1
268.9623	-152.4660	51.3508	-3.7866	-9.8511	-25.5350	22.7737	-17.6720	33.33	0
269.4408	-154.6630	50.8622	7.2342	-1.2787	-27.0241	15.9708	-10.5125	0	1

Table 3.6 Examples of the Combined Data Sets (t-SNE and GPF) for Dataset 2

tSNE1	tSNE2	tSNE3	GPF	Promotion
-10.9324	13.72186	13.63332	100	1
-2.59696	-9.47629	-6.67375	66.66	1
-5.97983	12.36096	-9.26217	33.33	0
9.776347	11.01716	21.40389	0	0

3.5 Data Balancing

To handle the issue of class imbalance, this study applied the Synthetic Minority Over-sampling Technique (SMOTE). This method was chosen to improve the performance of models without directly incorporating the promotion label into the training process. Instead of relying on the original class distribution, which typically includes far fewer promoted employees, SMOTE generates new, synthetic data points for the minority class to achieve a balanced dataset. The goal was to create an equal number of records for both promoted and non-promoted classes, enabling the model to treat both groups with equal importance during training.

In Dataset 1, the application of SMOTE resulted in the creation of 45,472 additional promotion records, effectively balancing the dataset. Similarly, for Dataset 2, SMOTE generated 31,830 synthetic promotion records to match the majority class. These balanced datasets provided a more stable foundation for unsupervised learning, thereby improving the interpretability of clustering results. Figures 3.11 and 3.12 illustrate the number of records before and after applying SMOTE for Dataset 1 and Dataset 2, respectively, following PCA feature extraction.

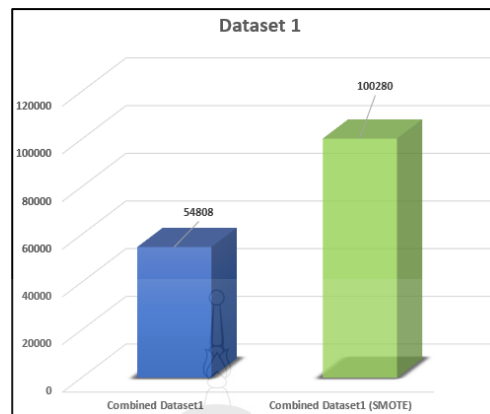


Figure 3.11 PCA Combined Dataset 1 before and after SMOTE

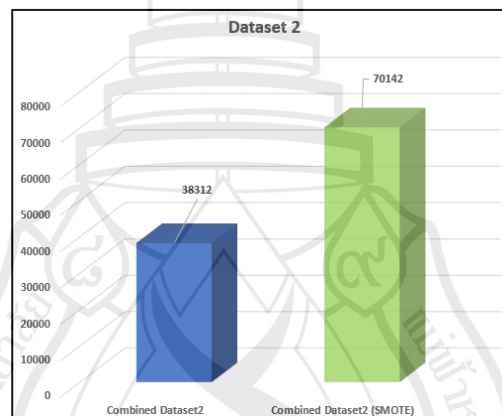


Figure 3.12 PCA Combined Dataset 2 before and after SMOTE

Table 3.7 Comparison of Selected Features of Dataset 1 before and after Applying SMOTE

Features	Before SMOTE	After SMOTE
KPI	(kpi >= 80%) = 19,291	(kpi >= 80%) = 50,361
	(kpi < 80%) = 35,517	(kpi < 80%) = 49,919
Awards won	received = 1,270	received = 4,801
	not received = 53,538	not received = 95,479
Average training score	(score >= 90) = 764	(score >= 90) = 6,253
	(score < 90) = 54,044	(score < 90) = 94,027

Table 3.8 Comparison of Selected Features of Dataset 2 before and after Applying SMOTE

Features	Before SMOTE	After SMOTE
Targets met	meet = 13,524	meet = 35,051
	not meet = 24,788	not meet = 35,091
Previous award	received = 887	received = 3,434
	not received = 37,425	not received = 66,708
Training score average	(score \geq 90) = 55	(score \geq 90) = 466
	(score < 90) = 38,257	(score < 90) = 69,676

3.6 Model Construction

3.6.1 Classification

In this study, six widely recognized classification models were employed to construct predictive models for employee promotion. These include Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Neural Network (NN). These models were selected based on their diversity in underlying algorithmic approaches, which encompass tree-based (RF and DT), margin-based (SVM), instance-based (KNN), statistical (LR), and deep learning (NN) techniques. This diversity enables a comprehensive evaluation of the effectiveness and adaptability of the proposed features of engineering techniques across various learning paradigms.

Before constructing the model, the datasets underwent extensive preprocessing. This process involved integrating the proposed Generated Promotion Feature (GPF), a feature engineering derived from performance-related indicators, as well as dimensionality reduction using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). Additionally, class imbalance issues were addressed using the Synthetic Minority Oversampling Technique (SMOTE), ensuring that the minority class (promoted employees) was sufficiently represented during model training.

To ensure the robustness and generalizability of the classification models, a stratified 10-fold cross-validation technique was applied. In this approach, the dataset

was randomly divided into 10 equal-sized folds while preserving the original class distribution within each fold. During each iteration, one fold was held out for validation while the remaining nine were used for model training. This process was repeated ten times, with each fold serving as the validation set once. The final performance metrics were computed as the average of the results across all folds.

The model evaluation was conducted using four primary metrics: Accuracy, Precision, Recall, and F1-score. These metrics provide a balanced view of each model's performance, particularly in the context of class-imbalanced data, where reliance on accuracy alone can be misleading. The evaluation focused on assessing each model's ability to accurately identify promotable employees while minimizing false positives and false negatives. Hyperparameter tuning for each classifier was performed using grid search methods within the cross-validation framework to optimize model performance.

This systematic modeling and validation framework was designed to assess how effectively each classification algorithm leveraged the enhanced features derived from GPF and dimensionality reduction techniques. The results obtained from this process were later compared and analyzed to identify the most suitable model configurations for predicting employee promotions in human resource analytics.

3.6.2 Clustering

K-means clustering was a partition-based algorithm that assigns each data point to the cluster with the nearest centroid, which represented the average position of all data points within that cluster. The process began by initializing a predefined number of clusters (k) and then calculating the centroids. Each data point was grouped with the cluster whose centroid was closest, typically measured using Euclidean distance. Then, the centroids were recalculated based on the updated cluster members, and this process of assignment and update continued iteratively until the cluster assignments stabilized or reached convergence. Furthermore, K-means worked well on datasets where the data naturally form well-separated groups, such as circular, elliptical, or linearly distributed patterns. Its effectiveness was enhanced when the number of clusters was known in advance, as this helped ensure that the grouping process was structured and consistent with the data's underlying patterns.

Fuzzy clustering, also known as fuzzy c-means, enabled each data point to have partial membership across multiple clusters rather than being strictly assigned to just

one. This could be achieved by assigning membership values ranging from 0 to 1, which reflected the degree to which a data point belonged to a given cluster. Points that were closer to a cluster's centroid received higher membership values, while those farther away were assigned lower values. This technique was beneficial when the boundaries between clusters were unclear or overlapping, making it suitable for datasets with ambiguous or gradual transitions between groups. Fuzzy clustering differed fundamentally from K-means, which forced each point into a single cluster. Such challenging assignments in K-means could lead to misclassification when a data point lies between two or more clusters and does not distinctly belong to just one. By allowing flexible memberships, fuzzy clustering provided a more nuanced and realistic interpretation of data groupings, especially in complex, real-world scenarios where strict cluster separation was not always present.

In this study, K-means clustering was configured to use two clusters ($n_clusters = 2$, $K = 2$). Moreover, fuzzy clustering was configured to use three clusters ($n_clusters = 3$). This setup was designed to facilitate the identification and separation of employees based on their promotion status. As previously stated, multiple versions of combined datasets were prepared for both Dataset 1 and Dataset 2, incorporating different combinations of original features, GPF, PCA, and t-SNE. These dataset variations were used to explore which combination would deliver the most effective clustering results in representing employee promotion potential. Through the application of various clustering methods to a range of dataset configurations, this research aimed to evaluate the advantages of each technique and determine the most suitable set of features for effectively categorizing employees into promotable and non-promotable groups.

3.7 Result Evaluation

3.7.1 Classification Evaluation

In this study, four commonly used evaluation metrics, including accuracy, precision, recall, and F1-score, were employed to assess the performance of classification models. These metrics were selected to provide a comprehensive

understanding of each model's effectiveness, especially in the context of imbalanced datasets commonly found in human resource data.

Accuracy was measured by the proportion of correct predictions over the total number of instances. It was a straightforward metric and valuable when the class distribution was relatively balanced. However, it may be misleading when the data is imbalanced, as high accuracy can be achieved by predicting only the majority class.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$

Precision evaluates how many of the predicted positive instances are positive. It is significant in the context of employee promotion prediction, where false positives (predicting promotion for a non-promotable employee) may result in wasted resources or unfair decisions.

$$Precision = \frac{TP}{TP+FP} \quad (3.2)$$

Recall, also known as sensitivity, measures the model's ability to identify all actual positive instances correctly. A high recall was crucial in this study, as it reflected the model's ability to identify promotable employees, a key objective in HR decision-making.

$$Recall = \frac{TP}{TP+FN} \quad (3.3)$$

F1-Score is the harmonic mean of precision and recall. It balanced the trade-off between these two metrics and was especially useful when the dataset was imbalanced. A high F1-score indicated that the model had both high precision and high recall, which was ideal in promotion decision-making.

$$F1 - Score = \frac{Precision * Recall}{Precision + Recall} \quad (3.4)$$

By combining these four metrics, this study provided a more comprehensive evaluation of classification performance across various datasets and model configurations. This approach facilitated the selection of the most suitable model for making predictions related to promotions in human resource analytics.

3.7.2 Clustering Evaluation

To assess the quality of the promotion clustering results, four evaluation metrics were used: Rand Index (RI), Mutual Information (MI), V-measure (V), and Fowlkes–Mallows Index (FMI). Each of these indices provided a different perspective on how closely the clustering output matched the actual promotion labels.

The Rand Index (RI) evaluated the similarity between the clustering results and the ground truth by considering pairs of data points. Specifically, it measured the proportion of data point pairs that were either correctly grouped in the same cluster or correctly separated into different clusters. A higher RI value indicated a greater degree of alignment between the predicted clusters and the actual labels. In this study, the RI was calculated using the formula shown in Equation (3.5) and served as one of the primary indicators of clustering accuracy and consistency.

$$RI = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.5)$$

In the context of clustering evaluation, the classification of pairwise data points can be described using four standard terms: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

1. True Positive (TP) occurs when a pair of data points is correctly grouped into the same cluster by the algorithm and is also classified together in the ground truth.
2. True Negative (TN) refers to a pair of points that are correctly identified as belonging to different clusters, both by the algorithm and the ground truth.
3. False Positive (FP) happens when the algorithm incorrectly places a pair of data points into the same cluster, even though they are not grouped based on the ground truth.
4. False Negative (FN) occurs when the algorithm fails to assign a pair of related data points to the same cluster, despite them being grouped in the ground truth.

These four values form the basis for computing clustering evaluation metrics, such as the Rand Index, which helps assess how closely the clustering results match the actual class labels.

Mutual Information (MI) is used to evaluate the amount of information shared between the clustering results and the actual labels. It reflects the degree of mutual dependence between the predicted clusters and the ground truth categories. A higher

MI value indicates that the clustering output captures a greater portion of the underlying data structure and aligns more closely with the actual labels. In other words, the more the predicted clusters can explain or represent the actual classification, the higher the MI score will be. In this study, MI is computed using Equation (3.6) and serves as a key indicator for assessing the alignment and effectiveness of clustering outcomes.

$$MI(Y; C) = H(Y) - H(Y|C) \quad (3.6)$$

In the context of clustering evaluation, MI (Y; C) denotes the Mutual Information score.

1. Y refers to the actual class labels
2. C represents the cluster assignments produced by the algorithm.
3. $H(Y)$ indicates the entropy of the actual class labels, reflecting the overall uncertainty or unpredictability in the label distribution.
4. $H(Y|C)$ represents the conditional entropy of Y given C, which measures the remaining uncertainty about class labels when the cluster assignments are known.

Essentially, the difference between $H(Y)$ and $H(Y|C)$ quantifies how much information the clustering results provide about the actual labels. A lower conditional entropy implies that the clustering effectively captures the actual class structure.

V-measure is a clustering evaluation metric that assesses the agreement between two independent label assignments, such as the predicted cluster labels and the actual class labels, when applied to the same dataset. This metric combines two important aspects including homogeneity and completeness, into a single balanced score. Homogeneity evaluates whether each cluster contains only data points that belong to a single class, ensuring internal consistency within clusters. On the other hand, completeness ensures that all data points from the same class are assigned to the same cluster, thereby promoting consistency across class labels. A high V-measure score indicates that the clustering structure aligns well with the ground truth, both in terms of internal purity and overall class coverage. In this study, the V-measure was computed using Equation (3.7) and served as a key metric to assess the quality and reliability of the clustering results.

$$v = 2 \left(\frac{(\text{homogeneity})(\text{completeness})}{\text{homogeneity} + \text{completeness}} \right) \quad (3.7)$$

In this context, “v” denotes the V-measure score, which serves as a balanced evaluation of clustering performance. The score is derived from two core components: homogeneity and completeness.

1. Homogeneity ensures that each cluster is composed of data points that belong to only one actual class, indicating that the clustering process maintains internal consistency within each group.

2. Completeness assesses whether all data points that belong to the same class are grouped into a single cluster, reflecting the algorithm’s ability to preserve class integrity across clusters.

When both homogeneity and completeness are high, the V-measure score increases, signifying that the clustering solution closely matches the actual class distribution.

The Fowlkes–Mallows Index (FMI) is a clustering evaluation metric that assesses how well the predicted clusters align with the actual class labels by comparing the distribution of data points within clusters to the overall distribution across classes. Specifically, FMI measures the ratio of variance within clusters relative to the total variance observed between the predicted clusters and the actual class labels. Lower FMI values suggest that the clusters are more compact and tightly grouped, which generally indicates better clustering performance. Unlike other metrics such as Rand Index (RI) or Mutual Information (MI), FMI is less sensitive to differences in cluster sizes, making it a more robust option in cases of class imbalance. A high FMI value, particularly one that approaches 1, implies a strong correspondence between the predicted clustering and the ground truth classification. In this study, FMI was calculated using Equation (3.8) as part of the comprehensive evaluation of clustering quality.

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \quad (3.8)$$

In the context of clustering evaluation, the classification of data points can be understood using four key terms: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

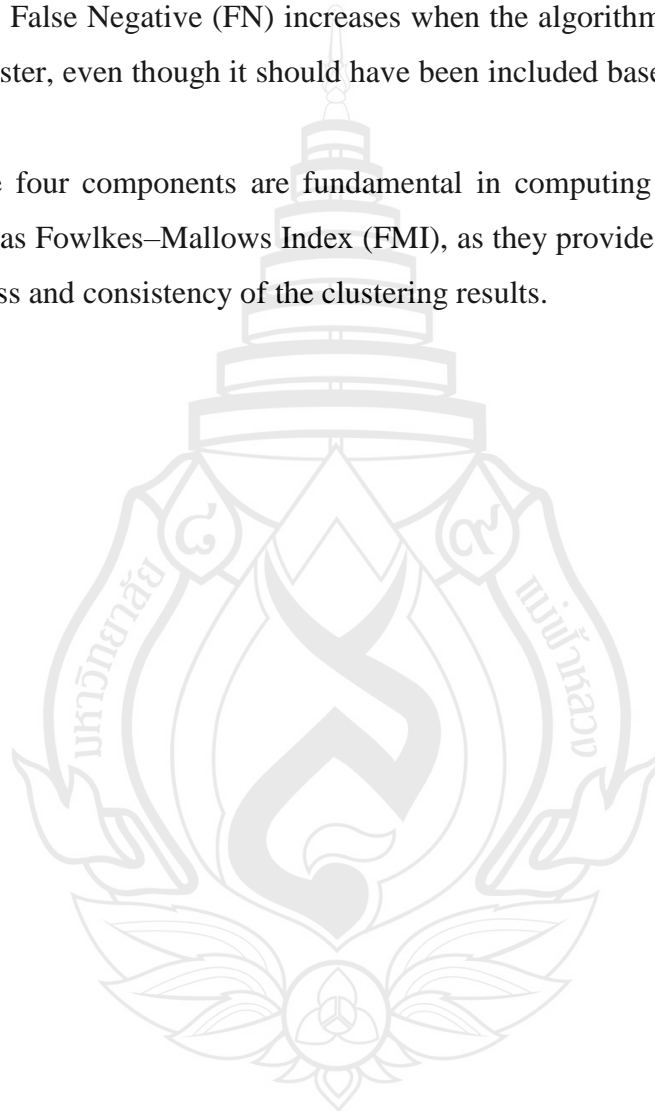
1. True Positive (TP) occurs when a data point is correctly grouped into a cluster by the algorithm and is also classified within the same cluster in the ground truth.

2. True Negative (TN) refers to a data point that is correctly identified as not belonging to a particular cluster by both the algorithm and the ground truth.

3. False Positive (FP) is recorded when the algorithm incorrectly assigns a data point to a cluster, despite the ground truth indicating that it does not belong to that cluster.

4. False Negative (FN) increases when the algorithm fails to assign a data point to a cluster, even though it should have been included based on the ground truth labels.

These four components are fundamental in computing clustering evaluation metrics such as Fowlkes–Mallows Index (FMI), as they provide a basis for measuring the correctness and consistency of the clustering results.



CHAPTER 4

EXPERIMENTAL RESULTS

In this study, the experimental design is divided into two major groups based on learning type: Supervised Learning and Unsupervised Learning. Starting with Supervised Learning, the experiments are further divided into two subgroups: Models trained on the original imbalanced datasets (without any balancing techniques), and Models trained on datasets that have been balanced using SMOTE (Synthetic Minority Oversampling Technique). Similarly, in the Unsupervised Learning group, clustering experiments are conducted in two groups: Clustering with the original imbalanced datasets, and Clustering with datasets that have been preprocessed using SMOTE for class balance.

The main objective of this structure is to evaluate the impact of GPF and PCA on model performance across both learning paradigms. Additionally, the experiments are conducted using two distinct HR datasets to validate the generalizability and robustness of the proposed framework in different organizational contexts.

4.1 Classification Results

Classification applied six algorithms: Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Neural Network (NN).

4.1.1 Random Forest Classification

Table 4.1 Performance Comparison of Random Forest for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.5360	0.0787	0.1921	0.0610
2	Dataset with GPF	0.4705	0.0524	0.1981	0.0509
3	Dataset with PCA	0.5703	0.0971	0.1852	0.0643
4	Dataset with t-SNE	0.5606	0.1007	0.1852	0.0618
5	Dataset with GPF and PCA	0.6024	0.3007	0.1673	0.0665
6	Dataset with GPF and t-SNE	0.5008	0.0655	0.1981	0.0527
7	PCA	0.7283	0.2226	0.0846	0.0281
8	t-SNE	0.7559	0.1466	0.0548	0.0281
9	PCA and GPF	0.6859	0.0204	0.1351	0.0324
10	t-SNE and GPF	0.6738	0.2244	0.1398	0.0371

The results of the Random Forest classification without applying SMOTE indicate that while accuracy was highest when using PCA or t-SNE alone, these configurations received poor recall and F1-scores due to class imbalance. Notably, the integration of GPF and PCA in the original dataset (without dimensionality reduction) yields the most balanced result, significantly improving precision while maintaining acceptable accuracy. This suggests that GPF, when combined with PCA in a feature-augmentation role rather than feature-reduction, enhances the model's ability to identify promotable employees.

Table 4.2 Performance Comparison of Random Forest for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.7272	0.7103	0.9744	0.8053
2	Dataset with GPF and SMOTE	0.7246	0.7099	0.9734	0.8043
3	Dataset with PCA and SMOTE	0.7239	0.7061	0.9780	0.8033
4	Dataset with t-SNE and SMOTE	0.7162	0.6947	0.9755	0.7960
5	Dataset with GPF, PCA, and SMOTE	0.7262	0.7099	0.9789	0.8058
6	Dataset with GPF, t-SNE, and SMOTE	0.7289	0.7115	0.9778	0.8068
7	PCA and SMOTE	0.7731	0.7368	0.9268	0.8140
8	t-SNE and SMOTE	0.7486	0.7183	0.8858	0.7869
9	PCA, GPF and SMOTE	0.7692	0.7349	0.9277	0.8125
10	t-SNE, GPF and SMOTE	0.7463	0.7186	0.8981	0.7904

The results of the Random Forest classification on Dataset 1 after applying SMOTE demonstrated a significant overall improvement across all evaluation metrics. While the highest accuracy was achieved with the PCA-only configuration (0.7731), this setup showed slightly reduced recall compared to other combinations. Notably, configurations combining GPF with PCA or t-SNE, such as Dataset with GPF, t-SNE, and SMOTE (Accuracy = 0.7289, F1 = 0.8068) and Dataset with GPF, PCA, and SMOTE (Accuracy = 0.7262, F1 = 0.8058), offered a better balance of precision and recall, resulting in strong and stable F1-scores.

The integration of GPF with dimensionality reduction and balanced data allowed the model to identify promotable employees more effectively, as reflected in recall values consistently above 0.96. Although PCA and t-SNE alone yielded slightly higher accuracy, their overall performance was less balanced compared to GPF-enhanced configurations.

These findings confirmed that while accuracy remained an important indicator, combining GPF with SMOTE and PCA or t-SNE yielded more consistent performance across all metrics. This emphasized the benefit of domain-informed features and data balancing techniques in real-world HR promotion prediction tasks.

Table 4.3 Performance Comparison of Random Forest for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.5701	0.1069	0.1751	0.0563
2	Dataset with GPF	0.5195	0.0844	0.1969	0.0523
3	Dataset with PCA	0.5984	0.1496	0.1751	0.0654
4	Dataset with t-SNE	0.8780	0.0596	0.0151	0.0204
5	Dataset with GPF and PCA	0.5211	0.0837	0.1945	0.0507
6	Dataset with GPF and t-SNE	0.5094	0.0712	0.1954	0.0502
7	PCA	0.8231	0.1285	0.0481	0.0258
8	t-SNE	0.8780	0.0596	0.0151	0.0204
9	PCA and GPF	0.6950	0.1253	0.1402	0.0333
10	t-SNE and GPF	0.7529	0.0266	0.0752	0.0270

The results of Random Forest classification on Dataset 2 without applying SMOTE showed significant performance variation across different feature combinations. While the highest accuracy was achieved with t-SNE transformations

(87.80%), the recall and F1-score are extremely low, indicating a failure to identify promotable employees. The most balanced result among all combinations was observed when PCA was applied to the original dataset ($F1 = 0.0654$), though performance remained weak due to the unresolved class imbalance. These findings reinforced the need to apply data balancing techniques, SMOTE, to enhance model generalizability and fairness.

Table 4.4 Performance Comparison of Random Forest for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.7727	0.7464	0.9662	0.8294
2	Dataset with GPF and SMOTE	0.7507	0.7308	0.9681	0.8182
3	Dataset with PCA and SMOTE	0.7508	0.7281	0.9799	0.8198
4	Dataset with t-SNE and SMOTE	0.7778	0.7491	0.9636	0.8312
5	Dataset with GPF, PCA, and SMOTE	0.7390	0.7215	0.9808	0.8145
6	Dataset with GPF, t-SNE, and SMOTE	0.7478	0.7289	0.9660	0.8162
7	PCA and SMOTE	0.7776	0.7352	0.9197	0.8121
8	t-SNE and SMOTE	0.7649	0.7372	0.8302	0.7802
9	PCA, GPF and SMOTE	0.7524	0.7195	0.9234	0.8008
10	t-SNE, GPF and SMOTE	0.7340	0.7129	0.8663	0.7750

After applying SMOTE to Dataset 2, the classification performance of the Random Forest model significantly improved across all metrics. The best overall result was obtained when combining t-SNE with SMOTE ($F1 = 0.8312$), followed closely by the baseline SMOTE-only dataset and the combination of PCA with SMOTE. Models incorporating GPF also achieved competitive results, particularly in recall, with several configurations exceeding a 96% recall rate. These findings confirmed the effectiveness of SMOTE in addressing class imbalance and demonstrated the complementary strengths of dimensionality reduction and feature engineering in enhancing the performance of supervised learning.

4.1.2 Decision Tree Classification

Table 4.5 Performance Comparison of Decision Tree for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.3214	0.0221	0.2129	0.0398
2	Dataset with GPF	0.3156	0.0181	0.1983	0.0338
3	Dataset with PCA	0.3240	0.0221	0.2133	0.0394
4	Dataset with t-SNE	0.3224	0.0224	0.2150	0.0404
5	Dataset with GPF and PCA	0.3202	0.0186	0.1908	0.0341
6	Dataset with GPF and t-SNE	0.3191	0.0187	0.1914	0.0332
7	PCA	0.6645	0.0574	0.1456	0.0552
8	t-SNE	0.6910	0.0631	0.1386	0.0611
9	PCA and GPF	0.6405	0.1792	0.1760	0.0528
10	t-SNE and GPF	0.6480	0.1120	0.1752	0.0557

The performance of the Decision Tree classifier on Dataset 1 without any balancing techniques was consistently poor across all feature combinations. Although combinations such as PCA or t-SNE alone yielded relatively high accuracy, the recall and F1-score remained critically low due to the underlying class imbalance. The combination of GPF with either PCA or t-SNE yielded slight improvements in precision but did not significantly impact overall model effectiveness. These results emphasized the importance of applying data balancing methods, such as SMOTE, before training classifiers in imbalanced HR datasets.

Table 4.6 Performance Comparison of Decision Tree for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.7272	0.7103	0.9744	0.8053
2	Dataset with GPF and SMOTE	0.7246	0.7099	0.9734	0.8043
3	Dataset with PCA and SMOTE	0.7239	0.7061	0.9780	0.8033
4	Dataset with t-SNE and SMOTE	0.7162	0.6947	0.9755	0.7960
5	Dataset with GPF, PCA, and SMOTE	0.7262	0.7099	0.9789	0.8058
6	Dataset with GPF, t-SNE, and SMOTE	0.7289	0.7115	0.9778	0.8068
7	PCA and SMOTE	0.7731	0.7368	0.9268	0.8140
8	t-SNE and SMOTE	0.7486	0.7183	0.8858	0.7869
9	PCA, GPF and SMOTE	0.7692	0.7349	0.9277	0.8125
10	t-SNE, GPF and SMOTE	0.7463	0.7186	0.8981	0.7904

The Decision Tree model exhibited extremely low performance on Dataset 1 without SMOTE, with F1-scores of less than 0.06 across all feature configurations. However, after applying SMOTE, F1-scores increased dramatically, often exceeding 0.80. Among all combinations, PCA and SMOTE ($F1 = 0.8140$) and PCA, GPF, and SMOTE ($F1 = 0.8125$) yielded the best results, suggesting that dimensionality reduction with PCA, especially when supplemented by domain-informed feature engineering via GPF, plays a crucial role in model performance. These findings emphasize the importance of applying data balancing techniques, such as SMOTE, when working with imbalanced HR datasets.

Table 4.7 Performance Comparison of Decision Tree for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.4008	0.0338	0.2127	0.0521
2	Dataset with GPF	0.3986	0.0228	0.2056	0.0388
3	Dataset with PCA	0.4019	0.0327	0.2112	0.0511
4	Dataset with t-SNE	0.7996	0.0588	0.0907	0.0699
5	Dataset with GPF and PCA	0.3989	0.0229	0.2056	0.0389
6	Dataset with GPF and t-SNE	0.3978	0.0229	0.2059	0.0392
7	PCA	0.7179	0.0531	0.1206	0.0596
8	t-SNE	0.7995	0.0619	0.0882	0.0703
9	PCA and GPF	0.6342	0.0467	0.1933	0.0559
10	t-SNE and GPF	0.7025	0.0686	0.1307	0.0568

On Dataset 2, the Decision Tree classifier performed poorly across all configurations without the use of SMOTE. The highest F1-score observed was 0.0703 from the t-SNE transformation alone, which was marginally better than other configurations but still insufficient for identifying promotable employees. The use of GPF in combination with PCA or t-SNE showed little to no improvement, indicating that class imbalance severely hindered model learning. These results reinforce the necessity of applying balancing techniques prior to training when working with imbalanced HR data.

Table 4.8 Performance Comparison of Decision Tree for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.7275	0.7094	0.9729	0.8047
2	Dataset with GPF and SMOTE	0.7277	0.7084	0.9745	0.8041
3	Dataset with PCA and SMOTE	0.6941	0.6802	0.9660	0.7824
4	Dataset with t-SNE and SMOTE	0.6688	0.6428	0.9674	0.7599
5	Dataset with GPF, PCA, and SMOTE	0.6983	0.6791	0.9682	0.7834
6	Dataset with GPF, t-SNE, and SMOTE	0.7143	0.6895	0.9728	0.7933
7	PCA and SMOTE	0.6998	0.6723	0.8279	0.7398
8	t-SNE and SMOTE	0.7321	0.7107	0.7857	0.7449
9	PCA, GPF and SMOTE	0.6985	0.6740	0.8491	0.7466
10	t-SNE, GPF and SMOTE	0.6992	0.6802	0.8160	0.7370

After applying SMOTE to Dataset 2, the Decision Tree classifier demonstrated consistent and balanced performance across all evaluation metrics. The baseline configuration with SMOTE achieved the highest overall scores, with Accuracy = 0.7275, Precision = 0.7094, Recall = 0.9729, and F1-score = 0.8047, indicating a strong ability to identify promotable employees while maintaining low false positives correctly. Feature combinations such as Dataset with GPF, t-SNE, and SMOTE, as well as Dataset with GPF, PCA, and SMOTE, also worked well, with F1-scores above 0.79 and strong recall (above 0.96). This demonstrated their suitability for tasks that prioritize sensitivity to promotable employees. Although t-SNE and SMOTE alone achieved the highest accuracy (0.7321), they delivered lower recall and F1 compared to the top configurations. In addition, the results confirmed that SMOTE significantly improved model balance across all metrics, and that integrating GPF with dimensionality reduction methods further stabilized performance without losing precision. The combination of high recall and reasonably high precision indicated that the Decision Tree, when supported by SMOTE and informed by relevant features, was effective and reliable for promotion prediction in imbalanced HR datasets.

4.1.3 SVM Classification

Table 4.9 Performance Comparison of SVM for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.7426	0.0585	0.3085	0.0983
2	Dataset with GPF	0.5252	0.0448	0.5000	0.0823
3	Dataset with PCA	0.7477	0.5600	0.3162	0.1116
4	Dataset with t-SNE	0.7470	0.5584	0.3014	0.0999
5	Dataset with GPF and PCA	0.5348	0.0458	0.5000	0.0839
6	Dataset with GPF and t-SNE	0.5244	0.0448	0.5000	0.0822
7	PCA	0.8360	0.0421	0.0936	0.0580
8	t-SNE	0.8543	0.0388	0.0653	0.0487
9	PCA and GPF	0.6320	0.0447	0.3616	0.0796
10	t-SNE and GPF	0.6183	0.0558	0.5000	0.1004

Without applying SMOTE, the SVM model on Dataset 1 exhibited a typical imbalance-driven performance pattern, characterized by high accuracy but low recall and F1-score. While the datasets with PCA and t-SNE yielded the best F1-scores (0.1116 and 0.0999, respectively), these values remained too low for practical applicability. Configurations involving GPF achieved high recall but poor precision, indicating sensitivity to promotable instances without reliable discrimination. These results suggest that the model's ability to identify the minority class (i.e., promotable employees) remains limited without the use of a balancing technique such as SMOTE.

Table 4.10 Performance Comparison of SVM for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.6485	0.6620	0.8577	0.7254
2	Dataset with GPF and SMOTE	0.4460	0.4678	0.8322	0.5982
3	Dataset with PCA and SMOTE	0.5302	0.5260	0.8073	0.6308
4	Dataset with t-SNE and SMOTE	0.5408	0.5432	0.7747	0.6250
5	Dataset with GPF, PCA, and SMOTE	0.4652	0.4777	0.8390	0.6075
6	Dataset with GPF, t-SNE, and SMOTE	0.4464	0.4671	0.8314	0.5971
7	PCA and SMOTE	0.5121	0.5078	0.7907	0.6184
8	t-SNE and SMOTE	0.4569	0.4674	0.6021	0.5252
9	PCA, GPF and SMOTE	0.4967	0.4982	0.8053	0.6155
10	t-SNE, GPF and SMOTE	0.4906	0.4931	0.7527	0.5954

After applying SMOTE to Dataset 1, the SVM classifier demonstrated substantial improvements across all performance metrics. The baseline SMOTE configuration yielded a strong balance, with Accuracy = 0.6485, Precision = 0.6620, Recall = 0.8577, and an F1-score of 0.7254, indicating that the model could effectively identify promotable employees while maintaining an acceptable false positive rate.

Configurations involving dimensionality reduction techniques such as PCA and t-SNE combined with SMOTE also performed well. For instance, the Dataset with PCA and SMOTE achieved an F1-score of 0.6308, while the Dataset with t-SNE and SMOTE followed closely with 0.6250. These setups tended to lose some recall in exchange for improved precision, suggesting a more conservative decision boundary. Meanwhile, the inclusion of GPF with either PCA or t-SNE yielded additional gains in recall (often above 0.80), though this sometimes came at the expense of lower precision.

Overall, the results indicated that SMOTE played a crucial role in allowing SVM to function effectively in imbalanced settings. While PCA and GPF combinations provided flexibility and tuning options, the base SMOTE configuration alone already delivered a strong and reliable outcome for HR promotion prediction tasks.

Table 4.11 Performance Comparison of SVM for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.9154	0.0000	0.0000	0.0000
2	Dataset with GPF	0.9154	0.0000	0.0000	0.0000
3	Dataset with PCA	0.9154	0.0000	0.0000	0.0000
4	Dataset with t-SNE	0.9154	0.0000	0.0000	0.0000
5	Dataset with GPF and PCA	0.9154	0.0000	0.0000	0.0000
6	Dataset with GPF and t-SNE	0.9154	0.0000	0.0000	0.0000
7	PCA	0.9154	0.0000	0.0000	0.0000
8	t-SNE	0.9154	0.0000	0.0000	0.0000
9	PCA and GPF	0.6615	0.0622	0.4972	0.1106
10	t-SNE and GPF	0.6604	0.0619	0.4963	0.1101

Without applying SMOTE, the SVM classifier completely failed to identify promotable employees across almost all configurations. Despite a high accuracy of 0.9154, the model achieved zero precision, recall, and F1-score in most setups, clearly

indicating that it predicted only the majority class. This misleadingly high accuracy was a classic condition of class imbalance, where the minority class is entirely ignored.

Only the configurations involving PCA and GPF, as well as t-SNE and GPF, showed any meaningful detection, with recall values near 0.50 and modest F1-scores of around 0.11. However, precision remained low, 0.06, suggesting a high number of false positives among limited true positives. These combinations slightly improved balance but remained far from practically usable.

These results emphasized the significance of addressing class imbalance when using SVM in promotion prediction tasks. While the model achieved superficially high accuracy, its complete inability to generalize to the minority class without SMOTE made it ineffective for real-world HR applications.

Table 4.12 Performance Comparison of SVM for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.4940	0.4947	0.7734	0.5916
2	Dataset with GPF and SMOTE	0.5494	0.7706	0.4231	0.3919
3	Dataset with PCA and SMOTE	0.4939	0.4946	0.7757	0.5926
4	Dataset with t-SNE and SMOTE	0.4931	0.4939	0.7741	0.5914
5	Dataset with GPF, PCA, and SMOTE	0.5507	0.7710	0.4258	0.3952
6	Dataset with GPF, t-SNE, and SMOTE	0.6377	0.7705	0.5999	0.6200
7	PCA and SMOTE	0.4151	0.4443	0.6601	0.5281
8	t-SNE and SMOTE	0.4873	0.4883	0.5233	0.5044
9	PCA, GPF and SMOTE	0.7153	0.7685	0.7569	0.7441
10	t-SNE, GPF and SMOTE	0.6785	0.6770	0.7733	0.7151

After applying SMOTE to Dataset 2, the SVM model showed a dramatic improvement across all configurations. Unlike the pre-SMOTE results, where recall and F1-scores were zero, the balanced data allowed the model to learn from the minority class and generalize more effectively. The most balanced and high-performing configurations were PCA, GPF, and SMOTE, which achieved an Accuracy of 0.7153, a Precision of 0.7685, a Recall of 0.7569, and an impressive F1-score of 0.7441. This suggests excellent overall performance in correctly identifying promotable employees while minimizing false positives.

Interestingly, the Dataset with GPF, PCA, and SMOTE led to a high precision of 0.77 but a significantly lower recall of 0.42, suggesting a more conservative model that may miss positive cases. This highlighted the trade-off between capturing all promotable individuals and reducing false positives, depending on the feature set.

Overall, the results confirmed that combining GPF with dimensionality reduction techniques and applying SMOTE offered the best balance between accuracy, precision, and recall. These configurations significantly enhanced the effectiveness of SVM for promotion prediction in HR analytics.

4.1.4 K-Nearest Neighbor Classification

Table 4.13 Performance Comparison of KNN for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.8737	0.1502	0.1028	0.0951
2	Dataset with GPF	0.4330	0.0152	0.1525	0.0275
3	Dataset with PCA	0.9028	0.2687	0.1150	0.1252
4	Dataset with t-SNE	0.8844	0.3550	0.0900	0.0979
5	Dataset with GPF and PCA	0.6055	0.0328	0.1694	0.0414
6	Dataset with GPF and t-SNE	0.6876	0.1899	0.1777	0.0553
7	PCA	0.7880	0.0186	0.0683	0.0275
8	t-SNE	0.8174	0.0233	0.0306	0.0184
9	PCA and GPF	0.7309	0.0200	0.1043	0.0290
10	t-SNE and GPF	0.7199	0.0178	0.0929	0.0260

Without applying SMOTE, the KNN classifier on Dataset 1 produced inconsistent and overall weak performance across most configurations. Although certain combinations, such as Dataset with PCA and Dataset with t-SNE, yielded high accuracy (0.9028 and 0.8844, respectively), their F1-scores remained low (0.1252 and 0.0979) due to poor recall. This indicated that KNN predominantly predicted the majority class and failed to identify effectively the promotable employees, like other models under imbalanced data conditions.

The configuration using PCA as additional features (Dataset with PCA) yielded the highest F1-score overall (0.1252), and the original dataset and t-SNE also demonstrated a relatively better balance compared to the others. However, setups

involving GPF either alone or in combination generally resulted in lower precision and F1-scores, possibly due to sensitivity to minority samples in the absence of balancing.

These results highlighted that if KNN relied solely on high accuracy, it could be misleading in imbalanced classification problems. To make the model suitable for HR promotion prediction, where minority class identification was crucial, class balancing methods such as SMOTE were essential to improve recall and overall performance.

Table 4.14 Performance Comparison of KNN for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.8198	0.7404	0.9937	0.8477
2	Dataset with GPF and SMOTE	0.6797	0.6486	0.9959	0.7724
3	Dataset with PCA and SMOTE	0.7612	0.6818	0.9923	0.8074
4	Dataset with t-SNE and SMOTE	0.7885	0.7182	0.9778	0.8255
5	Dataset with GPF, PCA, and SMOTE	0.7213	0.6871	0.9940	0.7988
6	Dataset with GPF, t-SNE, and SMOTE	0.7614	0.7331	0.9832	0.8244
7	PCA and SMOTE	0.6858	0.6725	0.8127	0.7294
8	t-SNE and SMOTE	0.6914	0.6740	0.8094	0.7301
9	PCA, GPF and SMOTE	0.6825	0.6686	0.8263	0.7316
10	t-SNE, GPF and SMOTE	0.6782	0.6611	0.8356	0.7308

After applying SMOTE, the performance of the KNN model on Dataset 1 improved substantially across all configurations. The baseline SMOTE configuration achieved the highest F1-score (0.8477), with a strong balance in all metrics: Accuracy = 0.8198, Precision = 0.7404, and an exceptionally high Recall = 0.9937, indicating the model could correctly identify nearly all promotable employees.

Other configurations, such as the Dataset with t-SNE and SMOTE (F1 = 0.8255) and the Dataset with GPF, t-SNE, and SMOTE (F1 = 0.8244), also performed very effectively, demonstrating that dimensionality reduction techniques, when combined with SMOTE, could enhance model sensitivity while maintaining high precision. The inclusion of GPF slightly reduced precision in some cases; however, it maintained extremely high recall across the board (often above 0.98), making it suitable for cases where missing a promotable candidate was significant.

Although some Dataset setups with GPF, PCA, and SMOTE showed slightly lower accuracy, they still yielded strong F1-scores above 0.79, reinforcing the model's

reliability when addressing minority class imbalance. This confirmed that KNN, when paired with both data balancing and meaningful feature engineering, can serve as a highly effective classifier in predicting employee promotions.

Table 4.15 Performance Comparison of KNN for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.8844	0.2158	0.0983	0.0923
2	Dataset with GPF	0.5039	0.0191	0.1827	0.0342
3	Dataset with PCA	0.9119	0.3181	0.0836	0.1169
4	Dataset with t-SNE	0.9093	0.2374	0.0043	0.0079
5	Dataset with GPF and PCA	0.6365	0.0419	0.1803	0.0416
6	Dataset with GPF and t-SNE	0.7237	0.2214	0.1732	0.0430
7	PCA	0.8549	0.0334	0.0450	0.0300
8	t-SNE	0.9093	0.2374	0.0043	0.0079
9	PCA and GPF	0.7108	0.0178	0.1267	0.0312
10	t-SNE and GPF	0.7888	0.0115	0.0555	0.0190

Without applying SMOTE, the KNN classifier on Dataset 2 showed limited effectiveness in identifying promotable employees. While some configurations achieved high accuracy, such as the Dataset with PCA (0.9119) and the Dataset with t-SNE, these setups demonstrated very low recall and F1-scores, with F1 as low as 0.0079 in t-SNE, confirming that the model essentially predicted only the majority class.

The configuration with the highest F1-score was the Dataset with PCA as an added feature (F1 = 0.1169), showing a slight improvement in balance, but still insufficient for practical application. Combinations involving GPF generally increased recall (up to ~0.18) but caused severe drops in precision, indicating that GPF alone cannot compensate for imbalance in the dataset. This trend was consistent across most variations, where precision-recall trade-offs were unnecessary.

These results demonstrated that high accuracy on its own was misleading in the presence of class imbalance. The model failed to predict the minority promotion class with meaningful accuracy, which was significant in HR contexts. As seen in Dataset 1, SMOTE or other balancing techniques will be essential for enabling KNN to perform reliably in promotion prediction tasks.

Table 4.16 Performance Comparison of KNN for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.8357	0.7578	0.9951	0.8595
2	Dataset with GPF and SMOTE	0.6991	0.6627	0.9967	0.7835
3	Dataset with PCA and SMOTE	0.7790	0.6979	0.9922	0.8188
4	Dataset with t-SNE and SMOTE	0.7717	0.6922	0.9852	0.8126
5	Dataset with GPF, PCA, and SMOTE	0.7257	0.6881	0.9951	0.8007
6	Dataset with GPF, t-SNE, and SMOTE	0.7495	0.7119	0.9908	0.8154
7	PCA and SMOTE	0.6965	0.6579	0.8684	0.7452
8	t-SNE and SMOTE	0.6888	0.6681	0.7581	0.7096
9	PCA, GPF and SMOTE	0.6777	0.6464	0.8784	0.7394
10	t-SNE, GPF and SMOTE	0.6671	0.6495	0.8259	0.7206

The use of SMOTE significantly enhanced the performance of the KNN classifier on Dataset 2. The baseline configuration (SMOTE only) achieved the best balance across all metrics, with an accuracy of 0.8357, precision of 0.7578, recall of 0.9951, and an F1-score of 0.8595, demonstrating the model's strong ability to identify promotable employees while minimizing false positives.

Several other configurations also delivered high performance. For example, the Dataset with GPF, t-SNE, and SMOTE and the Dataset with PCA and SMOTE achieved F1-scores of 0.8154 and 0.8188, respectively, with substantial precision and recall values, indicating consistent prediction stability. GPF-inclusive setups often increased recall (frequently above 0.99), but in some cases, this was accompanied by a slight drop in precision.

While dimensionality reduction techniques like PCA and t-SNE slightly reduced accuracy in some combinations, their inclusion with GPF consistently yielded robust F1-scores in the range of 0.72–0.80. These results suggest that SMOTE alone can dramatically improve KNN performance. However, optimal results are achieved when SMOTE is combined with GPF and dimensionality reduction for feature enhancement and noise mitigation.

Overall, the findings confirm that KNN becomes a highly reliable model for promotion prediction when data imbalance is addressed and relevant performance features are effectively engineered.

4.1.5 Logistic Regression Classification

Table 4.17 Performance Comparison of LR for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.8040	0.2379	0.1529	0.0644
2	Dataset with GPF	0.8023	0.2428	0.2221	0.1034
3	Dataset with PCA	0.7757	0.2286	0.1595	0.0676
4	Dataset with t-SNE	0.8122	0.2269	0.1368	0.0533
5	Dataset with GPF and PCA	0.8169	0.2850	0.2105	0.1016
6	Dataset with GPF and t-SNE	0.8120	0.2428	0.2152	0.0985
7	PCA	0.9142	0.0484	0.0032	0.0060
8	t-SNE	0.9148	0.0000	0.0000	0.0000
9	PCA and GPF	0.8605	0.0343	0.1655	0.0568
10	t-SNE and GPF	0.8717	0.0367	0.1471	0.0588

Without applying SMOTE, the Logistic Regression model on Dataset 1 showed moderate accuracy across several configurations but struggled to detect promotable employees effectively due to class imbalance. The configuration that yielded the best F1-score (0.1034) was the Dataset with GPF, with a reasonable trade-off between Precision (0.2428) and Recall (0.2221), indicating that domain-informed features helped improve class sensitivity to some extent.

Combining GPF with PCA (Dataset with GPF and PCA) also produced a competitive performance (F1 = 0.1016), while retaining high accuracy (0.8169). On the other hand, configurations relying solely on dimensionality reduction, such as PCA only or t-SNE only, showed extremely poor recall and F1-scores, despite achieving the highest accuracy (~0.91). These results suggest that high accuracy was not necessarily indicative of actual model performance in imbalanced settings.

Overall, the model's low F1-scores and limited recall across most configurations demonstrate that Logistic Regression is particularly sensitive to imbalance. Although GPF improves performance marginally, balancing techniques like SMOTE are necessary to enable LR to generalize meaningfully in HR promotion prediction tasks.

Table 4.18 Performance Comparison of LR for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.8553	0.8517	0.8609	0.8537
2	Dataset with GPF and SMOTE	0.8574	0.8510	0.8814	0.8641
3	Dataset with PCA and SMOTE	0.7854	0.7923	0.8409	0.8088
4	Dataset with t-SNE and SMOTE	0.7060	0.6892	0.7967	0.7357
5	Dataset with GPF, PCA, and SMOTE	0.7645	0.7682	0.8312	0.7907
6	Dataset with GPF, t-SNE, and SMOTE	0.7556	0.7577	0.8191	0.7801
7	PCA and SMOTE	0.5655	0.5897	0.6247	0.5959
8	t-SNE and SMOTE	0.5084	0.5163	0.5802	0.5432
9	PCA, GPF and SMOTE	0.7232	0.7977	0.7728	0.7627
10	t-SNE, GPF and SMOTE	0.7240	0.7986	0.7740	0.7636

The application of SMOTE to Dataset 1 significantly enhanced the performance of the Logistic Regression model across all evaluation metrics. The best overall configuration was the Dataset with GPF and SMOTE, which achieved an F1-score of 0.8641, with balanced performance in Accuracy = 0.8574, Precision = 0.8510, and Recall = 0.8814. This suggests that the model was both precise and sensitive in identifying employees who are promotable.

Similarly, the baseline SMOTE configuration (without GPF) also performed well, with F1 = 0.8537, indicating that class balancing alone already produced a substantial improvement. Configurations combining SMOTE with PCA or t-SNE yielded slightly lower performance but still maintained strong F1-scores in the range of 0.73-0.81, particularly when combined with GPF.

Lower performance was observed in PCA or t-SNE-only configurations (e.g., F1 \approx 0.59-0.54), despite achieving decent accuracy, indicating that dimensionality reduction alone was insufficient for robust classification. The inclusion of GPF consistently improved recall and precision in these cases, restoring model effectiveness.

Overall, the results demonstrate that Logistic Regression can be a valuable and interpretable model for promotion prediction when supported by SMOTE and domain-informed features, such as GPF. The combination of GPF and SMOTE stands out as the most reliable configuration for achieving high recall without losing precision.

Table 4.19 Performance Comparison of LR for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.7929	0.4639	0.1455	0.0511
2	Dataset with GPF	0.8012	0.3344	0.2167	0.0767
3	Dataset with PCA	0.9050	0.8571	0.1335	0.1299
4	Dataset with t-SNE	0.9154	0.0000	0.0000	0.0000
5	Dataset with GPF and PCA	0.8033	0.4521	0.2195	0.0914
6	Dataset with GPF and t-SNE	0.7943	0.4214	0.1966	0.0428
7	PCA	0.9154	0.0000	0.0000	0.0000
8	t-SNE	0.9154	0.0000	0.0000	0.0000
9	PCA and GPF	0.8401	0.0238	0.1387	0.0406
10	t-SNE and GPF	0.9130	0.0847	0.0786	0.0815

Without the use of SMOTE, the Logistic Regression model on Dataset 2 demonstrated moderate to high accuracy across most configurations but failed to provide meaningful recall or F1-scores. Notably, the PCA-only and t-SNE-only configurations yielded the highest accuracy (both 0.9154), but produced zero precision, recall, and F1-score, indicating that the model ignored the minority class entirely.

The best F1-score (0.1299) was achieved by the Dataset with PCA, which also exhibited a very high precision (0.8571) but an extremely low recall (0.1335), indicating a highly conservative model that identified a small number of promotable employees correctly. Configurations using GPF, particularly the Dataset with GPF and PCA (F1 = 0.0914), improved balance slightly, but were still insufficient to handle the imbalance without oversampling.

These results emphasized the limitation of relying on accuracy alone in class-imbalanced HR datasets. Although some configurations showed promise in terms of precision, the consistently low recall and F1-scores across the board highlighted the need for techniques like SMOTE to enable the model to generalize effectively in promoting employees.

Table 4.20 Performance Comparison of LR for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.8515	0.8503	0.8899	0.8658
2	Dataset with GPF and SMOTE	0.8829	0.8834	0.9063	0.8925
3	Dataset with PCA and SMOTE	0.6820	0.6990	0.7393	0.7099
4	Dataset with t-SNE and SMOTE	0.7852	0.7871	0.8521	0.8117
5	Dataset with GPF, PCA, and SMOTE	0.7719	0.8175	0.8317	0.8051
6	Dataset with GPF, t-SNE, and SMOTE	0.8435	0.8546	0.8904	0.8650
7	PCA and SMOTE	0.5396	0.5822	0.5724	0.5642
8	t-SNE and SMOTE	0.4884	0.4887	0.5007	0.4945
9	PCA, GPF and SMOTE	0.7193	0.7974	0.7630	0.7575
10	t-SNE, GPF and SMOTE	0.7207	0.7969	0.7658	0.7596

After applying SMOTE, the Logistic Regression model on Dataset 2 showed a substantial improvement across all evaluation metrics. The most effective configuration was the Dataset with GPF and SMOTE, which achieved the highest overall performance with Accuracy = 0.8829, Precision = 0.8834, Recall = 0.9063, and F1-score = 0.8925. This indicated a well-balanced model with excellent generalization and low misclassification for both majority and minority classes.

Other top-performing configurations included a Dataset with GPF, t-SNE, and SMOTE (F1 = 0.8650) and the baseline SMOTE setup (F1 = 0.8658), demonstrating that both dimensionality reduction and domain-informed features, such as GPF, can complement SMOTE to enhance model robustness further. Even combinations like t-SNE and SMOTE, as well as PCA and SMOTE, while yielding lower accuracy, still showed reasonable F1-scores (~0.49–0.56), confirming the overall benefit of balancing.

In contrast, although configurations without GPF performed reasonably well, adding GPF consistently improved recall and stabilized F1-scores across variations. This emphasizes that GPF, when combined with SMOTE and optionally PCA or t-SNE, can yield a stable classification pipeline for predicting HR promotions.

Overall, the findings confirm that Logistic Regression, despite its simplicity, performs excellently when enhanced with GPF and SMOTE, offering an interpretable and accurate approach for modeling promotional decisions in imbalanced HR datasets.

4.1.6 Neural Network Classification

Table 4.21 Performance Comparison of NN for Dataset 1

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.5301	0.5453	0.5026	0.1004
2	Dataset with GPF	0.5222	0.5445	0.5165	0.1062
3	Dataset with PCA	0.5284	0.5347	0.5219	0.0942
4	Dataset with t-SNE	0.5247	0.5238	0.5114	0.1411
5	Dataset with GPF and PCA	0.5052	0.1490	0.5221	0.1148
6	Dataset with GPF and t-SNE	0.5248	0.4720	0.5165	0.1118
7	PCA	0.7051	0.0470	0.3126	0.0865
8	t-SNE	0.8147	0.0375	0.0844	0.0545
9	PCA and GPF	0.5679	0.0499	0.4998	0.0899
10	t-SNE and GPF	0.5913	0.0531	0.5000	0.0962

Without SMOTE, the Neural Network model on Dataset 1 yielded inconsistent results across different feature configurations. Despite moderate accuracy scores (typically around 0.52–0.53), the F1-scores remained low in most cases due to poor balance between precision and recall. For example, t-SNE only provided the highest accuracy (0.8147), but had an extremely low F1-score (0.0545) due to inadequate recall.

The most balanced configuration was t-SNE as an added feature (F1 = 0.1411), with more stable precision and recall. Combinations such as Dataset with GPF and PCA and Dataset with GPF and t-SNE offered slight improvements in recall, but F1-scores still hovered around 0.10–0.11, which was insufficient for practical use. While GPF inclusion did slightly boost model sensitivity to the minority class, it could not overcome the limitations imposed by imbalanced data.

These results emphasize the importance of applying class-balancing techniques when using deep models, such as neural networks, on imbalanced datasets. Without SMOTE, even complex models tended to predict the majority class, resulting in misleading accuracy and ineffective promotion prediction in HR scenarios.

Table 4.22 Performance Comparison of NN for Dataset 1 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.5542	0.5187	0.9751	0.6788
2	Dataset with GPF and SMOTE	0.5343	0.5199	0.9831	0.6789
3	Dataset with PCA and SMOTE	0.5759	0.5522	0.9386	0.6889
4	Dataset with t-SNE and SMOTE	0.5375	0.5280	0.9219	0.6617
5	Dataset with GPF, PCA, and SMOTE	0.5195	0.5131	0.9522	0.6613
6	Dataset with GPF, t-SNE, and SMOTE	0.5347	0.5103	0.9031	0.6565
7	PCA and SMOTE	0.5187	0.5103	0.7937	0.6330
8	t-SNE and SMOTE	0.4859	0.5035	0.7131	0.5774
9	PCA, GPF and SMOTE	0.4458	0.4671	0.7794	0.5816
10	t-SNE, GPF and SMOTE	0.4800	0.4783	0.7106	0.5542

After applying SMOTE, the Neural Network model on Dataset 1 showed substantial improvements in its ability to detect promotable employees. While accuracy across configurations ranged from 0.48 to 0.57, the F1-scores improved dramatically compared to the pre-SMOTE results. The most balanced configuration was the Dataset with PCA and SMOTE, achieving an F1-score of 0.6889, with Precision = 0.5522 and Recall = 0.9386, indicating a well-balanced model that effectively minimized false negatives.

Other strong configurations included a Dataset with GPF and SMOTE (F1 = 0.6789) and Baseline SMOTE (F1 = 0.6788), with recall values approaching 0.98, showing high sensitivity to minority class instances. Although these combinations sacrificed some precision (~0.51–0.52), they maintained respectable F1-scores due to excellent recall.

Dimensionality reduction with t-SNE and the inclusion of GPF tended to produce slightly lower F1-scores (~0.55–0.66) but still offered significantly better results than non-SMOTE counterparts. This suggests that while neural networks can overfit or become unstable with high-dimensional features, SMOTE remains a robust stabilizer, especially when paired with relevant performance-based features, such as GPF.

Overall, SMOTE enables the Neural Network to shift from underperforming (F1 \approx 0.1) to producing strong and generalizable promotion predictions. Feature

configurations involving PCA and/or GPF are the most effective in optimizing the trade-off between sensitivity and precision.

Table 4.23 Performance Comparison of NN for Dataset 2

No.	Dataset Combination	AC	PS	RC	F1
1	Original dataset	0.5193	0.5434	0.5059	0.1346
2	Dataset with GPF	0.5152	0.0436	0.5000	0.0849
3	Dataset with PCA	0.5175	0.5249	0.5062	0.0927
4	Dataset with t-SNE	0.9154	0.0000	0.0000	0.0000
5	Dataset with GPF and PCA	0.5105	0.5436	0.5120	0.0898
6	Dataset with GPF and t-SNE	0.5152	0.5436	0.5000	0.0912
7	PCA	0.8992	0.0335	0.1499	0.0230
8	t-SNE	0.9154	0.0000	0.0012	0.0000
9	PCA and GPF	0.6364	0.0605	0.5000	0.1061
10	t-SNE and GPF	0.6564	0.0622	0.5000	0.1109

Without the use of SMOTE, the Neural Network model on Dataset 2 struggled with class imbalance, resulting in unreliable performance despite achieving seemingly high accuracy in some configurations. Specifically, the t-SNE and PCA-only configurations achieved an accuracy of over 0.90, but their recall and F1-scores were close to zero, indicating that the model completely ignored the promotable class.

The configuration with the most balanced output was the original dataset (F1 = 0.1346), with precision and recall approximately 0.54 and 0.50, respectively. Other datasets, such as PCA and GPF, t-SNE and GPF, and the Dataset with GPF and t-SNE, maintained a recall near 0.50 but failed to convert this into meaningful F1-scores due to extremely low precision.

These results demonstrated that Neural Networks, despite their modeling power, are highly sensitive to class imbalance. Even with GPF integration, which slightly improved sensitivity, the absence of SMOTE caused the model to favor the majority class heavily, rendering high accuracy misleading.

Overall, without SMOTE, Neural Network models are unsuitable for promotion prediction tasks, as they fail to recognize minority instances even with feature engineering support. Proper balancing mechanisms were essential for leveraging the full potential of deep learning in HR analytics.

Table 4.24 Performance Comparison of NN for Dataset 2 with SMOTE

No.	Dataset Combination	AC	PS	RC	F1
1	Dataset with SMOTE	0.7614	0.7552	0.9744	0.6025
2	Dataset with GPF and SMOTE	0.7504	0.7528	0.9918	0.8208
3	Dataset with PCA and SMOTE	0.5192	0.7546	0.9849	0.8144
4	Dataset with t-SNE and SMOTE	0.7393	0.7529	0.8766	0.7693
5	Dataset with GPF, PCA, and SMOTE	0.7257	0.7530	0.9909	0.8016
6	Dataset with GPF, t-SNE, and SMOTE	0.7398	0.7402	0.9866	0.7634
7	PCA and SMOTE	0.4302	0.2310	0.5552	0.6417
8	t-SNE and SMOTE	0.4904	0.4920	0.5156	0.5328
9	PCA, GPF and SMOTE	0.3510	0.4912	0.5224	0.4008
10	t-SNE, GPF and SMOTE	0.5662	0.5070	0.7229	0.5757

With the application of SMOTE, the Neural Network on Dataset 2 demonstrated significant performance improvements in both recall and F1-score. The best overall result was achieved by the Dataset with GPF and SMOTE, yielding an F1-score of 0.8208, accuracy of 0.7504, precision of 0.7528, and recall of 0.9918. This balance suggests that the model was highly effective at detecting promotable employees with minimal loss in precision.

Other top-performing configurations included Dataset with PCA and SMOTE (F1 = 0.8144) and Dataset with GPF, PCA, and SMOTE (F1 = 0.8016). These setups maintained high recall (above 0.98) while varying slightly in precision, demonstrating that even simple transformations, such as PCA, when combined with SMOTE, helped the model become more generalizable to the minority class.

Overall, these findings confirm that SMOTE, especially when combined with GPF and dimensionality reduction, enables Neural Networks to become highly effective in detecting promotable employees. Among the various configurations, GPF-enhanced models showed consistent recall and high F1-scores, affirming the strength of the proposed feature in HR promotion prediction.

4.2 Clustering Results

Clustering used 2 algorithms: K-Means clustering and Fuzzy clustering.

Table 4.25 Performance Comparison of K-Means Clustering for Dataset 1

No.	Dataset Combination	RI	FMI	MI	VMS
1	Without all components	0.4239	0.5495	0.0017	0.0025
2	SMOTE	0.5016	0.4354	0.0164	0.0187
3	GPF and SMOTE	0.5043	0.485	0.0842	0.0948
4	PCA and SMOTE	0.5057	0.5121	0.012	0.0173
5	t-SNE and SMOTE	0.5	0.4082	0	0
6	GPF, PCA, and SMOTE	0.615	0.6156	0.1205	0.174
7	GPF, t-SNE, and SMOTE	0.5803	0.5146	0.1127	0.1277

Table 4.26 Performance Comparison of K-Means Clustering for Dataset 2

No.	Dataset Combination	RI	FMI	MI	VMS
1	Without all components	0.5034	0.6413	0.0002	0.0005
2	SMOTE	0.5075	0.4403	0.0192	0.0219
3	GPF and SMOTE	0.5909	0.4643	0.0373	0.0442
4	PCA and SMOTE	0.6215	0.6465	0.1565	0.2356
5	t-SNE and SMOTE	0.5	0.4082	0	0
6	GPF, PCA, and SMOTE	0.6839	0.683	0.1984	0.2834
7	GPF, t-SNE, and SMOTE	0.5727	0.5077	0.1011	0.1149

As illustrated in Tables 4.25 and 4.26, the datasets enriched with a combination of GPF, PCA, and SMOTE achieved the highest clustering performance when evaluated using the K-means algorithm. This configuration consistently outperformed all other combinations across both datasets. In general, incorporating any of the enhancement techniques, including GPF, PCA, or SMOTE, to investigate which technique could provide better clustering results compared to the use of the original dataset alone. However, one notable exception was the t-SNE and SMOTE combination, which did not yield significant performance improvements and, in some cases, slightly degraded clustering quality. These results suggest that PCA is a more effective dimensionality reduction method than t-SNE for the datasets used in this

study. The possible reason is related to PCA, which preserves global variance structures and provides a more stable feature representation for clustering algorithms. Moreover, the inclusion of GPF proved beneficial across all scenarios, as it helps enhance the quality of the feature space for both PCA and t-SNE combinations. Moreover, these results confirm that GPF played a valuable role in improving clustering performance, regardless of the dimensionality reduction method used.

Table 4.27 Performance Comparison of Fuzzy Clustering for Dataset 1

No.	Dataset Combination	RI	FMI	MI	VMS
1	Without all components	0.5006	0.6501	0.0002	0.0005
2	SMOTE	0.5003	0.5048	0.0002	0.0002
3	GPF and SMOTE	0.5165	0.5187	0.0167	0.0243
4	PCA and SMOTE	0.5085	0.5085	0.0086	0.0124
5	t-SNE and SMOTE	0.5001	0.5001	0.0001	0.0002
6	GPF, PCA, and SMOTE	0.6221	0.6226	0.1281	0.1849
7	GPF, t-SNE, and SMOTE	0.5591	0.5608	0.061	0.0883

Table 4.28 Performance Comparison of Fuzzy Clustering for Dataset 2

No.	Dataset Combination	RI	FMI	MI	VMS
1	Without all components	0.5142	0.6627	0.0006	0.0014
2	SMOTE	0.543	0.5433	0.0437	0.0631
3	GPF and SMOTE	0.6662	0.6667	0.1776	0.2565
4	PCA and SMOTE	0.591	0.5911	0.094	0.1357
5	t-SNE and SMOTE	0.5	0.5	0	0
6	GPF, PCA, and SMOTE	0.6728	0.6729	0.1846	0.2663
7	GPF, t-SNE, and SMOTE	0.596	0.5975	0.1003	0.1451

As presented in Tables 4.27 and 4.28, the fuzzy clustering results revealed that the combination of GPF, PCA, and SMOTE delivered the best clustering performance across both datasets. This finding was consistent with the results observed from the K-means clustering model, further validating the effectiveness of the proposed data enrichment approach. In general, all enhanced dataset configurations outperformed the original dataset, except for the t-SNE combined with SMOTE, which failed to produce meaningful improvements in clustering performance. For the fuzzy clustering model, PCA once again proved to be a more effective dimensionality reduction technique

compared to t-SNE. This supported the earlier observation that PCA's ability to capture global variance provided a more reliable foundation for clustering, particularly when used in combination with data balancing. Notably, the results also confirm that incorporating the Generated Promotion Feature (GPF) consistently improves clustering quality, regardless of whether PCA or t-SNE is used. This emphasizes GPF's value in enhancing feature representation and improving the model's ability to form more meaningful clusters.



CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Classification

5.1.1 Discussion of Classification Results Without SMOTE

In the absence of SMOTE, all classification models suffered from poor recall and F1-scores despite appearing to have high accuracy in some configurations. The imbalance in class distribution significantly influenced model behavior, resulting in the majority class dominating predictions. Among the six models tested, including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Neural Network (NN), Logistic Regression proved to be the most valuable under imbalanced conditions. Specifically, the combination of GPF and PCA helped improve recall while maintaining moderate precision, resulting in the most balanced F1-score among all models in this group. The GPF feature, when used without SMOTE, provided some benefit by increasing recall in specific models (e.g., LR, NN). Although it often came at the cost of reduced precision. PCA, when used as an additional feature (rather than for dimensionality reduction), helped specific models, such as KNN and NN, gain slight improvements. On the other hand, t-SNE, while improving accuracy in some cases, generally failed to enhance predictive power and often led to instability or overfitting, especially in models such as SVM and NN. Overall, the experiments demonstrate that while GPF and PCA can marginally support model performance under imbalanced conditions, the lack of SMOTE causes most classifiers to be ineffective in reliably identifying promoted employees. Specifically, SVM and DT perform the worst across both datasets and should not be used without data balancing, regardless of the feature engineering.

5.1.2 Discussion of Classification Results with SMOTE

After applying SMOTE to both datasets, all classification models showed significant improvements in their ability to identify promotable employees. Across almost all models, recall and F1-scores became substantially more balanced. This confirms the effectiveness of SMOTE in addressing class imbalance and enhancing model generalizability. Among the six models evaluated, LR and NN demonstrated the most consistent and well-balanced results across datasets. The inclusion of GPF was especially effective in these models, contributing to both high recall and stable precision. Combinations such as (GPF and SMOTE), or (GPF, PCA, and SMOTE), yielded top F1-scores, often above 0.85. PCA also proved valuable, especially when used as an additional feature rather than for dimensionality reduction. Models like KNN, NN, and LR benefited from PCA-enhanced representations, improving F1-scores in configurations like (PCA and SMOTE) or (GPF, PCA and SMOTE). However, PCA should be used with care in models such as RF or DT, where performance gains were marginal or inconsistent. t-SNE, on the other hand, yielded mixed results. While it improved recall in some models (e.g., KNN, NN), it did not consistently enhance the F1-score. It sometimes introduced instability, especially in models prone to overfitting, like NN or less robust ones like DT. t-SNE performed best when used together with GPF and SMOTE. However, it should not be relied upon as a standalone enhancer. In conclusion, GPF was the most universally beneficial feature, as it improved recall and overall interpretability in nearly every model. PCA is recommended especially for KNN, LR, and NN, while t-SNE should be used cautiously and only in supportive roles. When SMOTE is used in tandem with GPF and dimensionality-aware transformations, classification models become substantially more reliable for promotion prediction tasks in human resource analytics.

5.2 Clustering

5.2.1 Effectiveness of Dimensionality Reduction Techniques

The experimental results indicate that integrating Principal Component Analysis (PCA) into the clustering pipeline significantly improves performance compared to using the original dataset alone. PCA outperformed t-SNE in terms of cluster quality for both K-means and Fuzzy C-means models. This distinction can be attributed to the fundamental differences between the two techniques. PCA, as a linear dimensionality reduction method, captures large-scale patterns and maximizes variance across the dataset, which is particularly effective for identifying global structures. In contrast, t-SNE is a non-linear approach that preserves local similarity at the expense of distorting global relationships.

To assess the underlying structure of the dataset, PCA was first applied to examine how variance was distributed among principal components. The results indicate that a small number of components explain the majority of the variance, suggesting that the dataset lies on a lower-dimensional linear subspace. This observation, paired with the superior clustering performance of PCA over t-SNE, provides strong evidence that the dataset predominantly exhibits linear characteristics.

5.2.2 Role of Generated Promotion Feature (GPF)

The inclusion of the Generated Promotion Feature (GPF) significantly enhances clustering performance. GPF is designed as a composite performance-based variable derived from the top three performance-oriented features that exhibit the highest Pearson correlation with the promotion target variable. These features typically include KPI achievement, awards, and average training performance metrics commonly used in HR assessments.

A trade-off between interpretability and model generalizability guided the decision to use only the top three features. This limited but informative set of features prevents redundancy and reduces the risk of overfitting. A sensitivity analysis confirmed that the addition of GPF improved performance metrics, particularly recall and F1-score, across multiple models, especially under class-imbalanced conditions.

This demonstrates the value of GPF in promoting more accurate identification of promotable employees.

5.2.3 Enhancing Clustering with Balanced Data Using SMOTE

To further improve model stability and reduce bias, the Synthetic Minority Oversampling Technique (SMOTE) was applied to address the inherent class imbalance in the dataset. By generating synthetic instances of underrepresented promotion cases, SMOTE ensured that both majority and minority classes were adequately represented during clustering.

The combined use of PCA, GPF, and SMOTE yielded a well-structured and enriched feature space, consistently improving clustering outcomes. This integrated approach not only enhances model accuracy but also improves interpretability and robustness, affirming its effectiveness for promotion analysis in human resource contexts.

5.3 Conclusion

This study presents a comprehensive data enrichment framework designed to enhance the effectiveness of classification and clustering models in identifying promotable employees.

For classification tasks, the results demonstrate that handling class imbalance with SMOTE alone significantly improves model performance, particularly in terms of recall and F1-score. Therefore, the choice of SMOTE is strongly supported by its consistent effectiveness across models and datasets. The effectiveness of SMOTE was observed across both datasets, indicating that the success was not data-specific but rather a robust approach to imbalance. PCA and GPF may still be helpful in certain situations, but they are not required for achieving strong classification results in promotion prediction tasks. Although techniques like PCA or GPF were explored, they were not essential for improving classification outcomes in all cases. PCA offered limited additional value, and GPF was only helpful when further increasing recall was needed.

For clustering tasks, the proposed methodology was evaluated using K-means and Fuzzy clustering, known as clustering algorithms, applied to two publicly available HR datasets. A variety of dataset combinations were tested, involving different configurations of feature extraction (PCA/t-SNE), GPF augmentation, and SMOTE balancing. Experimental results show that the combination of GPF, PCA, and SMOTE consistently outperformed other configurations across both datasets. This combination leads to significant improvements in clustering quality, as measured by multiple evaluation metrics including the Rand Index (RI), Fowlkes–Mallows Index (FMI), Mutual Information (MI), and V-measure (V). Overall, the findings support the conclusion that combining performance-driven feature engineering with dimensionality reduction and class balancing strategies can substantially enhance the ability of clustering models to identify meaningful patterns in employee promotion data.

5.4 Suggestions and Future Work

This study provides practical implications for utilizing machine learning in human resource management, particularly in promotion analysis and decision support. The developed framework can be applied in two main approaches: classification and clustering.

In the context of classification, the model can be employed to predict promotion outcomes, which requires a well-prepared training dataset containing historical promotion records and relevant performance features. Before applying, organizations should ensure that data preprocessing, feature selection, and model validation are conducted thoroughly to ensure reliable predictions. This application is especially beneficial for supporting fair and transparent promotion decisions.

In the clustering context, the proposed model can be used without the need for labeled outcome data. Instead, organizations can structure input datasets using either performance features (e.g., KPI scores, awards, training results) or personal attributes (e.g., years of service, education, skills). By feeding these features into the clustering model, the algorithm will automatically group employees based on performance levels or similarity patterns. This unsupervised approach supports strategic workforce

planning, such as identifying high-potential employee clusters or segmenting training needs.

The proposed model offers a promising decision-support tool for human resource analytics. With proper adaptation and integration into HR practices, it can support organizations in identifying promotable employees, promoting fairness, and optimizing talent management strategies. To mitigate potential biases in promotion decisions, the proposed model is designed with a strong emphasis on performance-driven and explainable criteria. Specifically, the model integrates a domain-informed feature called the Generated Promotion Feature (GPF), which is derived solely from quantifiable performance-related attributes such as KPIs, training scores, and award recognition. These features are objectively measurable and are selected based on their statistical correlation with historical promotion outcomes, rather than personal or demographic factors.

One notable limitation of this study is that the feature selection process for GPF construction was based on predefined correlation scores, which were derived through manual observation and analysis. Although this approach may not capture the full complexity of the data, it is still effective. Therefore, future research should explore automated, data-driven feature selection techniques to enhance the objectivity and adaptability of GPF creation. Additionally, extending the experimentation to include other HR datasets, different clustering algorithms, and alternative methods for handling imbalanced data would help validate and strengthen the generalizability of the proposed approach.

REFERENCES

- Adnan, M., Habib, A., Ashraf, J., Shah, B., & Ali, G. (2020). Improving M-learners' performance through deep learning techniques by leveraging features weights. *IEEE Access*, 8, 131088–131106.
<https://doi.org/10.1109/ACCESS.2020.3007727>
- Alqahtani, F. A., & Almaleh, A. (2022). Analysis and prediction of employee promotions using machine learning. In *5th International Conference on Data Science and Information Technology (DSIT)* (pp. 1–9). IEEE.
<https://doi.org/10.1109/DSIT55861.2022.00005>
- Aottiwerch, N., & Kokaew, U. (2018). The analysis of matching learners in pair programming using K-means. In *5th International Conference on Industrial Engineering and Applications (ICIEA)* (pp. 362–366). IEEE.
<https://doi.org/10.1109/IEA.2018.8387125>
- Asim, Y., Raza, B., Malik, A. K., Rathore, S., & Bilal, A. (2018). Improving the performance of professional blogger's classification. In *International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICOMET.2018.8346342>
- Assiri, A., Berri, J., & Chikh, A. (2012). Classification and tendencies of evaluations in e-learning. In *International Conference on Education and e-Learning Innovations* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICEELI.2012.6360570>
- Bagdadli, S., Roberson, Q. M., & Paoletti, F. (2006). The mediating role of procedural justice in responses to promotion decisions. *Journal of Business and Psychology*, 21(1), 83–102. <https://doi.org/10.1007/s10869-005-9026-2>
- Chan, D., Rao, R., Huang, F., & Canny, J. F. (2018). T-SNE-CUDA: GPU-accelerated T-SNE and its applications to modern data. In *30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)* (pp. 330–338). IEEE. <https://doi.org/10.1109/SBAC-PAD.2018.00055>
- Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 875–886). Springer. https://doi.org/10.1007/978-0-387-09823-4_45

- Chen, Q., & Gong, Z. (2013). Data mining modeling of employee engagement for IT enterprises based on decision tree algorithm. In *6th International Conference on Information Management, Innovation Management and Industrial Engineering* (pp. 305–308). IEEE. <https://doi.org/10.1109/ICIII.2013.6703145>
- Dang, Q., Truong, M., & Huynh, M. (2021). Studying the fuzzy clustering methods to understand employee performance. In *4th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 541–544). IEEE. <https://doi.org/10.1109/ICAIBD51990.2021.9458974>
- De Oliveira Goes, A. S., & De Oliveira, R. C. L. (2020). A process for human resource performance evaluation using computational intelligence: An approach using a combination of rule-based classifiers and supervised learning algorithms. *IEEE Access*, 8, 39403–39419. <https://doi.org/10.1109/ACCESS.2020.2975485>
- Dias da Silva, A., & Van der Klaauw, B. (2006). Wage dynamics and promotions inside and between firms. *Journal of Population Economics*, 24, 1513–1548. <https://doi.org/10.1007/s00148-009-0274-8>
- Ding, X., & Tang, Y. (2013). Exploring of clustering algorithm on class-imbalanced data. In *8th International Conference on Computer Science & Education (ICCSE)* (pp. 89–93). IEEE. <https://doi.org/10.1109/ICCSE.2013.6553903>
- Duan, Y., Niu, X., & Nie, G. (2018). Data augmentation based on interest points of feature. In *Proceedings of the International Conference on Digital Image Processing*. IEEE.
- Dutsinma, L. I. F., & Temdee, P. (2020). VARK learning style classification using decision tree with physiological signals. *Wireless Personal Communications*, 111, 2875–2896. <https://doi.org/10.1007/s11277-020-07196-3>
- Eminagaoglu, M., & Eren, S. (2010). Implementation and comparison of machine learning classifiers for information security risk analysis of a human resources department. In *International Conference on Computer Information Systems and Industrial Management Applications (CISIM)* (pp. 187–192). IEEE. <https://doi.org/10.1109/CISIM.2010.5643665>

- Feng, H., Duan, L., Liu, S., & Liu, S. (2020). Entity hierarchical clustering method based on multi-channel and t-SNE dimension reduction. *IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 9, 2155–2159. <https://doi.org/10.1109/ITAIC49862.2020.9339079>
- Gathungu, E., Iravo, M., & Namusonge, S. (2015). Effect of promotion strategies on the organizational commitment of banking sector employees in Kenya. *IOSR Journal of Business and Management*, 20(10), 36–45. <https://doi.org/10.9790/0837-201013645>
- Guo, Z., & Zhang, Y. (2010). The third-party logistics performance evaluation based on the AHP-PCA model. In *International Conference on E-Product E-Service and E-Entertainment* (pp. 1–4). IEEE. <https://doi.org/10.1109/ICEEE.2010.5660810>
- Guohao, Q., Bin, W., Bai, W., & Baoli, Z. (2019). Competency analysis in human resources using text classification based on deep neural network. In *IEEE Fourth International Conference on Data Science in Cyberspace (DSC)* (pp. 322–329). IEEE. <https://doi.org/10.1109/DSC.2019.00056>
- Hartanto, A. D., Utami, E., Adi, S., & Hudnanto, H. S. (2019). Job seeker profile classification of twitter data using the naïve bayes classifier algorithm based on the DISC method. In *4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 533–536). IEEE. <https://doi.org/10.1109/ICITISEE48480.2019.9003963>
- Hatanaka, S., & Nishi, H. (2021). Efficient GAN-based unsupervised anomaly sound detection for refrigeration units. In *IEEE 30th International Symposium on Industrial Electronics (ISIE)* (pp. 1–7). IEEE. <https://doi.org/10.1109/ISIE45552.2021.9576182>
- He, M., Zhu, Y., Lv, N., & He, R. (2022). A feature fusion-based representation learning model for job recommendation. In *2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (pp. 791–794). IEEE. <https://doi.org/10.1109/ICCECE54139.2022.9712783>

- Hoon, G. K., Min, G. K., Wong, O., Pin, O. B., & Sheng, C. Y. (2015). Classifly: Classification of experts by their expertise on the fly. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (pp. 245–246). IEEE. <https://doi.org/10.1109/WI-IAT.2015.63>
- Huang, Y. B. (2009). Study of college human resources data mining based on the SOM algorithm. *Asia-Pacific Conference on Information Processing, 1*, 324–327. <https://doi.org/10.1109/APCIP.2009.89>
- Huang, Z., & Jiang, D. (2011). Research and implementation of fuzzy ISODATA clustering algorithm based on gene expression programming in human resource management. In *International Conference of Information Technology, Computer Engineering and Management Sciences* (pp. 178–180). IEEE. <https://doi.org/10.1109/ICM.2011.361>
- Ilwani, M., Nassreddine, G., & Younis, J. A. (2023). Machine learning application on employee promotion. *Mesopotamian Journal of Computer Science*, 3(1), 45–55.
- Isaac, R. G., Zerbe, W. J., & Pitt, D. C. (2001). Leadership and motivation: The effective application of expectancy theory. *Journal of Managerial Issues*, 13(2), 212–226.
- Jaffar, Z., Noor, W., & Kanwal, Z. (2019). Predictive human resource analytics using data mining classification techniques. *International Journal of Computer (IJC)*, 32(1), 9–20.
- Jamil, M., Liu, H., Phatak, A., & Memmert, D. (2021). An investigation identifying which key performance indicators influence the chances of promotion to the Elite leagues in professional European football. *International Journal of Performance Analysis in Sport*, 21(4), 641–650. <https://doi.org/10.1080/24748668.2021.1933845>
- Jay, S. (2023). *Promotion rate: How to calculate & improve this key HR metric*. <https://www.aihr.com/blog/promotion-rate/#what>
- Jing, H. (2009). Application of fuzzy data mining algorithm in performance evaluation of human resource. In *International Forum on Computer Science-Technology and Applications* (pp. 343–346). IEEE. <https://doi.org/10.1109/IFCSTA.2009.90>

- Juvitayapun, T. (2021). Employee turnover prediction: The impact of employee event features on interpretable machine learning methods. In *13th International Conference on Knowledge and Smart Technology (KST)* (pp. 181–185). IEEE. <https://doi.org/10.1109/KST51265.2021.9415794>
- Kaewwiset, T., & Temdee, P. (2022). Promotion classification using decision tree and principal component analysis. In *Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)* (pp. 489–492). IEEE. <https://doi.org/10.1109/ECTIDAMTNCON54331.2022.1234567>
- Kaewwiset, T., Temdee, P., & Yooyativong, T. (2021). Employee classification for personalized professional training using machine learning techniques and SMOTE. In *Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference* (pp. 376–379). IEEE. <https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425754>
- Kok, Y. W. (2017). Investigation on promotional criteria based on human resource practices for better organizational effectiveness. *Journal of Contemporary Issues and Thought*, 7, 68–78. <https://doi.org/10.37134/jcit.vol7.7.2017>
- Kumar, S. (2021). *HR analysis case study datasets used for learning purpose*. <https://www.kaggle.com/shivan118/hranalysis>
- Kumar, S. (2020). *HR analysis case study*. <https://www.kaggle.com/shivan118/hranalysis>
- Lai, C., & Wei, M. (2007). A common weighted performance evaluation process by using data envelopment analysis models. In *IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 827–831). IEEE. <https://doi.org/10.1109/IEEM.2007.4419296>
- Li, X., & Zhou, Q. (2019). Research on improving SMOTE algorithms for unbalanced data set classification. In *International Conference on Electronic Engineering and Informatics (EEI)* (pp. 476–480). IEEE. <https://doi.org/10.1109/EEI48997.2019.00099>

- Ligare, B. S., Wanyama, K. W., & Aliata, V. L. (2020). Job promotion and employee performance among the administration police in Bungoma County, Kenya. *Cross Current International Journal of Economics, Management and Media Studies*, 2(2), 111–118. <https://doi.org/10.36344/ccijemms.2020.v02i02.002>
- Lingling, X. (2010). Applying grey relation clustering and PCA to performance evaluation of vendors in fresh milk supply chain. *International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, 2, 932–935. <https://doi.org/10.1109/ICLSIM.2010.5461169>
- Liu, J., Wang, T., Li, J., Huang, J., Yao, F., & He, R. (2019). A data-driven analysis of employee promotion: The role of the position of organization. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)* (pp. 4056–4062). IEEE. <https://doi.org/10.1109/SMC.2019.8914457>
- Liu, L., Wang, Q., Dong, M., Zhang, Z., Li, Y., Wang, Z., . . . Wang, S. (2020). Application of K-Means++ algorithm based on t-SNE dimension reduction in transformer district clustering. In *Asia Energy and Electrical Engineering Symposium (AEEES)* (pp. 111–114). IEEE. <https://doi.org/10.1109/AEEES48850.2020.9121438>
- Liu, Y. (2021). Analysis of human resource management mode and its selection factors based on clustering algorithm. In *IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)* (pp. 535–538). IEEE. <https://doi.org/10.1109/AEECA52637.2021.9509594>
- Liu, Y., Xie, C., Chen, X., Zhu, Q., & Sun, Y. (2023). Intelligent collection system of human resource information based on clustering algorithm. In *International Conference on Signal Processing and Communication Technology (SPCT 2022)*. IEEE. <https://doi.org/10.1117/12.2673991>
- Long, Y., Liu, J., Fang, M., Wang, T., & Jiang, W. (2018). Prediction of employee promotion based on personal basic features and post features. In *International Conference on Data Processing and Applications*. IEEE. <https://doi.org/10.1145/>

- Mathew, V., Chacko, A. M., & Udhayakumar, A. (2018). Prediction of suitable human resource for replacement in skilled job positions using Supervised Machine Learning. In *8th International Symposium on Embedded Computing and System Design (ISED)* (pp. 37–41). IEEE.
<https://doi.org/10.1109/ISED.2018.8704120>
- Mousavian, S. A., Haeri, A., & Moslehi, F. (2021). Providing a hybrid clustering method as an auxiliary system in automatic labeling to divide employee into different levels of productivity and their retention. *Iranian Journal of Management Studies*, 14(1), 115–132.
<https://doi.org/10.22059/IJMS.2021.313448.674062>
- Muhammad, A., Asad, H., Jawad, A., Babar, S., & Gohar, A. (2020). Improving M-learners' performance through deep learning techniques by leveraging features weights. *IEEE Access*, 8, 131088–131106.
<https://doi.org/10.1109/ACCESS.2020.3007727>
- Nedelcu, A., Nedelcu, B., Sgarciu, A. I., & Sgarciu, V. (2020). Data mining techniques for employee evaluation. In *12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1–6). IEEE.
<https://doi.org/10.1109/ECAI50035.2020.9223165>
- Noori, R. (2023). *The state of promotions at work: How companies can fuel employee growth in 2023*. <https://nectarhr.com/blog/workplace-promotion-statistics>
- Oraman, Y., Selen, U., & Unakitan, G. (2011). Measuring employee expectations in a strategic human resource management research: Job satisfaction. *Procedia - Social and Behavioral Sciences*, 24, 413–420.
<https://doi.org/10.1016/j.sbspro.2011.09.022>
- Ouyang, J., & Ge, H. (2020). Data analysis framework of human resource estimation system based on MySQL-SAAS and fuzzy clustering. In *International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 417–420). IEEE. <https://doi.org/10.1109/ICOSEC49089.2020.9215371>

- Pahmi, S., Saepudin, S., Maesarah, N., Solehudin, U. I., & Wulandari. (2018). Implementation of CART (classification and regression trees) algorithm for determining factors affecting employee performance. In *International Conference on Computing, Engineering, and Design (ICCED)* (pp. 57–62). IEEE. <https://doi.org/10.1109/ICCED.2018.00017>
- Pal, K., & Sharma, M. (2020). Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. In *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)* (pp. 131–135). IEEE. <https://doi.org/10.1109/I-SMAC49090.2020.9243502>
- Peng, F., Guo, M., Zheng, C., Wang, S., Wang, X., & Xu, M. (2023). An assessment model of digital literacy for the students in vocational education based on principal component analysis in machine learning. *IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 6, 1382–1386. <https://doi.org/10.1109/ITNEC57892.2023.10134009>
- Petkov, P. N., Helgason, H., & Kleijn, W. (2012). Feature set augmentation for enhancing the performance of a non-intrusive quality predictor. In *Fourth International Workshop on Quality of Multimedia Experience* (pp. 121–126). IEEE. <https://doi.org/10.1109/QoMEX.2012.6263880>
- Phelan, S., & Lin, Z. (2000). Promotion systems and organizational performance: A contingency model. *Computational & Mathematical Organization Theory*, 7(3), 207–232. <https://doi.org/10.1023/A:1009692624413>
- Qi, E., & Sun, W. (2011). Research on the performance evaluation system of private enterprise based on AHP and PCA. In *IEEE 18th International Conference on Industrial Engineering and Engineering Management, Part 1* (pp. 201–204). IEEE. <https://doi.org/10.1109/ICIEEM.2011.6035219>
- Qian, L. (2013). Human resources assessment based on standard triangular whitenization weight function. In *IEEE International Conference on Grey Systems and Intelligent Services (GSIS)* (pp. 288–291). IEEE. <https://doi.org/10.1109/GSIS.2013.6714801>

- Ramdhani, T. W., Purwandari, B., & Ruldeviyani, Y. (2016). The use of data mining classification technique to fill in structural positions in bogor local government. In *Conference on Advanced Computer Science and Information Systems (ICACSYS)* (pp. 536–541). IEEE.
<https://doi.org/10.1109/ICACSYS.2016.7872797>
- Sahinbas, K. (2022). Employee promotion prediction by using machine learning algorithms for imbalanced dataset. In *2nd International Conference on Computing and Machine Intelligence (ICMI)* (pp. 1–5). IEEE.
<https://doi.org/10.1109/ICMI55725.2022.00005>
- Sarker, A., Shamim, S. M., Zama, M. S., & Rahman, M. A. (2018). Employee's performance analysis and prediction using K-means clustering & decision tree algorithm. *Global Journal of Computer Science and Technology: C Software & Data Engineering*, 18(1), 1–7.
<https://computerresearch.org/index.php/computer/article/view/1682>
- Silva, I. E., & Krohling, R. A. (2018). A fuzzy sociometric approach to human resource allocation. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–8). IEEE.
<https://doi.org/10.1109/FUZZ-IEEE.2018.8491685>
- Stephanie, C., & Sarno, R. (2019). Classification talent of employee using C4.5, KNN, SVM. In *International Conference on Information and Communications Technology (ICOIACT)* (pp. 388–393). IEEE.
<https://doi.org/10.1109/ICOIACT46704.2019.8938508>
- Strand, R., Spaapen, J., Bauer, M. W., Hogan, E., & Pereira, Â. G. (2015). *Indicators for promoting and monitoring responsible research and innovation*.
<https://op.europa.eu/en/publication-detail/-/publication/306a7ab4-f3cb-46cb-b675-9697caf5df19/language-en>
- Sulaiman, A. M. (2019). *Data science staff promotion prediction*.
<https://www.kaggle.com/datasets/behordeun/data-science-staff-promotion-prediction>

- Sun, H., & Li, Q. (2019). Research on application of PCA and K-means clustering in enterprise human resources. In *24th International Conference on Industrial Engineering and Engineering Management 2018* (pp. 429–436). IEEE.
https://doi.org/10.1007/978-981-13-6460-4_54
- Sun, Q., Wu, T., & Hua, J. (2022). Design of distributed human resource management system of Spark framework based on fuzzy clustering. *Journal of Sensors*, 2022(1), 1–9. <https://doi.org/10.1155/2022/4592087>
- Sun, W., & Zhao, X. (2011). Evaluation of talent cultivation for colleges and universities based on principal component analysis and BP neural network. In *International Conference of Information Technology, Computer Engineering and Management Sciences* (pp. 185–187). IEEE.
<https://doi.org/10.1109/ICM.2011.344>
- Syafrudin, M., Alfian, G., Fitriyani, N. L., Sidiq, A. H., Tjahjanto, T., & Rhee, J. (2020). Improving efficiency of self-care classification using PCA and decision tree algorithm. In *International Conference on Decision Aid Sciences and Application (DASA)* (pp. 224–227). IEEE.
<https://doi.org/10.1109/DASA51403.2020.9317243>
- Tallo, T. E., & Musdholifah, A. (2018). The implementation of genetic algorithm in SMOTE (Synthetic Minority Oversampling Technique) for handling imbalanced dataset problem. In *4th International Conference on Science and Technology (ICST)* (pp. 1–4). IEEE.
<https://doi.org/10.1109/ICSTC.2018.8528625>
- Tang, A., Lu, T., Lynch, Z. S., Schaer, O., & Adams, S. (2020). Enhancing promotion decisions using classification and network-based methods. *Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1–6). IEEE.
<https://doi.org/10.1109/SIEDS49339.2020.9106647>
- Tarusov, T., & Mitrofanova, O. (2019). Risk assessment in human resource management using predictive staff turnover analysis. In *1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)* (pp. 194–198). IEEE.
<https://doi.org/10.1109/SUMMA48161.2019.8947527>

- Thakur, D., Guzzo, A., & Fortino, G. (2021). t-SNE and PCA in ensemble learning based human activity recognition with smartwatch. In *IEEE 2nd International Conference on Human-Machine Systems (ICHMS)* (pp. 1–6). IEEE.
<https://doi.org/10.1109/ICHMS53169.2021.9582726>
- Todeschini, B. V., Rodrigues, C. M., Anzanello, M. J., & Tortorella, G. L. (2016). Clustering tool usage to align a company strategy to its talent management needs. *Journal of Industrial Engineering and Management*, 9(1), 1–20.
- Walter, W. (2021). *Human resource dataset*. <https://www.kaggle.com/colara/human-resource>
- Wang, Q., Li, B., & Hu, J. (2009). Feature selection for Human resource selection based on Affinity Propagation and SVM sensitivity analysis. In *World Congress on Nature & Biologically Inspired Computing (NaBIC)* (pp. 31–36). IEEE. <https://doi.org/10.1109/NABIC.2009.5393596>
- Wang, X., Yang, Y., Chen, M., Wang, Q., Qin, Q., Jiang, H., . . . Wang, H. (2020). AGNES-SMOTE: An oversampling algorithm based on hierarchical clustering and improved SMOTE. *Scientific Programming*, 2020(1).
<https://doi.org/10.1155/2020/8837357>
- Wang, Z. (2021). Research on digital economy and human resources based on fuzzy clustering and edge computing. *Security and Communication Networks*, 2021(1), 1–8. <https://doi.org/10.1155/2021/5583967>
- Wu, Q., & Shen, Y. (2023). Human resource attendance mechanism based on the Internet of Things: A method based on data fusion. *Internet Technology Letters*, 6(6). <https://doi.org/10.1002/itl2.391>
- Yu, M., Zhang, S., Zhao, L., & Kuang, G. (2017). Deep supervised t-SNE for SAR target recognition. In *2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST)* (pp. 265–269). IEEE.
<https://doi.org/10.1109/ICFST.2017.8304476>
- Zhao, Y. (2020). Application of K-means clustering algorithm in human resource data informatization. In *The 2020 International Conference on Cyberspace Innovation of Advanced Technologies* (pp. 45–49). IEEE.
<https://doi.org/10.1145/3431296.3431330>

CURRICULUM VITAE

NAME Theeramet Kaewwiset

EDUCATIONAL BACKGROUND

2018 Master of Computer Engineering
Mae Fah Luang University

2012 Bachelor of Software Engineering
Mae Fah Luang University

WORK EXPERIENCE

2019-Present System Analyst
Summit Computer

