CONDUCTING AN ARTIFICIAL INTELLIGENCE SUPPORTED STUDY ON HUMAN RESOURCES COMPETENCY ASSESSMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ANKARA UNIVERSITY

by

Nağme CÖMERTLER

IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN
ARTIFICIAL INTELLIGENCE TECHNOLOGY

ANKARA
2026

ABSTRACT

Master's Thesis

CONDUCTING AN ARTIFICIAL INTELLIGENCE SUPPORTED STUDY ON HUMAN RESOURCES COMPETENCY ASSESSMENT

Nağme CÖMERTLER

Ankara University
Graduate School of Natural and Applied Sciences Department of Artificial Intelligence Technologies

Supervisor: Prof.Dr. Asım Egemen YILMAZ

This study comprehensively examines the innovative approaches and effects of AI-supported qualification assessment systems implemented in large-scale organizations. Traditional qualification assessments are often conducted by managers or a limited group of peers, which restricts the evaluation process and fails to capture the full potential and individual competencies of all employees within an organization. The proposed AI-based system aims to derive results by analyzing a comprehensive dataset that includes individuals related to employees in technical and, if applicable, social club members, with managers also participating in the hierarchical evaluation. This method is especially crucial in large-scale organizations, where the high number of employees and the limited capacity of Human Resources departments to perform individualized analyses pose significant challenges. With AI support, employee performance, competency levels, and career development needs can be identified more rapidly and accurately, contributing critically to the achievement of organizational strategic goals. Additionally, compared to conventional methods, this system offers advantages in data analysis and interpretation processes, thereby enhancing overall efficiency and reducing costs. The findings of this study indicate that AI-supported qualification assessment models serve as a more inclusive, accurate, and effective management tool for organizations. Furthermore, the study's results are expected to pioneer innovative approaches in internal performance management and contribute to the development of new strategies that can be integrated into organizational transformation processes, thereby informing future research.

April 2026, 119 pages

Key Words: AI-supported Qualification Assessment, Competency Analysis, Data-driven Decision Making, Scalable Evaluation Models, Big Data Analysis, Natural Language Processing

ÖZET

Yüksek Lisans Tezi

İNSAN KAYNAĞI YETKİNLİK DEĞERLENDİRMESİNE YÖNELİK YAPAY ZEKA DESTEKLİ BİR ÇALIŞMA YÜRÜTÜLMESİ

Nağme CÖMERTLER

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Yapay Zeka Teknolojileri Anabilim Dalı

Danışman: Prof. Dr. Asım Egemen YILMAZ

Bu çalışma, büyük ölçekli organizasyonlarda uygulanan yapay zekâ destekli yetkinlik değerlendirme sistemlerinin yenilikçi yaklaşımlarını ve etkilerini kapsamlı bir biçimde incelemektedir. Geleneksel yetkinlik değerlendirmeleri genellikle yöneticiler veya sınırlı sayıdaki çalışma arkadaşı tarafından yapılmakta olup, bu durum değerlendirme sürecini kısıtlamakta ve organizasyon içindeki tüm çalışanların potansiyelini ve bireysel yetkinliklerini tam olarak yansıtamamaktadır. Önerilen yapay zekâ tabanlı sistem, çalışanın teknik olarak ilişkili olduğu kişiler ile (varsa) sosyal kulüp üyelerini ve hiyerarşik olarak yöneticileri de dahil eden kapsamlı bir veri setini analiz ederek sonuç üretmeyi amaçlamaktadır. Bu yaklaşım özellikle çalışan sayısının fazla olduğu büyük ölçekli kurumlarda, insan kaynakları birimlerinin bireysel analiz yapma kapasitesinin sınırlı olması nedeniyle kritik önem taşımaktadır. Yapay zekâ desteği sayesinde çalışan performansı, yetkinlik düzeyleri ve kariyer gelişim ihtiyaçları daha hızlı ve doğru bir biçimde belirlenebilmekte; bu da kurumların stratejik hedeflerine ulaşmasına önemli katkılar sağlamaktadır. Ayrıca geleneksel yöntemlerle karşılaştırıldığında, bu sistem veri analizi ve yorumlama süreçlerinde avantaj sunarak genel verimliliği artırmakta ve maliyetleri azaltmaktadır. Araştırma bulguları, yapay zekâ destekli yetkinlik değerlendirme modellerinin kurumlar için daha kapsayıcı, doğru ve etkili bir yönetim aracı olduğunu göstermektedir. Bununla birlikte, çalışmanın sonuçlarının iç performans yönetimi alanında yenilikçi yaklaşımlara öncülük etmesi ve kurumsal dönüşüm süreçlerine entegre edilebilecek yeni stratejilerin geliştirilmesine katkı sağlaması beklenmektedir.

Nisan 2026, 119 sayfa

Anahtar Kelimeler : Yapay Zekâ Destekli Yetkinlik Değerlendirmesi, Yetkinlik Analizi, Veri Odaklı Karar Verme, Ölçeklenebilir Değerlendirme Modelleri, Büyük Veri Analizi, Doğal Dil İşleme

FOREWORD AND ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to everyone who contributed to the preparation and execution of this thesis. I extend my special thanks to my advisor, Prof. Dr. Asım Egemen Yılmaz, whose guidance, suggestions, and unwavering support at every stage of my research have been instrumental in my academic and personal development. The education I received at the Artificial Intelligence Technologies Department of Ankara University has broadened my scientific vision and laid the foundation for this work. Moreover, the opportunities and institutional support provided by Havelsan have significantly enhanced the practical application dimension of my research.

I am also grateful to the esteemed faculty, colleagues, and technical staff who generously shared their knowledge and experience during my academic journey. Their continuous support and motivation played an essential role in the successful realization of this thesis. Finally, I offer my deepest appreciation to my spouse and my mother for their unwavering financial and emotional support, as well as for their love, patience, and understanding throughout my journey.

Nağme CÖMERTLER
Ankara, April 2026

TABLE OF CONTENTS

SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
ANOVA	Analysis of Variance
Avg	Average
BARS	Behaviourally Anchored Rating Scales
BERT	Bidirectional Encoder Representations from Transformers
CopulaGAN	Copula-based Generative Adversarial Network
CTGAN	Conditional Tabular GAN
Δ (Delta)	Change or difference between two values
DI	Disparate Impact
DL	Deep Learning
DP	Differential Privacy
DVC	Data Version Control
EURES	European Employment Service
ESCO	European Skills, Competences, Qualifications and Occupations
F1-Score	Harmonic mean of precision and recall
GAN	Generative Adversarial Network
GE	General Electric
GDPR	General Data Protection Regulation
HiPo	High-Potential Employee
HR	Human Resources
HR DSS	Human Resources Decision Support System
HRM	Human Resource Management
HTML	HyperText Markup Language
KPI	Key Performance Indicator
IBM	International Business Machines Corporation
ID	Identification
IT	Information Technology
JD–CV	Job Description - Curriculum Vitae
LightGBM	Light Gradient Boosting Machine
LLM	Large Language Model
LR	Logistic Regression
MA	Moving Average
MA2	Two-Year Moving Average
Macro-F1	F1-score computed by averaging across classes
MAE	Mean Absolute Error
MBO	Management by Objectives
MedGAN	Medical Generative Adversarial Network
Micro-F1	F1-score computed globally across all instances
MIA	Membership Inference Attacks

MICE	Multiple Imputation by Chained Equations
ML	Machine Learning
NLP	Natural Language Processing
Norm	Normalized value
PCA	Principal Component Analysis
PhD	Doctor of Philosophy
PSI	Population Stability Index
QA	Quality Assurance
R&D	Research and Development
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
SHAP	SHapley Additive exPlanations
TF-IDF	Term Frequency–Inverse Document Frequency
TL	Turkish Lira
VAE	Variational Autoencoder
WLB	Work–Life Balance
XGBOOST	Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Human resources (HR) management has undergone a profound transformation over the last decade, driven by digitalization, datafication, and the increasing availability of employee-generated textual information. Organizations today collect vast amounts of structured and unstructured HR data including performance records, employee feedback forms, peer evaluations, engagement surveys, and internal communication channels. While these data streams offer great potential to enhance decision-making, traditional appraisal systems remain limited in their capacity to extract meaningful insights from large-scale qualitative information. This limitation creates challenges in ensuring fairness, consistency, and evidence-based evaluation across the workforce.

Artificial intelligence (AI), particularly Natural Language Processing (NLP) and Machine Learning (ML), has emerged as a powerful enabler for modern HR analytics. NLP models can interpret the semantic and emotional content of employee feedback, while ML models can detect patterns relating to employee performance, potential, satisfaction, burnout risk, and promotion readiness. These technologies have reshaped appraisal mechanisms by moving from subjective, manager-centric evaluations toward more data-grounded and holistic decision frameworks.

However, despite significant research progress, several challenges persist. Many HR datasets are incomplete, imbalanced, or limited to numerical indicators, reducing their ability to capture behavioural and competency-related dimensions. Additionally, organizations rarely possess multi-year, longitudinal datasets needed for predictive modelling, fairness analysis, or trend detection. As a result, empirical HR research frequently suffers from data scarcity, limited reproducibility, and insufficient benchmarking.

To address these gaps, this study proposes an AI-supported study on human resources competency assessment, designed to integrate NLP-based textual analysis with advanced machine learning models across both single-year and multi-year HR datasets. The study is conducted in two main stages:

Stage 1 NLP-Based Competency and Sentiment Extraction (Single-Year Framework): Employee feedback texts from the 2025 cycle are analyzed using two core NLP tasks:

- *Competency Classification*: Identifying key HR competencies such as communication, teamwork, leadership, innovation, problem solving, and ownership.
- *Sentiment Analysis*: Assigning positive, neutral, or negative sentiment labels, with contextual sensitivity and calibration improvements.

  These outputs generate interpretable text-derived indicators (Sentiment_Index, Competency_Score vectors), enriching the HR evaluation pipeline.

Stage 2 Multi-Year HR Decision Engine (2021–2025 Panel Data): A synthetic yet realistic five-year HR panel dataset is constructed, incorporating structured indicators (performance, experience, training, promotion outcomes, salary changes) together with the NLP-derived behavioural measures. Multiple ML models (Random Forest, LightGBM, XGBoost, Logistic Regression) are used to answer 14 critical HR questions including promotion prediction, salary raise determinants, burnout risk estimation, fairness assessment across departments, and year-over-year sentiment/competency trends.

By combining text analytics and multi-year predictive modelling, this study demonstrates a comprehensive AI-supported HR assessment approach capable of enhancing decision quality, improving transparency, and reducing subjective bias in traditional appraisal systems. The methodology is particularly relevant for organizations that lack longitudinal data or advanced analytical infrastructure, offering a reproducible and scalable framework for competency-based HR analytics.

The remainder of the thesis is structured as follows:

- **Section 2** presents the theoretical background, including performance management evolution, NLP applications in HR, synthetic data generation methods, and risks of AI-enabled appraisal systems.
- **Section 3** describes the data pipeline, preprocessing steps, NLP modelling workflow, and the construction of the multi-year HR decision engine.
- **Section 4** reports empirical findings from competency modelling, sentiment analysis, and predictive modelling across the 14 HR research questions.
- **Section 5** concludes the study and outlines future research directions.

This integrated framework aims to contribute to both academic literature and practical HR analytics by offering a transparent, explainable, and data-driven approach to competency assessment and decision support.

2. LITERATURE REVIEW

2.1 History of Performance Management System

In contemporary organizations, employee performance evaluation has evolved from traditional hierarchical assessments to multidimensional and technology-supported systems. Historically, most companies have relied on managerial (top-down) reviews, in which supervisors rate subordinates based on predefined criteria. While this approach offers administrative simplicity and clear accountability, it often suffers from subjectivity and a limited perspective. Over time, organizations began adopting more participatory models, such as self-appraisals, which encouraged employees to reflect on their strengths, weaknesses, and career goals, fostering self-awareness and engagement. However, the inherent bias in self-ratings prompted the integration of multi-rater systems to enhance objectivity.

Among these, the 360-degree feedback approach has become one of the most widely adopted methods, especially in large corporations such as GE(General Electric), IBM, and Novo Nordisk(Stahl et al., 2011). As seen in Figure 2.1., this model combines feedback from managers, peers, subordinates, and sometimes clients, to provide a holistic view of an employee's performance. Research indicates that while 360-degree systems can uncover hidden strengths and development areas, their success depends heavily on design quality, anonymity, and follow-up mechanisms (Bracken et al., 2016). Simultaneously, Behaviourally Anchored Rating Scales (BARS) and Management by Objectives (MBO) frameworks were developed to align evaluations more closely with observable behaviours and measurable outcomes. BARS enhances inter-rater consistency by attaching behavioural examples to numerical scales, whereas MBO links performance to strategic objectives and quantifiable targets (Jafari et al., 2009).

Figure 2.1 Degree feedback schema

Recently, organizations have moved toward continuous and real-time feedback systems enabled by digital platforms that allow ongoing performance tracking rather than annual reviews. Companies such as Adobe and IBM have replaced rigid rating systems with iterative "check-ins" and dashboard-based tools that visualize progress and foster ongoing dialogue between employees and managers to enhance employee performance. These systems often integrate data analytics to detect performance trends, productivity patterns, and collaboration dynamics, making performance management more dynamic and data driven.

In parallel with these developments, AI-powered performance evaluation has emerged as a transformative innovation in modern human resource management (HRM). Machine learning algorithms can analyze large datasets, including quantitative KPIs and qualitative feedback, to identify performance patterns, competency gaps, and developmental needs with greater accuracy (Qin et al., 2023). NLP and sentiment analysis techniques are now being applied to interpret open-ended comments in 360-degree surveys, enabling richer insights into behavioural competencies such as teamwork, leadership, and communication (Thirunagalingam et al. 2025). Furthermore, Explainable AI (XAI) models, such as SHAP and LIME, are being used to ensure that algorithmic decisions remain transparent and interpretable to HR practitioners (Dwivedi et al., 2023).

Overall, the trajectory of performance appraisal systems reflects a shift from subjective and episodic assessments to continuous, data-driven, and AI-supported evaluation ecosystems. Organizations are increasingly combining traditional behavioural frameworks with digital and analytical tools to achieve greater scalability, fairness, and developmental impact. Nevertheless, scholars emphasize that the most effective systems maintain a human-centered balance, using AI as an augmentation tool rather than a replacement for managerial judgment (Pan et al.,2025).

2.2 Technical Integration of Machine Learning and NLP in HR Analytics

Recent literature highlights the transformative role of AI and ML in reshaping HRM processes. A growing number of studies have demonstrated that AI-driven ML models not only enhance analytical capabilities but also contribute to more data-informed, proactive, and personalized decision-making in HR practices.

One key line of research focuses on the application of ensemble machine learning algorithms for predicting employee turnover. These studies emphasize that a variety of data sources such as employees' demographic profiles, job-related information, and performance metrics should be integrated to improve turnover prediction accuracy. The findings reveal that ensemble-based models significantly outperform traditional statistical approaches, establishing AI-driven ML as a powerful paradigm for detecting early warning signs and enabling proactive HR interventions (Pessach et al., 2020).

In the context of individual performance evaluation, recent studies propose combining NLP with Deep Learning (DL) techniques to analyze unstructured textual data from employee feedback and performance reviews. Such advanced analytics methods are capable of capturing the subtle, subjective nuances of employee behaviour, offering a more holistic understanding of performance drivers. Furthermore, this information can inform the formulation of personalized talent management policies, ultimately enhancing employee engagement and overall organizational effectiveness. Literature also emphasizes the importance of embedding behavioural science insights into AI-driven HR frameworks to ensure contextual relevance and human-centric design, as AI-ML

models can support talent identification, workforce planning, and personalized career development in a data-driven manner, while also requiring alignment with human-centric principles and context-sensitive organizational frameworks (Chowdhury et al., 2023; Budhwar et al., 2022).

Another important contribution in the literature concerns the comparison of various ML algorithms used for turnover prediction. These studies consistently identify data quality, feature engineering, and model selection as the three most critical determinants of prediction accuracy. By combining structured and unstructured multimodal data, researchers have achieved more robust and reliable performance forecasting results. These studies also stress the necessity of a multidisciplinary collaboration between HR professionals, data scientists, and behavioural experts to enhance both the technical validity and the practical relevance of AI applications in HRM. (Jin et al., 2020).

There are studies showing that leadership development approaches have moved away from traditional classroom-based and standardized training models toward more personalized, digitalized, socially driven, and individually responsible structures. There are also approaches suggesting that the contemporary understanding of leadership development has been reshaped through more flexible, individual-oriented, technology-supported, and relationship/network-based learning structures compared with classical executive education models (Moldoveanu et al., 2019).

Employee attrition is defined as the situation in which employees leave their current organization and move to another organization due to better career or compensation opportunities. It is stated that attrition rates tend to increase particularly in cases such as mass retirements, sectoral growth, and rising labor demand. In this context, it is argued that machine learning methods can be used to predict employee turnover, and that the logistic regression model applied on the IBM HR dataset produced successful results (Ponnuru et al., 2020).

The literature continues to evolve by presenting competency-based frameworks for identifying high-potential employees. Such approaches argue that talent management

strategies must align with the organization's core competencies and long-term strategic direction. The proposed models integrate performance metrics, behavioural indicators, and developmental progress data to accurately predict individual leadership and development potential. This kind of personalized approach allows organizations to make targeted investments in employee growth, leading to a more skilled and engaged workforce (Juhdi et al., 2012).

Recent advancements also explore ML-based feedback systems designed to improve employee engagement. Studies demonstrate that integrating NLP and sentiment analysis techniques into HR systems enables real-time monitoring and personalization of employee feedback. Personalized, timely responses not only foster transparency and collaboration but also enhance engagement, productivity, and job satisfaction levels across the organization (Yanamala, 2022).

In addition, literature addressing AI-driven career development interventions highlights frameworks for leveraging ML algorithms to identify personal aspirations, skill gaps, and professional growth opportunities (Ekuma, 2024, Chuang et al., 2024). Importantly, researchers underline the need to customize career development plans by incorporating behavioural and psychological factors such as personality traits and motivational patterns to create more meaningful and sustainable career pathways.

Collectively, these studies underscore that the integration of AI and ML in HRM goes beyond automation it enables adaptive, evidence-based, and human-centric decision-making. As AI systems mature, their capacity to analyze behavioural data, predict performance, and personalize employee experiences positions them as indispensable tools in the next generation of human capital management.

The advancement of NLP techniques has significantly improved job matching systems, enabling more accurate and efficient alignment between job seekers and relevant positions. One of the traditional approaches, Term Frequency–Inverse Document Frequency (TF-IDF), assigns weights to words based on their frequency in a document and their rarity across the corpus. While TF-IDF is effective for identifying key terms, it

fails to capture semantic relationships between words, which limits its ability to process contextual information in job-related texts such as resumes and job postings (Bafna et al., 2016).

This limitation has been addressed through embedding-based methods, which represent words and sentences as dense vectors that encode semantic meaning. These embeddings allow the computation of similarity between resumes and job postings at the semantic level, rather than relying solely on lexical overlap. Among such models, BERT (Bidirectional Encoder Representations from Transformers) has demonstrated strong performance in learning contextual word representations that capture both syntactic and semantic structure.

However, since BERT is not inherently optimized for direct semantic similarity measurement between sentences or documents, Sentence-BERT (SBERT) and other Siamese-based architectures have been developed. These models employ a dual-encoder structure to generate embeddings that enable efficient semantic comparison, proving effective for resume–job matching tasks (Reimers & Gurevych, 2019).

Building upon these advancements, CareerBERT (Rosenberger et al., 2025) introduces a transformer-based framework that represents resumes and job descriptions in a shared embedding space. Trained on positive and negative JD–CV (Job Description - Curriculum Vitae) pairs, the model learns to position semantically similar documents closer together while distancing unrelated pairs. A key innovation of CareerBERT is its taxonomy-driven design: instead of matching resumes to specific job postings, it aligns them with general job categories derived from the ESCO taxonomy, thereby reducing potential human biases present in historical data.

CareerBERT was trained using large-scale data from EURES and ESCO, achieving over 20 % improvement compared to classical TF-IDF and Doc2Vec approaches. Furthermore, its outputs demonstrated a strong correlation with expert human judgments ($r \approx 0.83$). Consequently, CareerBERT provides a robust foundation for competency-based placement, internal mobility, and AI-driven qualification assessment systems.

2.3 Preparing for AI-Enabled Appraisal Process

The effective implementation of AI-driven performance evaluation systems requires organizations to be comprehensively prepared, both technologically and organizationally. The first step in this transformation is establishing a robust and integrated information technology infrastructure capable of operating seamlessly with AI tools and handling increased data processing demands. The reliability and validity of AI-generated insights depend heavily on data quality; therefore, organizations must focus on cleaning, updating, and structuring their existing databases to ensure accurate and consistent results.

Selecting scalable and reliable AI tools that align with organizational goals is a critical factor for success. Moreover, leadership support plays a vital role in driving this transformation; senior management must actively endorse the adoption of AI, allocate the necessary resources, and encourage employee adaptation. Employee engagement is equally important; informing staff about upcoming changes, addressing potential concerns, and clearly communicating the benefits of AI-supported appraisal systems help strengthen the organizational commitment to the process.

Continuous training and development programs should be implemented to enhance HR professionals and managers' data literacy, analytical reasoning, and technological proficiency, enabling them to accurately interpret AI-generated insights and act upon them effectively. Alongside technological transformation, performance criteria should be redefined to ensure that metrics are measurable by AI systems and truly reflect the value that employees contribute to the organization. Real-time data analytics further enable continuous feedback mechanisms that foster employee development and performance improvement. From an ethical standpoint, clear guidelines must be established to ensure transparency in AI decision-making and protect against algorithmic bias (Thirunagalingam et al., 2025).

Finally, organizations should remain vigilant toward emerging AI advancements, maintain strategic agility, and regularly evaluate system effectiveness to ensure that AI-

driven performance management remains sustainable, fair, and aligned with their long-term objectives.

2.4 Risks and Drawbacks of AI-Enabled Performance Appraisal Systems

The risks associated with AI-enabled performance appraisal processes can be examined under several distinct categories. These categories include,

Algorithmic Bias and Cross-Group Unfairness: AI-based performance appraisal systems can replicate historical biases present in training data, thereby systematically disadvantaging certain demographic groups. AI-Driven performance appraisal systems highlights the inherent risk of algorithmic bias, emphasizing that such systems may create disparities based on gender, race, or job position. Similarly, the article 3 key risks when using AI for performance management underscores that AI tools often reinforce pre-existing biases embedded within the datasets used for training.

Another important dimension of bias is cross-group unfairness. In AI-driven appraisal systems, this refers to situations where a model exhibits high overall accuracy but consistently produces errors or lower scores for specific subgroups such as male versus female, younger versus older or across different departments and cultural backgrounds (Mehrabi et al., 2021). In other words, even if the system performs well on average, it may remain unfair at the micro level.

Example: Consider a performance evaluation model designed to predict "leadership potential" scores based on historical company data. If most previous leaders in the dataset were male, the model may implicitly learn an association between leadership and male behavioural patterns. Consequently, despite maintaining 90% overall accuracy, the system could systematically assign lower leadership scores to female employees. This phenomenon is known as disparate impact or group fairness violation in fairness-aware machine learning literature.

Opacity and Lack of Explainability: AI models can behave like a "black box" when their decision-making logic lacks transparency. This may lead employees to question their evaluation results without receiving satisfactory explanations for why they received a specific score. Numerous studies have been conducted to promote transparent AI use in HR processes. One such study proposes the adoption of XAI models in HR systems to enhance transparency. XAI has emerged as a vital research domain aimed at improving transparency, fairness, and accountability in AI-driven employment decisions. However, many existing XAI techniques produce outputs that, while mathematically grounded, remain difficult for non-expert users to interpret in practice (Salih et al., 2025).

On the other hand, the high level of transparency advocated by XAI is not always correctly understood by users. Another study emphasizes that the benefits of XAI in recruitment depend significantly on users' level of AI literacy. It highlights the necessity of personalized explanation strategies and targeted literacy training within HR departments to ensure the fair, transparent, and effective adoption of AI (Kalff& Simbeck, 2025).

Privacy, Data Protection and Legal Regulations: Employee-related data are highly sensitive, and the processing, storage, and analysis of such information during AI model development introduce significant privacy risks. The General Data Protection Regulation (GDPR) published by the European Parliament examines the legal bases for applying AI technologies to personal data, particularly emphasizing the obligations related to profiling and automated decision-making systems. The regulation elaborates on data subjects' rights, including access, erasure, portability, and objection. The study provides a comprehensive analysis of automated decision-making processes, addressing the extent to which such decisions are acceptable, the protective measures that should be implemented, and whether data subjects are entitled to receive individual explanations regarding AI-driven outcomes (European Parliament & Council of the European Union, 2016).

Employee Perception, Dignity and Trust Deficiency: Research has indicated that the use of AI models in human resource processes may be perceived by some employees as

degrading or disrespectful. However, the opposite can also occur. If employees believe that a manager might act with bias, AI systems may be viewed as more trustworthy; conversely, when employees expect a positive bias from the manager, human evaluation may be preferred (Cha, 2025).

Contextual Fit, Anchoring Bias and Interaction Effects: AI recommendations can influence human evaluators, leading to an "anchoring effect," whereby the evaluator adopts the AI's suggested score as a reference point and adjusts their own assessment around it. The study by Carter and Liu demonstrates that when a low-level AI recommendation is presented, managers tend to assign similar scores without deviating significantly from the AI's suggestion (Carter & Liu, 2025).

Error Propagation, Automation Bias and Lack of Oversight: AI tools are known to occasionally produce outputs that are nonsensical or inaccurate commonly referred to as AI hallucinations. Organizations should avoid relying exclusively on AI-generated outputs and must independently verify their accuracy. For instance, incorrect statements within an employee's performance evaluation could expose the organization to considerable legal and reputational risk (Alon-Barkat et al., 2023).

Corporate Risk, Reputation and Lack of System Audit: Errors in AI systems, particularly within highly visible decision-making processes such as promotion or disciplinary actions, can severely damage an organization's reputation (Lukaszewski et al., 2024). Large enterprises therefore conduct audits, red teaming (ethical stress testing), and scenario analyses prior to AI system deployment. Media and public discourse frequently warn that AI-based evaluation systems may dehumanize employees, reduce morale, and foster a sense of alienation in the workplace (Sharma et al., 2025).

2.5 Synthetic Data Generation

Synthetic data is defined as artificially generated information that preserves the statistical characteristics and relationships of real-world datasets without directly exposing sensitive

records. It serves as a viable alternative when the collection, sharing, or utilization of real data is limited by privacy, legal, or accessibility constraints. In data-driven fields such as healthcare, finance, and human resources, data availability is frequently restricted due to confidentiality concerns and regulatory frameworks like the GDPR. Synthetic data mitigates these issues by enabling the creation of representative datasets that maintain utility while minimizing privacy risks. While high utility ensures that synthetic datasets support accurate downstream analyses, maintaining privacy prevents the re-identification of individuals from generated samples. Balancing these two objectives remains a key research concern. Synthetic data is therefore positioned as a central enabler of privacy-preserving machine learning and data democratization, where secure data sharing and collaborative research become possible without compromising sensitive information (Shi et al., 2025).

Compared to other data modalities such as image or text data tabular data presents unique structural complexities, including mixed feature types (categorical, numerical), correlations among attributes, and diverse distributions. These challenges necessitate specialized generative models capable of accurately learning and replicating the underlying data-generating process.

2.5.1 Taxonomy of generative models

The generation of synthetic tabular data can be categorized into three main methodological families: statistical models, deep generative models, and hybrid approaches. This taxonomy is based on how the model represents dependencies within the data, learns the underlying probability distributions, and synthesizes new samples that mimic real data characteristics (Shi et al., 2025).

Statistical and probabilistic models generate synthetic observations by explicitly modelling the probabilistic relationships among variables. Representative examples include Gaussian Copula, Bayesian Networks, and Decision Tree Sampling. These models offer a high degree of interpretability and transparency; however, their ability to

accurately capture complex, high-dimensional, and nonlinear relationships is limited, making them less effective for heterogeneous tabular datasets.

Deep generative models provide a more powerful framework for capturing complex data structures and nonlinear dependencies.

- GAN-based models (e.g., CTGAN, CopulaGAN, MedGAN) operate through an adversarial learning process between a generator and a discriminator, resulting in highly realistic synthetic samples. Despite their success, they often suffer from training instability and poor performance when learning from imbalanced or sparse data distributions.
- VAE-based models learn latent probabilistic representations of the data and generate new samples by sampling from these latent spaces. They generally exhibit stable training behaviour but may produce overly smoothed or less diverse samples compared to GANs. Autoregressive models generate features sequentially, conditioning each step on previously generated variables, allowing them to capture strong inter-variable dependencies.
- Diffusion-based models, by contrast, progressively denoise data from random noise, yielding samples with high fidelity and complex structural consistency.
- Unlike traditional statistical or deep learning-based models, Large Language Models (LLM-based models), it leverages the language understanding, pattern recognition, and contextual generation capabilities of LLMs. It can be implemented in two ways: Prompt-based and Fine-Tuning.

Hybrid approaches integrate statistical modelling with deep learning techniques, aiming to combine the interpretability of probabilistic methods with the expressive power of neural networks.

The relevant methods are detailed in Figure 2.2.

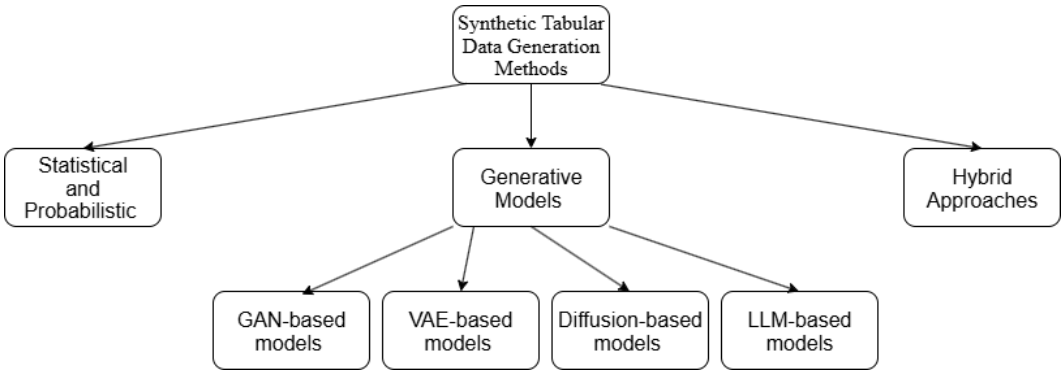


Figure 2.2 Overview of synthetic tabular data generation methods

Overall, this taxonomy highlights that there is no universally superior method for synthetic tabular data generation. The choice of model depends on multiple factors such as the type of data, desired privacy level, computational resources, and the intended downstream application.

2.5.2 Evaluation metrics

The evaluation of synthetic tabular data quality is conducted across three major dimensions: resemblance, utility, and privacy. Together, these dimensions form a comprehensive framework for assessing how closely synthetic data reproduces the statistical and functional properties of real data.

Resemblance metrics measure the statistical similarity between synthetic and real datasets. Commonly used indicators include distributional distance measures, correlation difference analysis, Principal Component Analysis (PCA) alignment, and statistical hypothesis tests such as the Kolmogorov–Smirnov test. These metrics evaluate the extent to which the synthetic data preserves the underlying data distributions, correlations, and feature dependencies.

Utility metrics quantify the practical usefulness of synthetic data in downstream analytical or machine learning tasks. The most common approach compares the performance of models trained on synthetic data to those trained on real data using metrics

such as Accuracy, F1-score, ROC-AUC, and Mean Absolute Error (MAE). High utility indicates that synthetic data can effectively substitute for real data in training, validation, or benchmarking contexts.

Privacy metrics assess the resistance of synthetic data to privacy attacks and the risk of re-identifying individuals from generated samples. The most frequently applied methods include Membership Inference Attacks (MIA), Attribute Inference Attacks, and Differential Privacy (DP)-based testing. These techniques evaluate whether synthetic data reveals sensitive information or individual participation in the original dataset.

A fundamental trade-off exists among these three dimensions: increasing resemblance and utility often raises privacy risks. Therefore, finding an optimal privacy–utility balance is one of the core challenges in synthetic data research. An effective synthetic data generation process aims to achieve high fidelity and representativeness while maintaining strong privacy guarantees (Shi et al., 2025).

3. MATERIALS AND METHODS

The evaluation of employee performance and the subsequent decisions regarding seniority, promotion, and role assignments constitute a critical aspect of human resources. In this study, it is aimed to automate these human-dependent processes and generate more accurate results by utilizing artificial intelligence models. Considering factors such as the limited number of evaluators and the abundance of data, conducting a precise and unbiased analysis becomes increasingly challenging. Therefore, the main objective of this research is to analyze the dataset by employing machine learning and language models, and to produce meaningful and reliable outcomes related to individual performance evaluations.

In order to present a clear and detailed explanation of the study, this section is structured into several subsections, each describing different stages of the research process. Collecting real data from corporate organizations has become nearly impossible due to data privacy and confidentiality constraints. In addition, considering the possibility of missing fields or manual interventions in existing datasets, synthetic data was utilized in this study. The process of generating synthetic data, the methods employed, and the content of the resulting dataset are described in detail in the following sections.

Following the acquisition of the data, the next stage involves model selection and training. This phase represents one of the most critical components of the study. The use of an appropriate model and the determination of optimal training parameters are of great importance to ensure the accuracy and reliability of the results.

Considering the potential applicability of the developed system by corporate organizations and the necessity of presenting the evaluation results in a more comprehensible visual format, various dashboard visualization methods were employed. These methods and their implementation details are explained in the corresponding subsection.

Ultimately, performance metrics were established to evaluate the outcomes of the study. The evaluation aimed not only to assess the model and the dataset but also to analyse the overall effectiveness and validity of the results.

3.1 Overview of the AI-Supported Competency Assessment Framework

The AI-Supported Competency Assessment Framework developed in this study was designed to establish an objective, sustainable, and scalable evaluation infrastructure for human resources processes. By integrating traditional performance appraisal practices with AI-driven data analytics, the framework enables a multidimensional examination of employees' competency levels. This integration facilitates a more precise, explainable, and measurable alignment between individual performance indicators and organizational strategic objectives.

All data used throughout the study were synthetically generated to simulate a large-scale corporate environment while avoiding privacy violations. In addition, a manually constructed gold-standard test set was used to provide an unbiased evaluation of model performance.

The proposed system was constructed in two main phases. In the first phase, a comprehensive NLP pipeline was developed using employee feedback data exclusively from the year 2025. This pipeline begins with data cleaning, normalization, and the transformation of raw textual content into computational representations suitable for machine learning models.

3.2 Data Generation and Processing Chain

The primary objective of this study is to support performance evaluation processes conducted within corporate environments through an objective, repeatable, and scalable artificial intelligence framework. Due to privacy constraints, incomplete or potentially

biased labelling, and the risk of manual intervention when accessing real corporate data, a synthetic data–driven approach was adopted in this study.

During the construction of the synthetic dataset, the following key principles were strictly observed:

- Privacy-preserving data generation strategies were employed to ensure that the generated records cannot be linked to real individuals and to minimize the risk of membership inference attacks.
- Although the dataset is synthetic, it was carefully designed to realistically reflect organizational roles, employees' professional experience profiles, academic backgrounds, and 360-degree evaluation scores that are commonly observed in real corporate environments. As a result, the generated dataset exhibits statistical and structural characteristics comparable to real-world corporate data.
- The synthetic data were constructed to enhance the effectiveness of downstream machine learning and natural language processing tasks, providing representative and informative samples for model training and evaluation.
- In line with the principles of reproducibility and scalability, the dataset was designed to allow future extensions such as increasing the number of records or introducing additional variables without compromising data consistency or structural integrity.

To enable a systematic examination of the generated synthetic data, a data dictionary was created and is presented in detail in the corresponding section. In the first phase of the study, employee feedback comments were analyzed using trained language models to perform sentiment classification, categorizing the feedback as positive, negative, or neutral. The same textual data were subsequently utilized to identify employees' competency labels**,** specifically Communication, Innovation, Leadership, Ownership, Problem Solving, and Teamwork, through language model–based classification.

3.2.1 Data dictionary

Table Data Dictionary presents the data dictionary describing all variables included in the AI based performance evaluation dataset. Each variable is listed with its corresponding variable name, data type, allowed values, and a short description. In this section, the rules used to generate each piece of information contained in the dataset, as well as the considerations taken into account to ensure that the synthetic data reflects a real dataset from a corporate organization, will be explained.

The synthetic dataset includes N=1000 employee records to reflect large scale corporate structures. For scalability testing, additional experiments with N=100 and N=500 subsets were conducted to evaluate model stability.

Table 3.1 Data dictionary

Variable	Variable Name	Data Type	Allowed Values	Description
Employee_ID	ID	Categorical	Unique integer	Unique identifier assigned to each employee.
Employee_Name	Name	Text	Free text	Full name of the employee. (Excluded from model input)
Gender	Gender	Categorical	{Male, Female}	Gender of the employee.
Age	Age	Numerical	Integer (e.g., 20-65)	Age of the employee in years.
Education_Level	Education Degree	Categorical	{Bachelor, Master, PhD}	Highest level of education completed.
University_Entry_Rank	University Rank	Numerical	Integer (e.g., 1000-100000)	Rank or score obtained upon university admission.
Language_Score	Language Score	Numerical	Float [0.0-100.0]	Language proficiency score, usually standardized on a 0-5 scale.
Life_Event	Life Event	Categorical	{Marriage, Divorce, Relocation,None}	Recent major personal life event.
Number_of_Child	Children Count	Numerical	Integer (0-10)	Total number of children the employee has.
Start_Date	Starting Date	Categorical / Datetime	Date format (YYYY-MM-DD)	Date the employee joined the company.
Department	Department	Categorical	{HR, IT, R&D, Finance, SAT, Engineering, Sales, etc.}	Organizational department where the employee currently works.
Role	Role	Categorical	{Software Engineer, Systems Engineer, Data Analyst, Manager, etc.}	Job role or title held by the employee.
Current_Position	Position	Categorical	Junior, Mid-level, Senior, Lead, etc	Current position level within the organization hierarchy.
Total_Work_Experience	Work Experience	Numerical	Float (0.0-40.0)	Total professional work experience in years, including prior jobs.
Years_At_Company	Company tenure	Numerical	Float (0.0-30.0)	Total number of years the employee has been with the company.
Years_With_Curr_Manager	Employee Manager tenure	Numerical	Float (0.0-30.0)	Duration in years working under the current manager.
Promotion_Status	Potential of Promotion	Categorical	{Yes, No}	Indicates whether the employee received a promotion during the period.
Career_Potential	Potential of Career	Categorical	{Low, Medium, High}	Assessment of the employee's potential for future career growth.
Succession_Ready	Succession Ready	Categorical	{Yes, No}	Whether the employee is ready for succession planning or leadership.
Goal_Attainment_Score	Score of goal attainment	Numerical	Float [0.0-5.0]	Degree to which the employee achieved set goals.
Strategic_Initiative_Impact	Impact of Strategic Initiative	Categorical	{Low, Medium, High}	Level of impact the employee had on strategic initiatives.
Deliverables_Quality_Score	Score of Deliverables Quality	Numerical	Float [0.0-5.0]	Average quality rating of the employee's deliverables.
Deadline_Adherence_Rate	Rate of Deadline Adherence	Numerical	Float [0.0-100.0]	Percentage of tasks completed within the assigned deadlines.
Problem_Resolution_Count	Count of Problem Resolution	Numerical	Integer (0-100)	Number of issues or problems successfully resolved.
Task_Efficiency_Score	Task Efficiency Score	Numerical	Float [0.0-5.0]	Efficiency in completing assigned tasks.
Ownership_Level	Ownership Level	Numerical	Float [0.0-5.0]	Degree of initiative and responsibility taken by the employee.
Documentation_Contribution	Contribution of Document	Numerical	Integer (0-100)	Number of documentation tasks or knowledge base contributions.
Code_Commit_Count	Count of Code Commit	Numerical	Integer	Total number of code commits made by the employee.
Annual_Training_Count	Number of training	Numerical	Integer (0-20)	Number of training programs attended within a year.
Annual_Certificates_Count	Number of certificate	Numerical	Integer (0-10)	Number of certificates earned during the year.
Annual_Papers_Patent_Count	Number papers or patent count	Numerical	Integer (0-5)	Number of research papers or patents published in a year.
Annual_Working_Hours	Annual Working Hours	Numerical	Float (1000.0-2500.0)	Total number of working hours in a year.
Current_Salary	Salary	Numerical	Float (e.g., 50000-500000)	Current salary of the employee.
Salary_Increase_Status	Increase Rate of Salary	Categorical	{Yes, No}	Indicates whether the employee received a salary increase.
Travel_Frequency	Frequency of Travel	Categorical	Rare, Occasional, Frequent	Frequency of business-related travel.
Remote_Work_Ratio	Ratio of Remote work	Numerical	Float [0.0-1.0]	Ratio of time worked remotely compared to total working time.
Peer_Feedback_Score	Score of Peer Feedback	Numerical	Float [0.0-5.0]	Average feedback score given by peers.
Manager_Feedback_Score	Score of Manager Feedback	Numerical	Float [0.0-5.0]	Evaluation score provided by the employee's direct manager.
Subordinate_Feedback_Score	Score of Subordinate Feedback	Numerical	Float [0.0-5.0]	Feedback score from subordinates or team members.
Customer_Feedback_Score	Score of Customer Feedback	Numerical	Float [0.0-5.0]	Evaluation score given by clients or external stakeholders.
Self_Evaluation_Score	Score of Self Feedback	Numerical	Float [0.0-5.0]	Self-assessment score provided by the employee.
Feedback_ID	Feedback ID	Numerical	Integer (>0), unique	Unique ID for each feedback entry.
Comment_Text	Comment	Text	5-5000 chars	Open-ended feedback content (PII should be masked on ingest).
Likert_Score	Score of Likert	Numerical	[1-5] or org scale	Numeric rating in the same form (if present).
Comment_Date	Comment date	Categorical / Datetime	Date format (YYYY-MM-DD)	Date/time the comment was submitted.
Comment_Language	Language of Comment	Text	{tr, en} (extensible)	Language of the comment (ISO 639-1 codes).

As shown in the table 3.1. above, the dataset consisting of 1000 employees includes various attributes related to each individual. Dataset was constructed using a combination of rule based assignments, randomly generated variables, and interdependent features to ensure realistic data behaviour. Data dictionary presented earlier represents the multi

year dataset structure. The component utilized by the HR engine corresponds to panel dataset. In addition to the multi year records, panel dataset includes prediction outputs produced by the sentiment and competency models, as well as labels corresponding to the expected outcomes for the models. In subsequent stages, these labels were concealed from the respective models during training, and performance evaluations were conducted based on the resulting predictions.

Detailed information on how each variable in the dataset was generated is provided below:

- **Employee_ID:** Employee_ID is a unique identifier assigned to each employee within the dataset. To protect the confidentiality of personal information, it was generated completely at random in a manner that prevents any association with real individuals. The IDs were not regenerated on a yearly basis; instead, each employee retained the same identifier across all years, enabling the longitudinal tracking of the employee's lifecycle. This approach ensured the creation of deterministic yet randomly assigned unique identifiers.

- **Year:** Year is a temporal feature that indicates the specific year to which each employee record belongs and was generated to construct a panel (longitudinal) data structure. A five year range (2021, 2022, 2023, 2024, and 2025) was defined. This time span provides the necessary temporal foundation for models such as promotion prediction, salary raise modelling, performance, trend analysis and sentiment/competency trend extraction. By assigning a deterministic sequence of years to each employee, the chronological progression of an employee's career can be accurately traced without temporal ambiguity. This variable also enables the identification of events such as employees who left the organization and those who joined in a given year.

- **Gender:** Gender is a categorical variable representing the sex information of employees. It consists of two categories (Male and Female). A random distribution was generated based on a 55% Male and 45% Female ratio. This

distribution was selected to avoid producing an overly balanced dataset and to simulate real world departmental gender diversity. The Gender attribute is entirely synthetic, not derived from any real employee data and cannot be associated with any identifiable individual. It was generated via random assignment, ensuring zero re identification risk under GDPR. Gender also serves as a supporting demographic feature in HR Engine models such as burnout risk estimation, performance modelling, department level fairness analysis, and sentiment trend difference analysis.

- **Age:** Age is a numerical variable representing the employee's age. It was generated using a statistically controlled method designed to mimic realistic corporate demographics. Ages were produced within the 22 to 60 range. Since age distributions in real organizations typically follow a bell shaped curve, a normal distribution with a mean of approximately 34–35 (μ) and a standard deviation of 7–8 (σ) was applied. No rules linking age to gender were defined; however, Education_Level values were assigned in relation to age. The Age variable serves as an important demographic input for several components of the HR Engine, such as burnout risk estimation and fairness analysis. Additionally, the natural increase in age across years provides a time dependent characteristic within the panel data structure.

- **Education_Level:** Education_Level is a categorical variable representing the employee's level of education. The dataset includes the following categories: High School, Associate Degree, Bachelor's Degree, Master's Degree, and PhD. These levels were selected to reflect the diversity of educational backgrounds typically observed in corporate environments. The distribution of education levels was generated randomly but under statistical control to mimic the demographic patterns commonly found in technology oriented and corporate organizations. The proportions were set as follows:
  Bachelor's Degree → 55% (largest group),
  Master's Degree → 25%,
  Associate Degree → 10%,

High School → 5% and

PhD → 5%.

Rather than using a purely random assignment, a logic consistent with age was applied. For example, ages 22–24 were more likely to receive a Bachelor's Degree, ages 25–30 had an increased likelihood of holding a Master's Degree, and employees above 30 had higher probabilities of Master's or PhD assignments. Since no employee was below age of 22, no downward inconsistency in education level occurred. High School and Associate Degree categories were distributed with minimal age based restrictions. Importantly, this relationship was constructed probabilistically rather than deterministically meaning there were no strict rules (e.g., age 35 does not guarantee a Master's degree). This approach ensured that the data remained realistic while avoiding overly deterministic patterns. Education_Level was utilized in the HR Engine as a demographic and socio professional input for models such as performance modelling, high potential employee detection, and fairness analysis.

- **University_Entry_Rank:** University_Entry_Rank is a numerical variable representing the employee's university entrance ranking. To simulate centralized university placement systems, values were generated within the range of 1,000 to 300,000. A right skewed distribution was applied, in which lower (better) rankings occur less frequently and higher (weaker) rankings occur more frequently. Specifically:

  1,000–50,000 → low density

  50,000–150,000 → highest density

  150,000–300,000 → medium density

  A weak probabilistic correlation was established with Education_Level. Employees with High School or Associate Degree backgrounds were not assigned a university entry rank. University_Entry_Rank functions as a professional profile indicator within HR Engine components such as High potential Employee Detection and Performance modelling. Additionally, it serves as a time invariant static feature within the panel data structure.

- **Language_Score:** Language_Score is a numerical variable representing the employee's foreign language proficiency. Values were assigned within the range of 0 to 100. Since language competency profiles in real organizations typically exhibit a wide distribution with a small number of high proficiency employees and a larger portion of beginner or intermediate level employees a truncated normal distribution was applied using a mean (μ) of 55–60 and a standard deviation (σ) of 15–18. No hard coded rules were defined; however, a mild probabilistic relationship was incorporated to maintain realism:
  Bachelor's, Master's and PhD holders were given slightly higher average Language_Score values,
  High School and Associate Degree holders were assigned broader and slightly lower distributions.
  This relationship is not deterministic (e.g., holding a Master's degree does not guarantee a high language score).

  Language_Score was used as an input feature in HR Engine components such as High potential Employee Detection and Performance modelling. In technology oriented organizations, foreign language ability is an important differentiator for global project involvement and promotion decisions, making this variable a valuable modelling input.

- **Number_of_Child:** Number_of_Child is an integer variable representing the number of children an employee has. For employees with no marriage or divorce events recorded in the Life_Event field, a value of 0 children was assigned. To obtain a realistic company wide distribution, the following proportions were applied:
  0 children → 55%
  1 child → 25%
  2 children → 15%
  3 children → 4%
  These proportions are aligned with demographic patterns reported by the Turkish Statistical Institute (TÜİK) and typical corporate workforce distributions.

Although no strict rules were defined, a mild age dependent probabilistic adjustment was applied to ensure realism:
Ages 22–28 → predominantly 0 children, occasionally 1
Ages 29–35 → wider distribution between 0–2 children
Above age 35 → increased likelihood of 1–3 children.
This adjustment is entirely probabilistic rather than deterministic; for example, being 35 years old does not imply having exactly two children. Number_of_Child was used as an input feature in HR Engine components such as Burnout Risk Estimation and Work–Life Balance analysis. While it contributes meaningfully to these models, none of the models incorporates any deterministic decision rules based solely on the number of children.

- **Start_Date:** Start_Date is a datetime variable representing the employee's hiring date within the organization. To support the multi year (2021–2025) panel structure, hiring dates were generated within the range of 2010 to 2025. Start_Date is not a purely random date; it was produced in a manner that does not conflict with the employee's age or total work experience. An approximate hiring year was derived using the 3.1:

$$\mathrm{Start}_{\mathrm{Date}} = \mathrm{Current}_{\mathrm{Year}} - \mathrm{Years}_{\mathrm{at\_company}} \tag{3.1}$$

After which the month and day were assigned randomly. Since the panel structure includes employees who leave the organization, temporal consistency rules were applied:
An employee with a Start_Date in 2024 does not appear in records for 2021.
An employee who leaves in 2023 does not have records for 2024–2025.
This ensures that the dataset remains temporally coherent and logically consistent. Start_Date was used within the HR Engine as a key input for computing experience based features.

- **Life_Event:** Life_Event is a categorical variable representing significant personal events that occur in the employee's life outside of work. It was created to synthetically model commonly observed personal circumstances in corporate

HR datasets. Since it is not derived from real employee information, it is fully anonymous, randomly assigned, and generated under statistically controlled conditions. The Life_Event variable includes the following categories: None (no life event), Marriage, Divorce, and Relocation. Life_Event values were not assigned purely at random; to ensure realism, a weak probabilistic relationship with age was incorporated:

Ages 22–27 → Low likelihood of Marriage and New Child, high likelihood of None,

Ages 28–40 → Increased likelihood of Marriage, New Child, and Relocation,

Ages 40+ → Slightly higher likelihood of Divorce and Serious Illness.

However, this relationship is not deterministic; for example, being 36 years old does not imply a certain Life_Event category such as New Child or Marriage. The intention was simply to produce a realistic distribution within the dataset. Life_Event was designed to vary across years for some employees. For example:

2021 → None

2022 → Marriage

2023 → Marriage

2024 → Marriage

2025 → Relocation

This temporal variability enables the analysis of life event impacts in HR Engine components such as Burnout Risk Estimation and Work–Life Balance modelling.

- **Department:** Department is a categorical variable representing the organizational unit in which the employee works. The department list was constructed to simulate the core organizational structures commonly found in technology oriented and corporate companies, and includes the following categories:

  R&D,

  IT,

  QA / Test Engineering,

  Operations,

HR (Human Resources),

Finance,

Sales,

Product Management.

This structure reflects typical departmental distributions observed in medium and large scale technology firms. Since such organizations generally employ a higher number of technical personnel, the department distribution was generated using an intentionally imbalanced yet realistic ratio:

Engineering / R&D → 29%

IT → 21%

QA / Test Engineering → 12%

Operations → 15%

Sales → 8%

Finance → 5%

HR → 4%

Product Management → 5%

Department assignment was not fully random; a weak role based probabilistic relationship was applied to ensure consistency:

Software Engineer" and "Data Scientist roles → higher likelihood of being assigned to IT or Engineering

QA Engineer → QA department

HR Specialist → HR

Sales Executive → Sales

Financial Analyst → Finance

This relationship is not deterministic; it simply increases realism while preserving multiple possible department matches for each role. Since organizational transfers were also considered, Department is not treated as a static (time invariant) variable. The Department variable served as a key segmenting feature in several HR Engine models.

- **Role:** Role is a categorical (text) variable representing the employee's job title within the organization. The role list was constructed to simulate the typical position structure of a technology oriented company and includes the following:

  Technical Roles
  • Software Engineer
  • Data Scientist
  • QA Engineer
  • DevOps Engineer
  • System Engineer

  Business and Management Oriented Roles
  • Product Manager
  • Business Analyst
  • HR Specialist
  • Finance Specialist
  • Sales Executive
  • Operations Specialist

  This structure ensures a balanced representation of both technical and support functions. Since technical roles are generally more common in technology companies, the dataset was designed such that 60–65% of employees hold technical positions, while 35–40% are assigned to business or support roles.

  Roles were not assigned entirely at random; instead, mild probabilistic relationships with Education_Level and Department were incorporated. For example:

  • IT → Software Engineer, DevOps Engineer
  • QA → QA Engineer, Test Engineer
  • HR → HR Specialist

  Unlike Department, roles were allowed to change over the years, enabling the simulation of career progression. This made promotion pathways and temporal

modelling more realistic. Role serves as an important feature within the HR Engine, particularly in Performance modelling and Burnout Risk Estimation.

- **Current_Position:** Current_Position is a categorical variable representing the employee's present seniority level within the organization. Unlike the Role variable, Current_Position directly reflects seniority and is used to model employees' career stages. The categories include Junior, Mid Level, Senior, Lead, and Manager. In corporate organizational structures, the hierarchy typically follows a pyramid distribution, where the number of employees decreases as seniority increases. This pattern was preserved in the synthetic dataset as well for example, approximately 30% of employees were assigned as Junior, whereas only around 3% were assigned as Manager. Current_Position was not assigned completely at random; instead, it was generated with mild probabilistic relationships based on Years_At_Company and Total_Work_Experience:

  • 0–3 years of experience → higher likelihood of Junior / Mid Level
  • 3–7 years → Mid Level / Senior
  • 7–12 years → Senior / Lead
  • 12+ years → higher likelihood of Lead / Manager

  No deterministic rules were applied (e.g., not every employee with 10 years of experience is necessarily senior), but probability weights were adjusted to produce realistic distributions. Current_Position was designed as a dynamic (time varying) variable to simulate career progression. For example, an employee may follow the trajectory:

  2021 → Junior
  2022 → Mid Level
  2023 → Mid Level
  2024 → Senior
  2025 → Senior

Within the HR Engine, Current_Position served as a critical input for model components such as Salary Raise modelling and Leadership Potential Assessment.

- **Years_At_Company:** Years_At_Company is a numerical (integer) variable representing the total length of time an employee has spent within the organization. It was not generated randomly; instead, it was calculated deterministically using the 3.2:

$$\text{Years}_{\text{at}_{\text{company}}} = \text{Current}_{\text{Year}} - \text{Start}_{\text{Date}} \quad (3.2)$$

The value was recalculated for each year so that it increases consistently over time. Additionally, the dataset was designed such that Total_Work_Experience ≥ Years_At_Company, ensuring the prevention of unrealistic profiles (e.g., a 22 year old with 10 years of tenure) and maintaining overall career consistency. Since the panel structure includes employees who leave the organization, Years_At_Company does not extend beyond the year in which the employee exits.

- **Years_With_Curr_Manager:** Years_With_Curr_Manager is a numerical (integer) variable representing the amount of time an employee has spent under their current manager. This variable was synthetically generated under the assumption that managers may change over time. It was not produced randomly; instead, it follows logical constraints specifically, the number of years spent with a manager cannot exceed the employee's total years at the company. The data generation process reflects realistic corporate patterns in which most employees work with the same manager for 1–3 years, while a smaller portion remain under the same manager for 4–5 years.

- **Total_Work_Experience:** Total_Work_Experience is a numerical (integer) variable representing an employee's total years of professional experience. It was not generated randomly; instead, it was produced in a manner consistent with Age, Start_Date, and Years_At_Company, since an individual's total work

experience cannot exceed their age. In alignment with the panel data structure, Total_Work_Experience increases over the years for employees who continue working.

- **Performance_Score:** Performance_Score is a continuous numerical variable (ranging from 0 to 5) representing the employee's annual performance evaluation. This variable was generated using a fully rule based yet noisy synthetic model influenced directly by several other features in the dataset. Specific weights were assigned to selected variables, and a weighted average was computed:

  PS: Performance Score,

  G: Goal_Attainment_Score (~30%),

  Q: Deliverables_Quality_Score (~25%),

  MFB: Manager_Feedback_Score (~20%),

  PSS: Peer_Feedback_Score + Subordinate_Feedback_Score + Self_Evaluation_Score (~15%),

  E: Task_Efficiency + Deadline_Adherence_Rate (~10%)

  A noise term was added annually to reflect real world subjectivity. The 3.3 used is:

$$PS = 0.3 \times G + 0.25 \times Q + 0.2 \times MFB + 0.15 \times PSS + 0.10 \times E + \in \quad (3.3)$$

$$\in \sim U(0.15,0.25) \quad (3.4)$$

  Performance_Score were generated by adding a uniform noise term to the weighted base score according to 3.4. Performance_Score is not fixed or deterministic across years; it is probabilistic. For example, an employee with a low Manager_Feedback_Score does not necessarily receive a low performance score it only increases the likelihood of such an outcome. As one of the most critical variables in the HR Engine, Performance_Score functions as a hidden label for several downstream models.

- **Goal_Attainment_Score:** Goal_Attainment_Score is a continuous numerical variable (ranging from 0 to 5) that reflects the extent to which an employee completes the objectives assigned throughout the year. It was not assigned randomly; instead, it was generated by establishing weak to moderate correlations with variables such as Task_Efficiency_Score, Deadline_Adherence_Rate, and Deliverables_Quality_Score.

  G: Goal_Attainment_Score,
  Ef: Task_Efficiency,
  DAd: Deadline_Adherence_Rate,
  Q: Deliverables_Quality_Score

$$\mathrm{G} = 0.4 \times Ef + 0.3 \times DAd + 0.3 \times Q + \in \quad (3.5)$$

  The score in 3.5 was recalculated each year to account for factors such as varying goal difficulty, differences in evaluation styles across managers, and natural fluctuations in employee performance. Within the HR Engine, Goal_Attainment_Score serves as a core feature used in trend based calculations, including ΔGoal, MA2, and trend_up. A uniform noise approach, similar to the one applied in the performance score calculation, was adopted.

- **Strategic_Initiative_Impact:** Strategic_Initiative_Impact is a continuous numerical variable (ranging from 0 to 5) representing the employee's level of contribution to strategic projects during the year. Strategic initiatives include high impact activities such as innovation projects, R&D efforts, digital transformation programs, critical deliveries, and process improvement actions. For this reason, the variable was not assigned randomly; instead, weak to moderate correlations were established with Role, Department, Seniority/Current_Position, Deliverables_Quality_Score, and Ownership_Index. Within the HR Engine, this variable plays a significant role, particularly in leadership assessment and high potential employee identification.

- **Deliverables_Quality_Score:** Deliverables_Quality_Score is a continuous numerical variable (ranging from 0 to 5) representing the quality of the work outputs produced by an employee throughout the year. The score reflects indicators such as accuracy, consistency, the need for rework, error rate, and customer or cross team satisfaction. As a critical component of performance evaluation systems, this variable was generated using a rule based approach supplemented with noise to ensure realism.

  In real corporate environments, metrics such as Deliverables_Quality_Score, Goal_Attainment_Score, and Task_Efficiency_Score are commonly tracked through tools like JIRA. Since this dataset is synthetic, these values were produced in a way that reflects realistic organizational behaviour. The score was not assigned randomly; instead, weak to moderate correlations were established with Task_Efficiency_Score, Deadline_Adherence_Rate, Performance_Score, and Role/Department. None of these relationships are deterministic.

  Within the panel data structure, Deliverables_Quality_Score is recalculated each year, enabling trend based analyses such as ΔQuality and Moving Average (MA2). As one of the most influential components contributing approximately 25% to Performance_Score it serves as a critical input within the HR Engine.

- **Task_Efficiency_Score:** Task_Efficiency_Score is a continuous numerical variable (ranging from 0 to 5) representing how efficiently an employee completes assigned tasks throughout the year. The score reflects factors such as timeliness, workload handling, rework requirements, and adherence to sprint or task cycles. Since it is a key metric in performance evaluation systems, it was generated using a rule based structure enhanced with noise for realism. In real corporate environments, metrics like Task_Efficiency, Deliverables_Quality_ Score, and Goal_Attainment_Score are typically tracked through tools such as JIRA. Because this dataset is synthetic, the score was produced in a way that mimics realistic organizational behaviour. It was not assigned randomly; instead, weak to moderate correlations were established with Deadline_Adherence_Rate,

Deliverables_Quality_Score, Performance_Score, and Role/Department. These relationships are non deterministic. Within the panel data structure, Task_Efficiency_Score is recalculated annually, making trend analyses such as ΔEfficiency and MA2 possible. As one of the contributing components of Performance_Score, it serves as an important input within the HR Engine

- **Deadline_Adherence_Rate:** Deadline_Adherence_Rate is a continuous numerical variable (ranging from 0 to 5) that represents the extent to which an employee adheres to task deadlines throughout the year. As a KPI indicator, it is widely used in domains such as software development, QA, IT operations, and project management. Therefore, it was not assigned randomly; instead, it was generated using rule based, probabilistic, and noise infused methods designed to reflect realistic organizational behaviour. The variable was modeled with weak to moderate probabilistic correlations to Task_Efficiency_Score, Deliverables_ Quality_ Score, and Role/Department. Deadline adherence metrics are particularly important in software and QA environments, where they function as key behavioural KPIs. Consequently, Deadline_Adherence_Rate plays a significant role in HR Engine components such as Performance_Score and Goal_Attainment_Score.

- **Manager_Feedback_Score:** Manager_Feedback_Score is a continuous numerical variable (ranging from 0 to 5) that represents the annual evaluation score given to the employee by their manager. Since manager feedback is one of the most influential components in corporate performance systems, it was not generated randomly. Instead, it was produced using a rule based, probabilistic model with added noise to reflect realistic evaluation patterns. The score was weakly to moderately correlated with Deliverables_Quality_Score, Goal_Attainment_Score, Ownership_Level, and related behavioural indicators. Manager feedback values were recalculated annually to reflect changes in projects, performance fluctuations, and potential manager transitions. As it constitutes a significant portion of the Performance_Score, this variable plays an

important role within the HR Engine, particularly in performance modelling, and leadership potential assessment.

- **Peer_Feedback_Score:** Peer_Feedback_Score is a continuous numerical variable (ranging from 0 to 5) that represents the feedback an employee receives from colleagues within the same team or across cross functional teams. In real organizational settings, peer feedback evaluates behavioural competencies such as team communication, collaboration, supportiveness, and overall team cohesion. In this dataset, the score was generated by establishing weak to moderate probabilistic relationships with related behavioural indicators. In essence, Peer_Feedback_Score is one of the strongest variables modelling an employee's social and behavioural competencies.

- **Subordinate_Feedback_Score:** Subordinate_Feedback_Score is a continuous numerical variable (ranging from 0 to 5) representing the feedback an employee receives from team members in subordinate positions. This variable was generated only for managerial roles such as Senior, Lead, and Manager or for positions that involve mentorship or coordination responsibilities within the team.

- **Self_Evaluation_Score:** Self_Evaluation_Score is a continuous numerical variable (ranging from 0 to 5) that represents the employee's self assessment of their own performance. It corresponds to the 'self review' component of 360 degree evaluation systems. Within the HR Engine and metric calculations, it was used only for monitoring purposes. Since employees tend to rate themselves relatively high, this variable was not treated as a critical criterion in decision making models.

- **Customer_Feedback_Score:** Customer_Feedback_Score is a continuous numerical variable (ranging from 0 to 5) that represents the feedback an employee receives from interactions with customers, either internal or external. If an employee does not directly engage with external customers, the score reflects

evaluations received from cross functional teams. Therefore, it was not assigned randomly.

- **Ownership_Level:** Ownership_Level is a continuous behavioural competency indicator (ranging from 0 to 5) that reflects the extent to which an employee takes ownership of their work, responsibilities, and outcomes. Since it is a soft skill based variable, it was not assigned randomly. Instead, weak to moderate correlations were established with Strategic_Initiative_Impact, Deliverables_ Quality_Score, Task_Efficiency_Score and Current_Position/Seniority, although none of these relationships are deterministic. Ownership_Level was recalculated each year, as ownership may increase with experience, fluctuate when an employee changes teams, or vary depending on project intensity. Within the HR Engine, it provides strong signals for models such as Leadership Potential and Performance modelling.

- **Code_Commit_Count:** Code_Commit_Count is an integer variable representing the number of code commits an employee makes within a year (e.g., via version control systems such as Git). Since this metric is meaningful only for technical roles, specific rules were applied during synthetic data generation. Commit values were produced exclusively for roles such as Software Engineer, Data Scientist, DevOps Engineer, and QA Engineer. The assigned commit ranges were informed by real world commit behaviour and were loosely correlated with role and position; however, the relationship is not deterministic and reflects only weak to moderate probabilistic associations.

- **Documentation_Contribution:** Documentation_Contribution is a continuous numerical variable (ranging from 0 to 5) representing the employee's level of contribution to project documentation within a given year (e.g., technical documents, user manuals, test documentation, architectural drafts, API documentation, etc.). Similar to Code_Commit_Count, the score was modeled with weak to moderate probabilistic relationships based on role and seniority.

Although documentation contribution is not a primary performance criterion, it serves as a valuable supportive behavioural and technical indicator.

- **Problem_Resolution_Count:** Problem_Resolution_Count is an integer numerical variable (0–N) representing the number of problems, incidents, bugs, errors, support requests, or operational interventions an employee resolves within a given year. The distribution was modeled approximately as follows:

  0–20 → low workload (common in non QA teams)
  20–60 → medium workload
  60–150 → high workload (QA & Operations)
  150–300 → edge cases involving very high operational demand

  These values were assigned using probabilistic weighting. Within the HR Engine, this variable does not directly determine performance; instead, it is used as a supporting technical KPI.

- **Annual_Training_Count:** Annual_Training_Count is an integer variable (0–N) representing the number of trainings an employee completes within a given year. These trainings may include technical courses, personal development programs, mandatory corporate trainings, e learning modules, and certificate preparation sessions. This variable was assigned completely at random.

- **Annual_Certificates_Count:** Annual_Certificates_Count is an integer variable (0–N) representing the number of certificates an employee obtains within a given year. The values were determined with a small degree of randomness, loosely aligned with the employee's training activity.

- **Annual_Papers_Patent_Count:** Annual_Papers_Patent_Count is an integer variable (0–N) representing the number of academic publications (such as journal papers), technical reports, conference proceedings, or patent applications produced by an employee within a given year. The values were assigned in a way

that loosely correlates with Education_Level, Role, and Current_Position for example, employees holding a PhD were given a higher likelihood of producing such outputs.

- **Work_Life_Balance_Score:** Work_Life_Balance_Score is a continuous well being indicator (ranging from 0 to 5) that measures how effectively an employee maintains a balance between work and personal life. The score was generated by establishing weak to moderate probabilistic relationships with variables such as Monthly_Working_Hours, Remote_Work_Ratio, Travel_Frequency, Life_Event, and Burnout_Risk_Score. For example, if an employee has high average monthly working hours, frequent travel, or a challenging life event such as divorce in a given year, their well being score is probabilistically assigned at a relatively lower level. None of these relationships are deterministic; probabilistic weighting was used solely to create realistic distribution patterns. Within the HR Engine, this variable is used as a supportive and explanatory feature rather than a primary decision making criterion.

- **Current_Salary:** Current_Salary is a numerical variable representing an employee's annual salary for a given year. Salaries were updated annually by applying an 18% inflation adjustment, making this variable time varying across the years. Salary levels differ across roles, and both promoted employees and those receiving raises without promotion were modeled accordingly, resulting in salary growth over time for all employees. No employee experienced a salary decrease; the inflation based increase was consistently applied each year.

- **Monthly_Working_Hours:** Monthly_Working_Hours is an integer variable representing the total number of hours an employee works within a month. In real organizational settings, monthly working hours typically fall within the following ranges:

  Standard: 160–180 hours for a full time employee
  During high intensity periods: 180–210+ hours

During low intensity periods: 150–160 hours

In the synthetic dataset, this structure was preserved, and the values were generated with weak to moderate probabilistic relationships to factors such as changes in the employee's role over the years and variations in life events.

- **Remote_Work_Ratio:** Remote_Work_Ratio is a variable representing the proportion of time an employee works remotely within a given year. The ratio was regenerated each year, as company policies may shift, team cultures may evolve, and employees may change roles.

- **Travel_Frequency:** Travel_Frequency is a categorical or integer variable (0–N) representing how often an employee engages in work related travel within a given year. The values were not assigned randomly; instead, they were generated using a role based, department based, and probabilistic approach that reflects real corporate work patterns. In actual organizations, travel intensity tends to be higher in roles such as Sales Executive, Product Manager, and certain QA positions, so these roles were assigned higher probabilistic weights. Within the HR Engine, Travel_Frequency serves as a supporting signal for the Burnout Risk Estimation model, although it is not a direct determinant of performance.

- **Manager_Readiness_Score:** Manager_Readiness_Score is a continuous variable (ranging from 0 to 5) that represents an assessment of how prepared an employee is to assume a managerial role in the future. This variable was created within the synthetic dataset to support the leadership potential modelling process. The score was not generated randomly; instead, it was computed using weak to moderate probabilistic relationships with behavioural competencies, tenure, experience, and performance indicators. Manager_Readiness_Score was recalculated each year. Within the HR Engine, it serves as a strong signal specifically for the Leadership Potential model.

- **Burnout_Risk_Score:** Burnout_Risk_Score is a continuous variable (ranging from 0 to 5) representing an employee's risk of burnout within a given year. The score was generated using weak to moderate probabilistic relationships with variables such as Monthly_Working_Hours,Work_Life_Balance_ Score,Travel_ Frequency,Deadline_Adherence_Rate,Task_Efficiency,Life_Event,andSentime nt Trends. As with other variables in the dataset, the relationships are not deterministic; probabilistic weighting was applied solely to create realistic distribution patterns. Within the HR Engine, this score does not directly influence promotion or salary raise decisions, but it serves as a critical early warning signal for HR. Burnout_Risk_Score is a hidden variable for the burnout risk model, meaning the model does not access this value directly in order to ensure an unbiased evaluation of model performance.

- **Career_Potential:** Career_Potential is an ordinal variable (Low / Medium / High) representing an employee's long term career growth potential. The variable was not generated randomly; instead, it was calculated using weak to moderate probabilistic relationships with the following indicators: Performance_Score, Ownership_Level,Innovation_Index,Communication_Index,Manager_Feedback _Score,Education_LevelandYears_At_Company/Total_Work_Experience.
  None of these relationships are deterministic; probabilistic weighting was applied solely to ensure realistic distribution patterns. Career_Potential functions as a supporting signal within the HR Engine, particularly for Promotion, Succession Planning and High Potential modelling. Similar to Burnout_Risk_Score it is not directly exposed to predictive models.

- **Succession_Ready:** Succession_Ready is a binary variable (0/1) indicating how prepared an employee is to assume a critical role in the organization. The probability of Succession_Ready = 1 was intentionally kept low (approximately 10–15% of employees), reflecting real corporate succession planning dynamics. In practice, such readiness is often determined through systems similar to the 9 box grid. Accordingly, the variable was generated using probabilistic relationships with Manager_Readiness_Score, Leadership_Index, Performance_

Score, Ownership_Level, and Total_Work_Experience / Years_At_Company. Succession_Ready is used within the HR Engine to support Leadership Pipeline modelling and critical role succession analysis. Similar to Burnout_Risk_Score, it is not directly exposed to predictive models.

- **High_Potential_Score:** High_Potential_Score is a continuous variable (0–5) representing the likelihood that an employee is classified as a High Potential (HiPo) talent. The score was generated probabilistically based on several key behavioural and performance indicators, including Performance_Score, Innovation_Index,Ownership_Level,Manager_Readiness_Score,Communicatio n_Index, and Career_Potential. This variable serves as a strong behavioural signal in Promotion, Leadership Potential, and Succession Planning models. Similar to Burnout_Risk_Score, it is not directly exposed to predictive models.

- **Promotion_Status:** Promotion_Status is a variable (Promoted or Not Promoted) indicating whether an employee received a promotion within a given year. This variable was not assigned randomly. If an employee's position advanced in subsequent years for example, from Junior to Mid Level or from Lead to Manager the corresponding record was labeled as promoted. The number of promoted employees was designed to simulate realistic promotion rates in a corporate organization. Current_Position and Current_Salary were updated accordingly for employees who received a promotion. Within the HR Engine, Promotion_Status serves as the primary target (label) for the Promotion Prediction model and is therefore kept hidden from the model.

- **Salary_Increase_Status:** Salary_Increase_Status is a variable indicating whether an employee received an additional merit based salary increase (raise or no raise) in a given year. If an employee was promoted, an extra 20% increase was applied on top of the inflation adjustment. In line with realistic organizational structures, some employees also received an additional 20% merit increase independent of promotion, and these cases were marked as 'raise' in the corresponding column. The salary increase is not directly tied to the performance

model or the burnout model; however, it is used as an important input feature within the HR Engine.

- **Likert_Score:** This scale was used to quantify subjective expressions within peer and manager comments. It was incorporated into the dataset using the following categorical structure:
  1 → Strongly Negative / Strongly Disagree
  2 → Negative / Disagree
  3 → Neutral
  4 → Positive / Agree
  5 → Strongly Positive / Strongly Agree.

In addition to the multi year dataset described above, employee feedback was collected in the form of textual comments. These comments were analysed using XLM R–based sentiment modelling and One vs Rest Logistic Regression classifier competency classification. The feedback texts were generated in alignment with 360 degree evaluation outcomes such as Manager_Feedback_Score and Peer_Feedback_Score and contained positive, neutral, or negative expressions, as well as keywords reflecting the employee's competencies in Communication, Leadership, Teamwork, Innovation, Ownership, and ProblemSolving.

These comments were processed by the sentiment model, which produced annual indicators such as Sentiment_Mean, Sentiment_Variance, Pos/Neu/Neg Ratios, and inter year changes captured by Delta_Sentiment_Index. Similarly, the competency classifier generated behavioural competency indices Communication_Index, Leadership_Index, Teamwork_Index, Innovation_Index, Ownership_Index, and ProblemSolving_Index along with the Dominant_Competency and yearly change indicators, all of which were integrated into the panel data structure.

Higher level annual summary features, such as Num_Feedback_Norm, Manager_Pos_Ratio, Manager_Neg_Ratio, multi year trend measures (MA2 Moving Average), trend flags (Trend_Up/Down), and normalized index scores (..._Norm), were

also included. These variables transform raw NLP outputs into structured, year level quantitative signals and serve as critical inputs for enhancing the accuracy of performance, promotion, burnout, leadership potential, and organizational behaviour models within the HR Engine.

To make the dataset generation techniques and formulas described above more concrete, Figure 3.1 has been prepared for a single employee within the dataset:

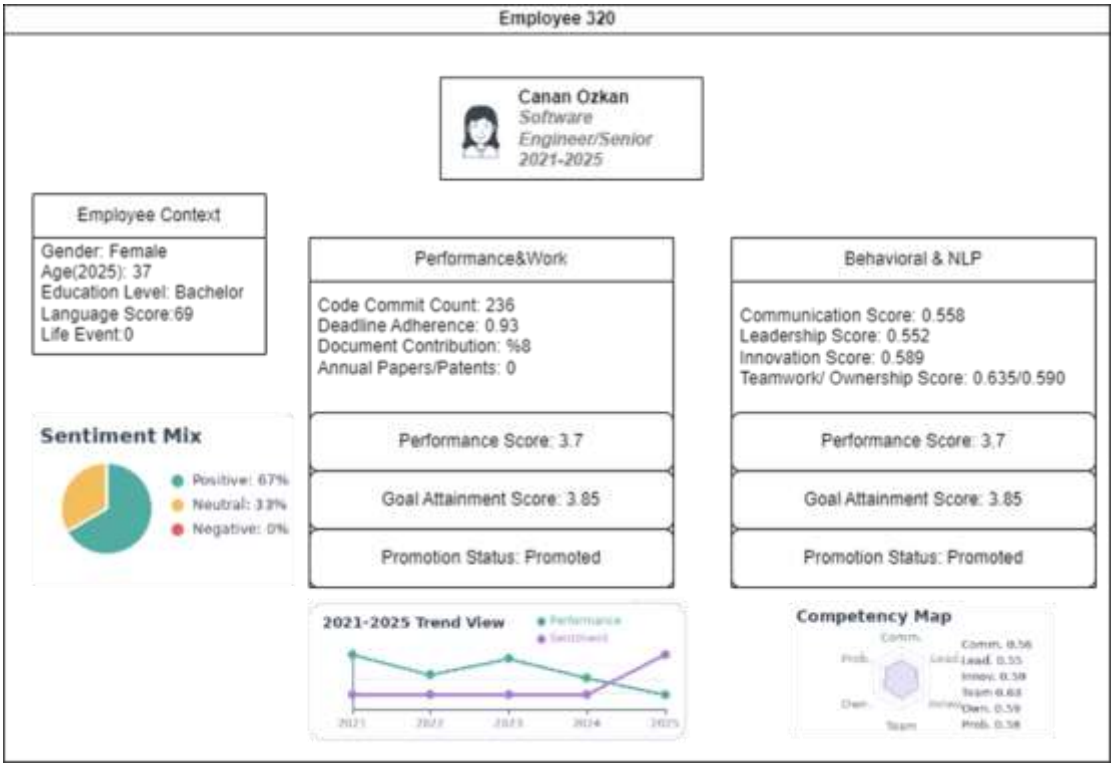


Figure 3.1 A sample employee from the dataset

3.2.2 Synthetic data generation protocol

This section describes the protocols and principles followed in the construction of the dataset presented in the Data Dictionary within the scope of this study. First and foremost, the data generation process was not conducted in a purely random manner. Instead, in order to realistically simulate an organizational structure and human resources processes, many variables in the dataset were generated using rule based mechanisms.

These rules were defined by considering established practices in human resources management, organizational hierarchies, and relevant academic literature.

For example;

- A logical dependency was established between Years_of_Experience and Role/Title, ensuring that higher level positions were represented by employees with greater levels of professional experience.
- The Education_Level variable was distributed in alignment with departmental and technical role requirements, with higher proportions of postgraduate education assigned to positions requiring advanced technical expertise.
- Performance scores were generated by taking into account employees' experience levels, position seniority, and the intensity of annual feedback, thereby preventing the performance variable from being distributed in a fully random manner.
- Outcome variables such as Promotion_Status and Salary_Increase were generated through indirect, multi factor rules rather than being directly derived from model input features. This design choice deliberately mitigated the risk of label leakage and circular inference.

Special care was taken in the generation of textual feedback (employee comments). These comments were created:

- Using predefined competency oriented vocabularies and phrase pools,
- By maintaining semantic consistency between the sentiment polarity (positive, neutral, negative) and the textual content,
- In a manner that allows each feedback entry to convey signals related to one or more competency dimensions.

This approach ensured that the natural language processing models learned contextual and thematic patterns rather than relying on superficial keyword repetition.

Privacy considerations were treated not merely as an ethical requirement but as a core design principle throughout the data generation process. No direct identifiers that could be associated with real individuals (such as names, email addresses, or identification

numbers) were generated. All records were constructed to be fully anonymous and synthetic. At the same time, the statistical distributions and inter variable relationships were preserved to closely resemble those observed in real corporate environments.

Finally, the dataset was designed in accordance with the principles of reproducibility and scalability. Using the same rule sets, datasets of varying sizes can be generated, and new variables or categories can be incorporated without compromising structural consistency. This design choice strengthens the methodological robustness of the study and supports the adaptability of the proposed AI supported competency assessment framework to different organizational scenarios.

The mathematical expressions presented in this section describe the general data generation protocol that represents the process through which all variables defined in detail in the data dictionary were constructed.

This 3.5 represents the core formulation that summarizes how the entire dataset is generated.

$$X_{i,t} = f(X_{(i,t)-1}) \in_{i,t} \tag{3.5}$$

$X_{i,t}$: A vector that represents all variables of the *i* th employee in year *t*, including demographic attributes, performance indicators, numerical scores, and text derived summary features.

$f(X_{(i,t)-1}) \in_{i,t}$: A rule based transition function that logically derives the current year values from the previous year's state. (Age increases over time, *Years_At_Company* increases consistently etc.)

$\in_{i,t}$: A random noise term representing human factors, managerial differences, measurement errors, and organizational uncertainties.

$$X_i \sim \text{Categorical}(\pi) \tag{3.6}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.7}$$

i: The index of the employee (e.g., 1st employee, 2nd employee, etc.)

X_i: The categorical variable to be generated (e.g., Department, Gender, Role)
K: The total number of categories (e.g., 8 departments in the Department variable)
π_k:The probability that the variable X_i takes the k th category

For each employee the value of is X_i determined by selecting one category according to predefined probabilities in the 3.6 and 3.7. The sum of these probabilities must equal 1, as the total probability mass must be conserved. For example, when defining a department distribution with eight categories, the probabilities assigned to all eight categories collectively sum to 1.

While generating the 360 degree evaluation comments, the following 3.8 was taken into consideration.

$$Y_i \sim G(C_i T_i L_i) \tag{3.8}$$

Y_i: The generated textual content (comment)
G: The text generation function (generator)
C_i: The selected competency label
T_i: The selected sentiment polarity
L_i: The length of the generated text

The above formulation was used to guide both the selection of words and phrases based on competency and the determination of tone based on sentiment. For example, if comments about an employee include expressions such as *"guidance," "decision making,"* and *"mentoring,"* the competency weighting results in a higher Leadership score. Similarly, the presence of expressions such as *"improvement," "concern,"* and related terms leads the sentiment polarity to be evaluated as negative. Consequently, the generated texts are not random; rather, they are produced by considering both *what* is being expressed and *how* it is expressed, with contextual coherence explicitly preserved.

In addition to the protocols described above, the following considerations were also taken into account during the data generation process:

- Stochastic noise terms were deliberately introduced to simulate human subjectivity, evaluator variability, and organizational uncertainty.
- No deterministic relationships were enforced between demographic attributes (e.g., gender) and performance related outcomes, in order to avoid introducing artificial bias into downstream models.
- Outcome variables such as promotion and salary increase were generated independently from textual features to prevent label leakage and circular inference.
- Following data generation, descriptive statistics and inter variable correlations were examined to ensure that the synthetic data exhibited realistic distributional properties and avoided pathological or degenerate patterns.
- All generation rules were parameterized, enabling reproducible dataset construction and controlled scalability.

3.2.3 Data preprocessing

The data pre processing process conducted in this study forms the foundation of the AI Supported Competency Assessment Framework. The process was systematically designed for both the first phase (single year data for 2025) and the second phase (multi year data for the 2021–2025 period). The goal is to transform textual, numerical, and structural human resources data into an analytically consistent, reliable, and reproducible format, thereby creating an integrated database capable of examining both current and temporal trends.

The data preprocessing process generally consists of the following subheadings: data aggregation and schema alignment, data cleaning and normalization, conversion of text to computational representation, labelling and class balance, temporal alignment and derived features, drift and leakage checks, scaling and selection, versioning, and quality assurance.

3.2.3.1 Data cleaning and normalization (textual + numeric)

The data cleaning process was applied to both textual and numerical data.

In the first phase, free form employee feedback was focused on; unnecessary spaces, HTML tags, and special characters were removed; stopwords were removed; text was converted to lowercase, accents were cleaned, and only meaningful sentences were included in the analysis. Duplicate records were merged, and missing or empty content was removed from the data set.

In the second phase, data inconsistencies between years were resolved; unit differences in fields such as date, salary, education, and performance scores were standardized. Interpolation or annual average estimation methods were used for missing annual values.

The goal in both phases was to obtain a database that was free of noise and ready for analysis.

3.2.3.2 Merging data sources and schema harmonization

Data collected from different corporate systems (performance evaluations, 360° feedback forms, training records, job changes, compensation information, etc.) was combined for both stages.

In the first stage (2025): Only data from 2025 was used; performance scores, text based feedback, and personnel information were combined under a common schema.

In the second stage (2021–2025): This structure was expanded to include data from the previous five years. Variable names, units of measurement, and formats that changed over the years were harmonized through schema standardization.

In both stages, Employee_ID was defined as the primary key; in the second stage, the Year variable was also added to ensure panel data integrity.

3.2.3.3 Conversion of texts into computational representation

The success of the natural language processing (NLP) based models used in this study is directly dependent on the quality of the digital representation of texts. The "Transformation of Texts into Computational Representations" phase constitutes one of the key components that determine both the semantic extraction and learnability levels of the AI powered competency assessment framework.

This process was conducted in two stages:

In the first phase (single year – 2025): texts were directly converted into model input data.

In the second phase (multi year – 2021–2025): the representations from the first phase were summarized on an annual basis and reconstructed to capture employee sentiment and competency trends over five years.
The goal is to transform natural language data obtained from employee feedback into multi layered vector representations containing statistical, semantic, and contextual information and to make these representations comparable over time.

- **Word Level Representation: TF–IDF Approach (First Stage):** In the first stage, the TF–IDF (Term Frequency–Inverse Document Frequency) method was used to capture the basic statistical properties of the texts. This method calculates the importance of a particular word in a document (e.g., an employee's feedback) by weighting it according to its overall prevalence across all documents.
  Mathematically:

$$TFIDF_{t,d} = TF_{t,d} \times \log(\frac{N}{\mathrm{D}F_t}) \quad (3.9)$$

  $TF_{t,d}$: Frequency of term t in document d,

DF_t : Number of documents in which the term occurs,
N : Total number of documents.

This process allows the model to suppress frequently occurring but low sense words ("and," "but," "like") and give more weight to rare but meaningful words ("leadership," "development," "innovative"). As a result, each feedback text is represented by a sparse vector consisting of thousands of dimensions. These representations form a particularly powerful input layer for linear classifiers (Logistic Regression, LinearSVC, and Ridge Classifier).

- **Word Embedding and Contextual Embedding (First Stage):** Word embedding based representations have also been used to go beyond statistical representations and model the semantic relationships of words. This approach represents words in a high dimensional continuous space based on their semantic similarity. Words with similar meanings (e.g., "leadership" and "team management") are located close to each other in this space. Two types of embedding methods were applied in the study:

- **Contextual embedding methods (XLM R, BERTurk):** These models take into account the fact that the same word can acquire different meanings in different contexts. For example, the word "strong" might convey a positive sentiment ("a strong leader") in one sentence and a negative sentiment ("an overly strong desire for control") in another. These transformer based models generate vectors by considering not only the word itself but also the entire context around it (the context window). Each feedback sentence was transformed into high dimensional dense embedding vectors using these methods, thus representing linguistic, semantic, and emotional information simultaneously.

- **Sentence and Document Level Representation (First Stage):** Since the majority of employee feedback consists of multiple sentences, sentence embedding and document level aggregation approaches were applied to ensure the model could grasp the entire text at the semantic level rather than individual

sentences. The embedding vectors for each sentence were combined using mean pooling or attention based pooling methods, thus representing each employee's feedback as a single "text vector." This process allowed us to capture not only word frequencies but also contextual tone, emotional coherence, and expression complexity of the texts. The representation obtained at this level was used as input for both sentiment analysis (sentiment classification) and competency classification models.

- **Hybrid Feature Representation and Fusion (First Stage):** At the end of the first stage, the resulting different text representations (TF–IDF and BERT based embeddings) were combined to create a hybrid feature matrix. Additionally, numerical and categorical variables related to employees (e.g., department, position, years of seniority, education level) were integrated into the text representations. One hot encoding was applied to categorical variables, and continuous variables were standardized using z score normalization As a result, each employee's utterance was transformed into a multidimensional, multimodal representation that included numerical indicators based on their organizational profile along with linguistic features (textual content) This combined structure allowed the model to learn textual, contextual, and structural factors simultaneously.

- **Creating Annual Representations (Stage Two):** In the second stage, all text representations obtained in the first stage were summarized at the annual level and integrated into the panel data structure (Employee_ID–Year). The purpose of this stage is to generate time series features (temporal features) from individual feedback and to numerically track employee development trends over the years. The following summary variables were derived for each year:

  Sentiment_Mean: Indicates the average sentiment score for that year.

  Sentiment_Variance: Measures annual sentiment variability (emotional fluctuation).

  Competency_Distribution: Indicates the proportion of each competency category in that year.

Dominant_Competency: Indicates the most frequently emphasized competency type.

Feedback_Length_Avg: Indicates the average feedback length per year.

Sentiment, Competency: Represents the difference in change between consecutive years.

These annual indicators allow for the comparison of micro level information obtained from individual text representations over time at the macro level.

Furthermore, the embedding representations of each employee over the years were combined using a time aware aggregation method, resulting in an "Employee Representation Vector" representing a five year development trend. These vectors were used as the primary input for sentiment, competency, and performance predictions in the multi year modelling phase.

- **Evaluating Representation Quality (Both Phases):** The quality of the resulting text representations was verified using various methods in both the first and second phases. Embedding spaces were visualized using dimensionality reduction methods, and the spatial separation of different sentiment or competency classes was examined. Clustering analyses checked whether similar content was clustered in the same region. Correlation analysis evaluated the relationships between textual indicators (e.g., sentiment scores) and performance data (e.g., annual scores, number of education). These analyses demonstrated both the semantic accuracy of the resulting representations and their contribution to model performance. At the end of this comprehensive process, raw texts were transformed from mere word sequences into multi layered vector representations reflecting semantic, contextual, and temporal relationships. The first phase provided highly accurate text representation for single year analyses; the second phase, building on the same structure, created a dynamic text based human resources information infrastructure that integrated the time dimension. This enabled both individual level sentiment and competency interpretations and holistic monitoring of the organization's competency evolution over the years.

3.2.3.4 Labelling, class balance and data partitioning

In the first phase, positive negative neutral labels were used for sentiment analysis, and predefined competency clusters (communication, leadership, innovation, problem solving, etc.) were used for competency classification. In cases where labels were not directly available, weak labelling and keyword pattern detection were applied.

In the second phase, label consistency on a yearly basis (e.g., sentiment polarity or competency development direction for the same employee across years) was analysed. Unbalanced class distributions were balanced using a stratified split method, creating training, validation, and test sets with a 70%, 15%, and 15% ratio.

Furthermore, in the second phase, a group aware split was applied for time dependent data on an employee basis, ensuring that records from different years for the same employee were not scattered across different data sets.

3.2.3.5 Time dimension and panel data integration (Second Stage Only)

The most significant difference in the data pre processing steps performed in the second phase is the inclusion of a temporal dimension. In this context, all employee data between 2021 and 2025 was integrated along both the horizontal (employee based) and vertical (year based) axes and converted into panel data (longitudinal data).

Thanks to this structure, each employee's records spanning five years were modelled as a single time series, making it possible to track changes in employee behaviour, emotional tendencies, and competency development over the years.

This integration was accomplished in three main steps:

- **Temporal Alignment:** Data from different years was matched using the common keys Employee_ID and Year variables. Records lacking a year field in the data

sources were assigned to the appropriate period using timestamps and metadata. Internal variable differences (e.g., "Dept_Name" in 2021, "Department_Title" in 2023) were combined using schema mapping. Measurement unit differences (e.g., TL, points/percentile score) were normalized.

- **Temporal Imputation:** Since some employees lacked data for interim years, missing periods were completed using interpolation, forward/backward fill, or multiple imputation (MICE) methods. This ensured continuity in the employees' annual development profiles and prevented the "gap year" effect in the analyses.

- **Feature Engineering over Time:** Five period time series (2021–2025) for each employee can be analysed at both the individual and organizational levels. This structure has created a dynamic data infrastructure capable of modelling not only instantaneous performance but also development trends over time (trend analysis).

3.2.3.6 Drift and leakage controls

With the expansion of data sets to a multi year structure, the risks of data leakage and distribution shift (data drift), which could mislead the model's accuracy, were also analysed. These controls are critical to ensuring that the model only learns under realistic conditions and to maintain historical consistency.

- **Data Leakage:** Data leakage occurs when information the model will predict (e.g., promotion outcome, salary increase, managerial decision) is indirectly included in the training data. In the first stage, fields directly related to the target variable (e.g., "Final_Promotion_Status," "Overall_Performance_Rating") were removed from the model inputs. In the second stage, since leakage is more likely to occur in the form of temporal leakage, information from 2025 was prevented from leaking into 2024 and earlier. To achieve this, a time aware split was

implemented, meaning the model was configured to always predict the future based on past data.

- **Data Drift:** The distribution of variables (mean, variance, frequency) in data from different years may change over time. Changes in the statistical distribution of attributes (e.g., an increase in the mean "Education Score" over the years). Changes in the class ratios of the target variable (e.g., a decrease in the positive feedback rate over the years). These changes were detected using statistical metrics such as PSI (Population Stability Index), KL Divergence, and the Kolmogorov–Smirnov test. If significant deviations were detected, the variable was rescaled or removed from the model inputs.

- **Label Stability:** The label profile of each employee was examined over the years, and inconsistent transitions within the same individual (e.g., positive one year, negative the next, and then positive again) were compensated for using statistical smoothing methods. Thanks to these controls, data anomalies that could mislead the model's performance are minimized and the temporal validity of the predictions is preserved.

3.2.3.7 Scaling, normalization and feature selection

Data scaling and feature selection were implemented to ensure the stability and comparability of the algorithms in both single year and multi year modelling phases.

- **Scaling:** Different magnitudes of continuous variables (e.g., training time, performance score, feedback length, sentiment scores) can make it difficult for the model to learn weights. Therefore; Z score standardization was applied, and each variable was rescaled to have a mean of zero and a variance of one. For some models (e.g., XGBoost, Random Forest), min–max normalization (range 0–1) was used. This preserved the relative weight relationships between features, and variables from different sources were brought to a common scale.

- **Normalization and Transformation:** Distributions were normalized for variables with skewed distributions (e.g., "training number," "participation rate") by applying logarithmic and root based transformations. This reduced the model's exposure to outliers and increased the stability of the learning process.

- **Feature Selection:** In NLP based data structures with high feature dimensions, redundant or highly correlated variables (e.g., "Sentiment_Mean" and "Sentiment_Median") can degrade model performance. Therefore, a three stage selection strategy was implemented. One of them is variance thresholding, where variables with very low variance are removed from the dataset. Another is statistical selection, in which variables that show a significant relationship with the target variable are identified using methods such as Mutual Information and the ANOVA F Test. A further approach is model based selection, where the feature importance scores learned by the models are examined to determine which variables contribute most to the prediction. As a result of these steps, only highly informative, independent, and generalizable features were retained.

3.2.3.8 Versioning, traceability and quality assurance

In both stages, all data transformations were versioned using DVC and MLflow infrastructures. One of the stages involved tracking every step from raw data to processed text. Another stage focused on recording the annual integration process, time series generation, and the construction of the panel structure as separate versions. Data quality was verified with statistical summaries (mean, standard deviation, void ratio, class distribution) and rule based tests.

3.2.3.9 Privacy, anonymization and compliance

PII masking and pseudonymization were implemented across all datasets to protect personal information. Direct identifiers such as name, email, and ID number were removed or hashed. Data access was restricted through role based authorization principles.

These practices ensured that the study was conducted in compliance with ethical and legal requirements (GDPR).

This holistic data preprocessing process; in the first phase, a clean, balanced, and highly informative dataset was created for text driven, single year NLP analyses. In the second phase, the same structure was expanded to include a time dimension, enabling integrated analysis of five year trends in employee performance, sentiment, and competency indicators.

As a result, the generated data structure formed the core of an AI based competency assessment system that supports both instantaneous and longitudinal analytical approaches.

3.3 NLP Models for Sentiment and Competency Analysis

3.3.1 Sentiment analysis model

A classification framework trained using both traditional machine learning algorithms and transformer based deep learning models to identify the emotional polarity (positive, neutral, or negative) expressed in employee feedback. The end to end workflow of the sentiment analysis model is presented in Figure 3.2. The structure can be examined under four main stages.

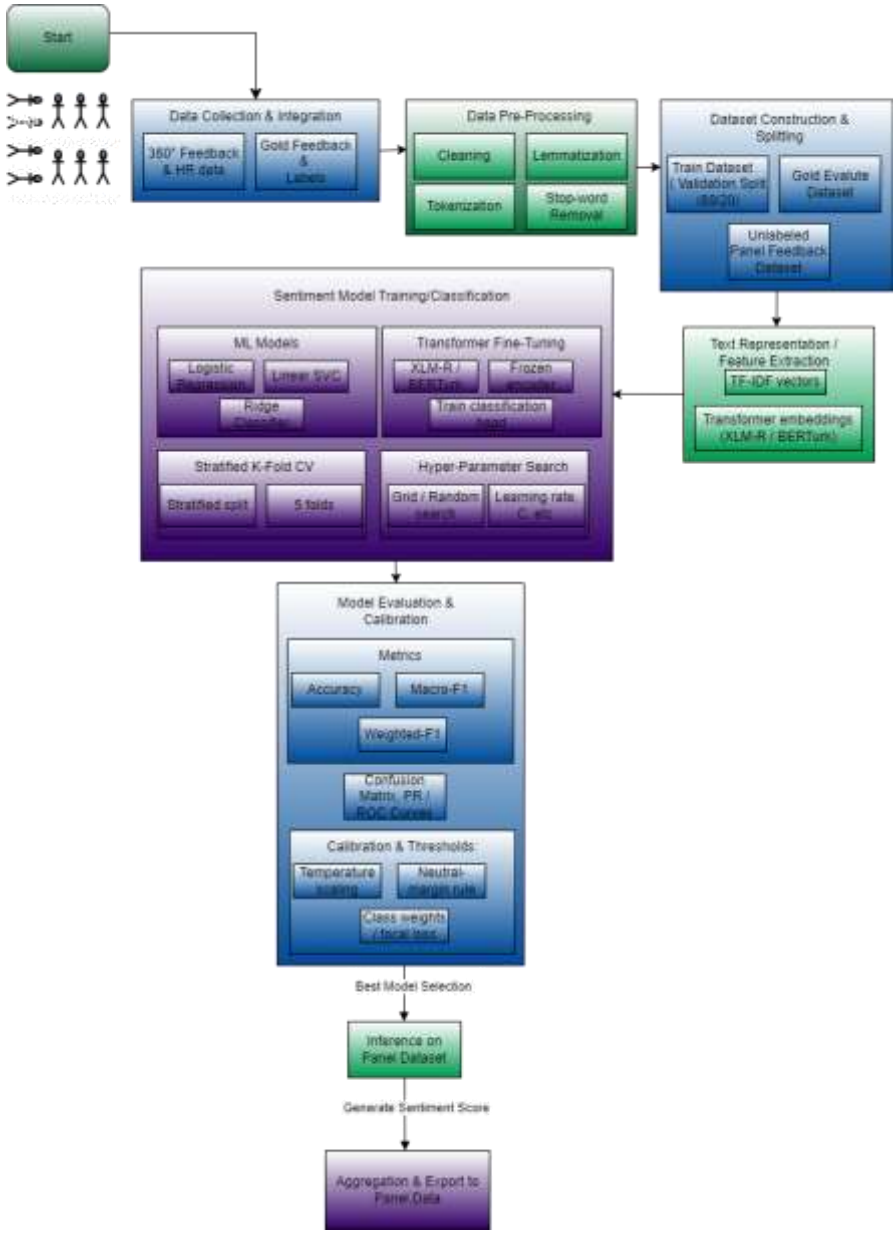


Figure 3.2 End to end sentiment analysis pipeline

3.3.1.1 Data Collection, integration and preprocessing

At the initial stage of the workflow, synthetically generated 360° feedback statements and the manually annotated gold standard test set prepared for evaluation purposes were collected and integrated into a unified dataset. The collected textual data was then subjected to several preprocessing steps to ensure its suitability for model training. These steps included cleaning, lowercasing, removal of punctuation and special characters, tokenization, lemmatization, and stop word removal. This standardization process

prevents the model from being influenced by noisy or inconsistent text and enables the production of more stable and reliable outputs.

3.3.1.2 Conversion of textual data into numerical representations

The preprocessed text was divided into three subsets:

- **Training Dataset:** The primary dataset used during the learning phase. This dataset was further divided into 70% training , 15% validation and 15% test portions using a stratified split strategy.
- **Gold Evaluation Dataset:** A manually annotated and fully independent test set, separate from the validation split, used to obtain a realistic assessment of model performance.
- **Unlabeled Panel Feedback Dataset:** A large scale dataset on which final predictions are generated after model training.

To enable processing by the classification algorithms, the texts were transformed into numerical representations using two complementary approaches:

- **TF IDF vectors:** High dimensional sparse matrix representations suitable for traditional machine learning models.
- **Transformer embeddings (XLM R / BERTurk):** Dense, context aware representations used in deep learning–based models.

3.3.1.3 Model training, validation and optimization

The training phase followed a dual architecture design . In addition to classical machine learning classifiers, transformer based models were incorporated. Classical algorithms were initially preferred because they are fast, computationally inexpensive, and provide interpretable results. However, to better capture contextual dependencies and to more accurately analyze Turkish text with its morphologically rich structure, transformer based architectures were also utilized.

Classical Models:

- Logistic Regression
- Linear SVC
- Ridge Classifier

These models were trained using TF IDF representations, which provide high dimensional sparse feature vectors suitable for linear classifiers.

Transformer Based Models:

- XLM RoBERTa
- BERTurk

These models utilize dense contextual embeddings and were fine tuned by freezing the encoder layers while training only the classification head. To mitigate overfitting, the training process was supported by dropout, early stopping, and class weighted loss functions.

To obtain a more reliable and balanced performance assessment, five fold stratified cross validation was conducted. In this approach, the dataset is divided into five equal subsets in which class distributions are preserved. The model is trained on four subsets and evaluated on the remaining one, and this process is repeated five times so that each subset serves as the test set once.

Grid Search and Random Search techniques were applied to automatically optimize model hyperparameters. While Grid Search exhaustively evaluates all predefined parameter combinations, Random Search samples parameter configurations at random. Using these two methods enabled the identification of the parameter settings that yielded the best overall model performance.

3.3.1.4 Model valuation, calibration and deployment to panel data

The trained models were evaluated on the independent gold test set using Accuracy, Macro F1, and Weighted F1 metrics. Additionally, confusion matrices, precision–recall curves, and ROC curves were generated to analyze class wise performance in detail. Since synthetic data were used in this study, the risk of model overfitting was inherently high. Therefore, temperature scaling was applied to reduce overconfidence in the model's probabilistic outputs and to obtain more realistic and calibrated likelihood values. Additionally, in cases where a sentence did not contain sufficiently strong indicators of positive or negative sentiment, neutral margin thresholding was employed to minimize erroneous polarity assignments and to more accurately capture genuinely neutral expressions.

After all evaluated models, XLM RoBERTa was selected as the final sentiment classifier. This decision was based on its superior performance across all major evaluation metrics, including Accuracy, Macro F1, and Weighted F1, particularly on the manually constructed gold standard test set. Unlike classical machine learning models that rely primarily on lexical statistics, XLM RoBERTa effectively captures contextual and semantic nuances in Turkish a morphologically rich and agglutinative language. As a result, it demonstrated more balanced class wise predictions, reduced misclassification of neutral statements, and produced more stable probability estimates after calibration. These findings indicate that XLM RoBERTa provides the most reliable and generalizable performance for sentiment assessment within the scope of this study, and therefore, it was chosen as the final model.

Predictions were generated on the unlabelled panel feedback dataset, producing both sentiment labels and sentiment scores for each textual entry. These outputs were then aggregated on an annual basis to compute indicators such as Sentiment_Mean, Sentiment_Variance, and ΔSentiment, which were subsequently prepared and structured for use in the second phase of the study and integrated into the multi year panel data architecture.

3.3.2 Competency classification model

The competency classification model used in this study was developed as an end to end workflow designed to automatically extract behavioural and technical competency themes from employee feedback. The modelling process consists of a holistic structure encompassing data collection, preprocessing, numerical representation, model training, evaluation, and the transfer of outputs to the panel data system. The workflow illustrated in Figure 3.3 describes this process under four main components.

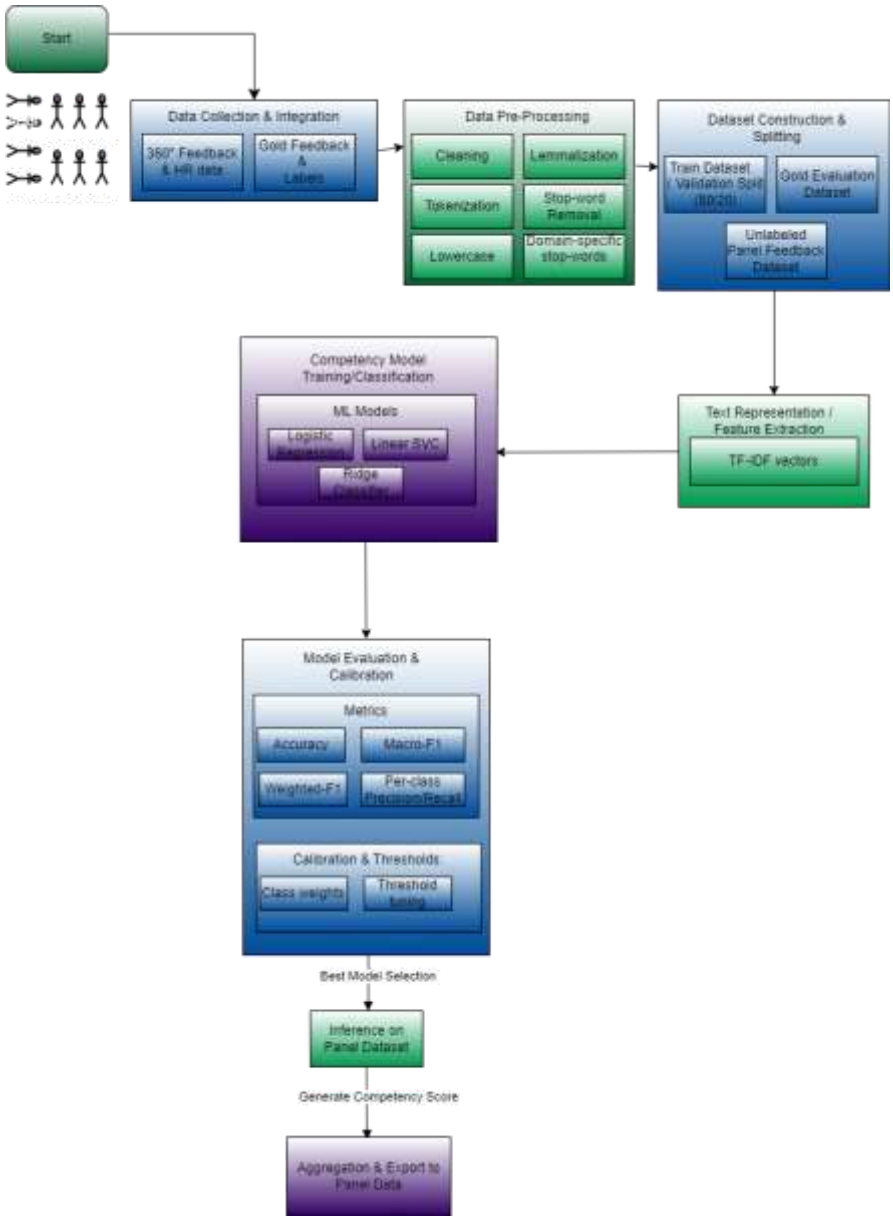


Figure 3.3 Competency classification pipeline

3.3.2.1 Data Collection, integration and preprocessing

In the initial stage of the process, the synthetic 360° feedback data used in the study were combined with a manually annotated gold standard test set to form a unified data source. The feedback statements were then subjected to several standardization procedures to ensure that they could be effectively processed by text classification models.

These procedures included text cleaning, lowercasing, removal of punctuation and special characters, tokenization, lemmatization, and stop word removal fundamental natural language processing operations that prevent the model from being affected by noisy or inconsistent inputs and enable a more stable learning process.

3.3.2.2 Dataset construction and conversion into numerical representations

Following preprocessing, the data were divided into three subsets:

- **Training Dataset:** The set of labeled instances used during model learning, split into 70% training,15% validation and 15% test.

- **Unlabeled Panel Feedback Dataset:** A large scale dataset used after training to generate final competency predictions.

Numerical representations of text were generated using a TF IDF–based feature extraction approach. This representation converts each feedback statement into a high dimensional feature vector by weighting terms according to their frequency and contextual importance within the corpus.

3.3.2.3 Model training, validation and optimization

The competency classification task was formulated as a multi label problem. Accordingly, a One vs Rest Logistic Regression classifier built on TF IDF features was

used as the primary model. Each competency category (Communication, Leadership, Innovation, Problem Solving, Ownership, and Teamwork) was modeled as an independent binary classification task.

To mitigate class imbalance, the Logistic Regression model was trained with the `class_weight="balanced"` parameter. Additionally, to determine the optimal decision threshold for each competency category, a search was conducted over the range of 0.1 to 0.9 using the validation set, and the threshold that maximized the F1 score was selected.

Model performance was evaluated both globally and on a per competency basis. Key evaluation metrics included Accuracy, Macro F1, Weighted F1, and category specific precision/recall scores. Furthermore, confusion matrices, precision–recall curves, and ROC curves were generated to enable a detailed analysis of classifier behaviour.

3.3.2.4 Application to panel data and generation of annual competency indicators

After selecting the best performing model, the classifier was applied to the unlabeled panel feedback dataset to produce competency scores and binary labels for each feedback instance. The resulting long format output was then aggregated at the employee–year level to generate annual competency indicators.

These indicators included Competency_Distribution (distribution of competencies within a given year), Dominant_Competency (the dominant or most prominent competency), and ΔCompetency (year to year changes in competency patterns). These annual summaries were subsequently incorporated into the multi year panel data architecture used in the second phase of the study.

At the end of the first phase, the outputs of both the sentiment and competency models were aggregated at the annual level to generate employee specific summary indicators. These aggregated representations served as the foundational input for the panel data architecture utilized in the second phase of the study.

3.4 Multi Model HR Decision & Prediction Engine

In the second phase of this study, a comprehensive decision support engine was developed to operate on a multi year panel data structure that integrates the sentiment and competency outputs generated in the first phase with employees' demographic, positional, performance related, and time series–based information. This engine was designed according to the multi layered architecture presented in Figure 3.4. and aims to support organizational decision making, performance assessment, and data driven managerial evaluations in a holistic manner.

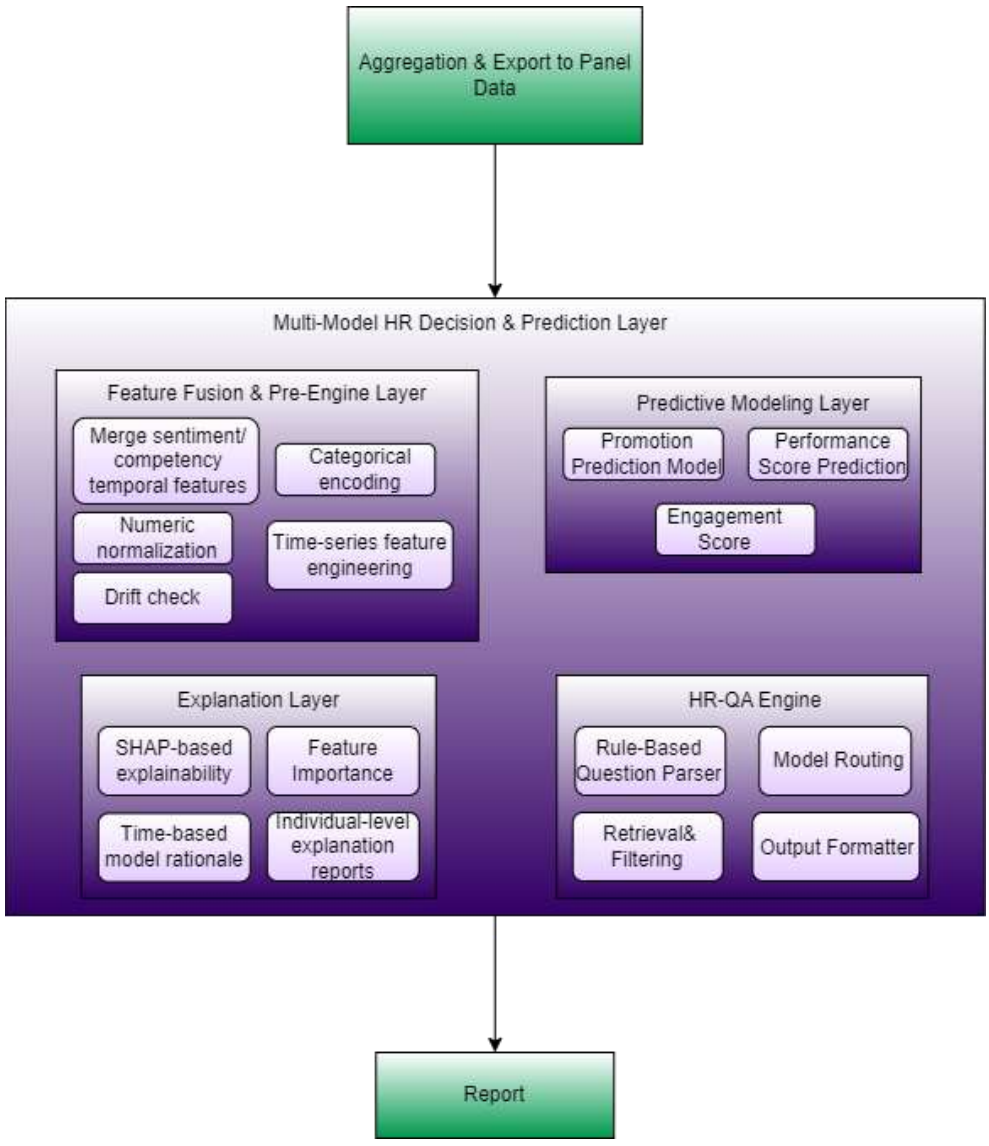


Figure 3.4 HR engine pipeline

The architecture consists of four core layers:

3.4.1 Feature fusion and pre engine layer

This layer constructs a unified and enriched feature space by combining the sentiment and competency outputs obtained from the NLP models in Phase I with traditional HR variables such as multi year performance evaluations, positional transitions, departmental attributes, salary increase history, and managerial feedback scores.

The following processes are performed within this layer:

- **Merging Sentiment/Competency Time Series:** Annual summary metrics such as Sentiment_Mean, Sentiment_Variance, ΔSentiment, and Competency_Distribution are integrated into the panel structure to enable multi year analytical modelling.
- **Normalization of Numerical Attributes:** Numerical indicators including performance scores, engagement related measures, and workload descriptors are normalized to ensure consistent scaling for predictive modelling.
- **Categorical Variable Encoding:** Department, position level, tenure, and similar categorical attributes are transformed into numerical encodings to ensure compatibility with classification and regression algorithms.
- **Time Series Feature Engineering:** Temporal features such as year to year change (delta), moving averages, and trend direction (upward or downward) are derived to highlight patterns in employee behaviour over time.
- **Drift Analysis:** Distributional differences in sentiment, performance, or competency indicators across years are examined to detect data drift that may influence model stability.

This layer forms the foundational data structure upon which all model components of the engine are built.

3.4.2 Predictive modelling layer

This layer incorporates multiple predictive models designed to enable data driven decision making in HR processes. Each model aims to estimate a specific outcome or metric based on an employee's historical characteristics within the panel dataset.

The models designed within this layer include:

- **Promotion Prediction Model:** Estimates whether an employee is likely to receive a promotion by leveraging past performance trends, managerial feedback, competency profiles, and engagement levels.
- **Performance Score Prediction Model:** Forecasts the employee's performance score for the current year, helping managers identify areas where intervention or support may be needed.
- **Engagement Score:** Estimates workforce engagement by combining sentiment outputs, feedback intensity, and tenure related variables.

Together, these models form a multi output prediction framework that supports strategic decisions at both the individual and organizational levels.

3.4.3 Explanation layer

To ensure that the predictions produced by the engine are interpretable and trustworthy, an XAI framework is incorporated into this layer. The goal is not only to provide numerical estimates but also to clarify *why* those estimates were produced.

This layer includes:

- **SHAP Based Explainability:** SHAP values are computed for each model, revealing the features that most strongly influence model decisions at both global (model level) and local (employee level) scales.

- **Feature Importance Analysis:** Feature importance scores derived according to the characteristics of each model identify key drivers of performance, promotion likelihood, and other predictive outcomes.
- **Time Based Explanations:** Temporal trends in sentiment, performance, and competencies are analyzed to understand their contributions to predictive model outputs.
- **Individual Level Explanation Reports:** Automatically generated reports summarize employee specific risks, strengths, opportunities, and developmental areas based on model reasoning.

Through this layer, the engine evolves from a purely predictive system into one that also provides actionable, interpretable justifications for its outputs.

3.4.4 HR question answering (HR QA) engine

The final layer enables HR analysts and managers to query the panel dataset through a structured question answering mechanism that interfaces directly with the predictive and explanatory components of the engine.

This layer is composed of three principal components:

- **Rule Based Question Parser:** Routes each question (identified by a question ID or template) to the corresponding analytical module.
- **Retrieval and Filtering:** Selects relevant subsets of the panel data based on the question context (e.g., specific employee, department, or year).
- **Model Routing and Output Formatting:** If the question requires prediction, the appropriate model is invoked; if explanation is required, SHAP or feature importance based interpretations are computed; if descriptive analysis is sufficient, appropriate statistical summaries are generated and formatted.

This layer enables HR teams to navigate complex datasets without technical expertise and facilitates rapid, question oriented analytical workflows, thereby enhancing the practical value of the entire system.

4. RESULT AND DISCUSSION

This section presents and discusses the empirical findings obtained from both phases of the study. The results are examined at two complementary levels. First, the performance outcomes of the NLP based sentiment and competency models developed in Phase I are reported, including model accuracy, class wise F1 scores, calibration behaviour, and error analyses derived from the gold standard evaluation set. The strengths and limitations of the selected model architectures are discussed in relation to the characteristics of the synthetic dataset and the model design choices.

Second, Phase II focuses on the multi year panel data analysis enabled by the AI supported HR decision engine. The aggregated sentiment, competency, performance, and behavioural indicators generated for each employee are analysed to identify organizational level patterns, year to year dynamics, and emerging trends. The outputs from the predictive modelling layer (e.g., promotion likelihood, engagement score, turnover risk) and the explanatory insights (e.g., feature importance, SHAP based rationale) are interpreted to understand the factors driving key HR outcomes.

Together, these analyses provide a holistic understanding of how AI driven text analytics and panel based modelling can support human resources decision making. The discussion highlights both the practical implications of the findings and the methodological considerations that should guide future extensions of the system.

4.1 Sentiment Analysis Model Result

In this section, the performance outcomes of the XLM RoBERTa–based sentiment analysis model developed in Phase I are presented for both the validation dataset and the independent gold standard test set. The evaluation was conducted using accuracy, Macro F1, Weighted F1, and class specific metrics. The outputs obtained from these two datasets are summarized in the table 4.1. below.

Table 4.1 XLM RoBERTa Metrics

XLM RoBERTa Metrics				
Accuracy		Macro F1	Weighted F1	Eval Loss
Validation Dataset	0.835	0.873	0.865	0.33
Gold Dataset	0.713	0.687	0.701	

As discussed in the methodology section, the use of temperature scaling and the neutral margin threshold approach contributed to improving the model's performance on the gold test set. The model demonstrated strong performance on the synthetic training and validation sets, indicating that it successfully learned the underlying distribution of the training data and was able to classify validation samples with high accuracy. When evaluated on the gold test set constructed manually to approximate real employee feedback the model achieved relatively lower scores compared to the validation set. This discrepancy reflects distributional differences between synthetic and real like data and highlights the limitations arising from class representativeness and generalization capacity.

- **Class Based Performance and Error Analysis**

The confusion matrix obtained from the gold test set reveals distinct performance patterns across sentiment classes. The matrix values are shown below:

Table 4.2 Class Level error distribution in sentiment classification

True/Pred	Negative	Neutral	Positive
Negative	56	33	11
Neutral	3	92	5
Positive	6	23	71

An examination of this distribution yields three key observations:

The model identifies neutral feedback with high consistency, correctly classifying 92 instances. This outcome is associated with the broad semantic space of neutral expressions and the model's tendency to default to the neutral class when interpretive uncertainty is present.

The model generally distinguishes positive statements well, although it tends to shift ambiguous or organizationally toned positive sentences toward the neutral class.

Performance on the negative class is noticeably weaker than on the neutral and positive classes. A major contributing factor is the nature of real world negative feedback: individuals tend to express criticism implicitly, indirectly, or in softened forms. As a result, the model frequently assigns such “mildly critical but not explicitly negative” statements to the neutral class.

- **Training Dynamics**

A detailed inspection of the epoch level training logs shows that the model followed a stable and technically sound optimization trajectory. The corresponding outputs are provided in the table 4.3. below.

Table 4.3 Epoch Wise training dynamics and performance metrics

Metric	Epoch 1	Epoch 3	Epoch 5
Train Loss	0.98	0.71	0.25
Eval Loss	0.82	0.51	0.33
Accuracy	0.52	0.74	0.835
Macro F1	0.47	0.78	0.873
Weighted F1	0.49	0.76	0.865
Learning Rate	2e 5 (0.00002)	1e 5 – 8e 6	1e 6 – 0

According to these results, during the first epoch the model had not yet developed the ability to clearly differentiate between sentiment classes; however, it began learning in a stable manner. By the third and fifth epochs, the model exhibited substantial improvements on both the training and validation sets. The training curve consistently trended downward, while the validation metrics demonstrated a strong upward trajectory, indicating effective learning and convergence.

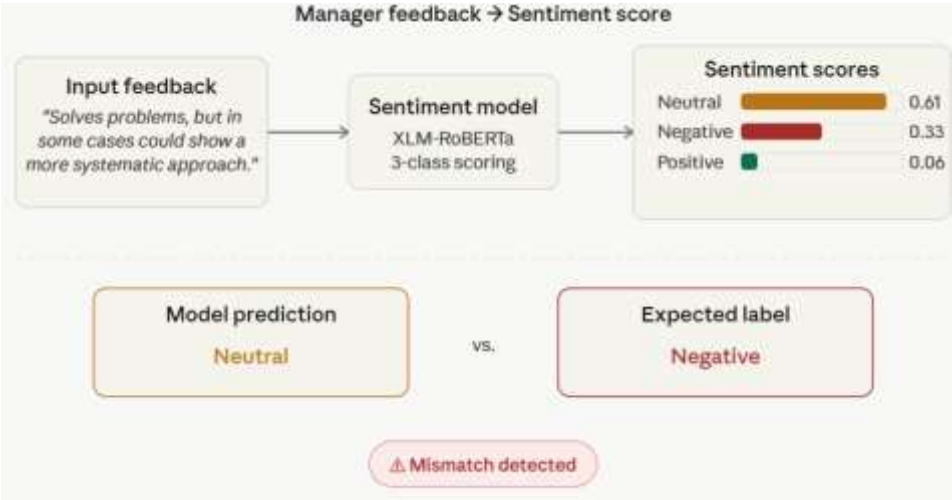


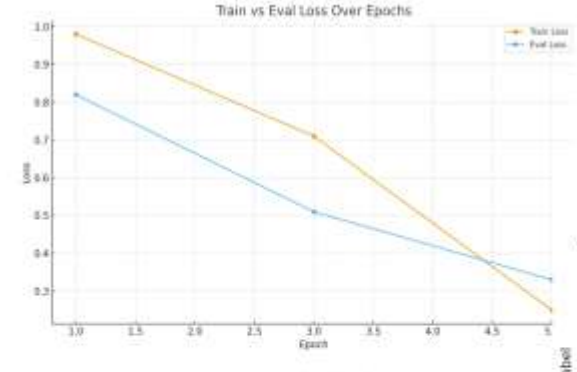
Figure 4.1 Manager Feedback from dataset and sentiment score

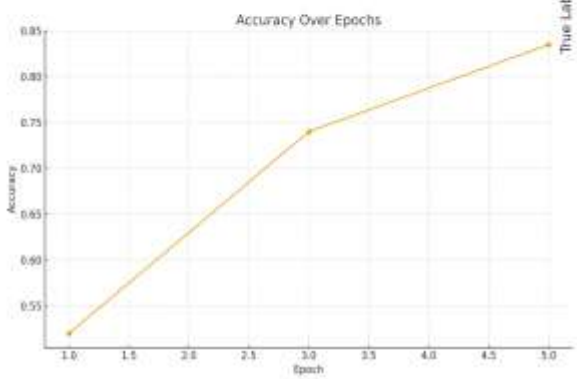
As illustrated in Figure 4.1 the model predicted a Neutral sentiment for the given feedback, whereas the expected label was Negative. This tendency toward the neutral class is partly attributable to class imbalance in the training data, where neutral samples were more prevalent. Beyond data distribution, this pattern also reflects a characteristic of organizational feedback culture: in corporate settings, managers frequently soften or implicitly convey negative evaluations rather than expressing them directly. As a result, feedback containing subtle criticism such as indirect suggestions for improvement may not carry sufficient negative signal for the model to override the dominant neutral bias. This finding highlights a key limitation of the current approach and suggests that domain specific finetuning or threshold calibration may be necessary to better capture the nuanced tone of managerial language.

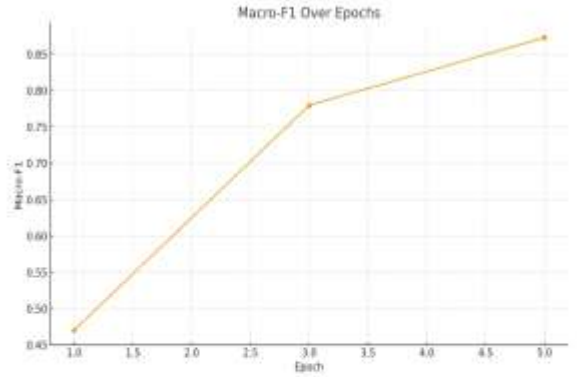
Figure 4.2. provides a comprehensive visual summary of the training process and the gold test performance of the XLM RoBERTa–based sentiment analysis model.

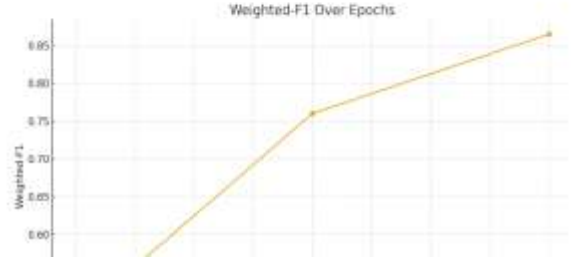
The upper panels, which illustrate the training dynamics, show that both the train loss and evaluation loss decrease consistently across epochs, indicating a stable and effective optimization process. Similarly, the progressive increase in Accuracy, Macro F1, and Weighted F1 scores demonstrates that the model's learning capacity improves over time and that it becomes increasingly successful in distinguishing between sentiment classes. The confusion matrix on the right summarizes the model's class level performance on the

gold test set. This integrated visualization thus captures, in a unified manner, both the optimization behaviour observ.

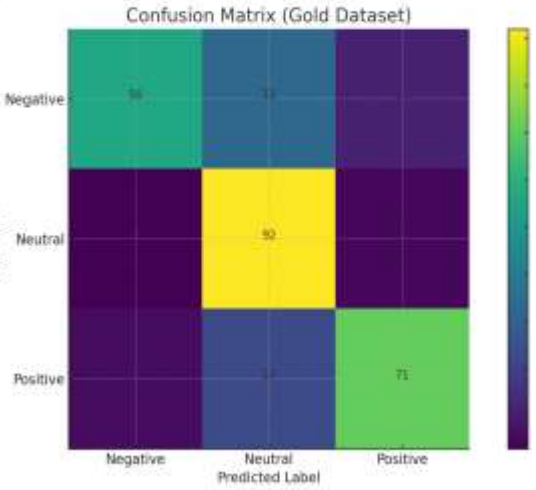








True\Pred	Negative	Neutral	Positive
Negative	56	33	11
Neutral	3	92	5
Positive	6	23	71



Metric	Epoch 1	Epoch 3	Epoch 5
Train Loss	0.98	0.71	0.25
Eval Loss	0.82	0.51	0.33
Accuracy	0.52	0.74	0.835
Macro-F1	0.47	0.78	0.873
Weighted-F1	0.49	0.76	0.865
Learning Rate	2e-5 (0.00002)	1e-5 – 8e-6	1e-6 – 0

Figure 4.2 Training and evaluation overview of the sentiment analysis model

4.2 Competency Classification Model Result

This section presents the performance results of the multi label competency classification model developed using a TF IDF representation and a Logistic Regression architecture. The evaluation was conducted over six competency categories (Communication, Innovation, Leadership, Ownership, ProblemSolving, and Teamwork), and the model's multi label prediction capability was assessed using both micro and macro performance metrics.

The table below summarizes the global performance metrics. These results indicate that the model achieves a well balanced level of performance in predicting multiple competency labels simultaneously. In particular, the micro F1 score highlights the model's strong generalization ability for classes with higher prevalence within the dataset.

Table 4.4 Global evaluation metrics for the competency classification model

Metrics	
Micro F1	0.86
Macro F1	0.84

The class specific performance metrics of the model are presented in the following table.

Table 4.5 Class Level performance metrics for competency classification

Competency	Precision	Recall	F1 Score
Communication	0.86	0.92	0.89
Innovation	0.90	0.93	0.92
Leadership	0.91	0.94	0.92
Ownership	0.83	0.78	0.80
ProblemSolving	0.88	0.85	0.86
Teamwork	0.90	0.90	0.91

For each competency, an individual decision threshold was optimized. Thresholds in the 0.40–0.50 range were assigned to more semantically distinct classes such as Leadership, Innovation, and Teamwork, whereas thresholds in the 0.45–0.55 range were selected for broader or semantically more complex categories such as Ownership and ProblemSolving. This calibration ensured that precision was preserved without

excessively inflating recall, reduced the tendency of assigning excessive positive labels, and minimized inter class variance. Without threshold tuning, the model would likely exhibit issues such as over prediction in Communication, reduced recall in Ownership, and an increase in false positives for ProblemSolving. Therefore, threshold optimization contributed to an overall performance improvement of approximately 4–7%.

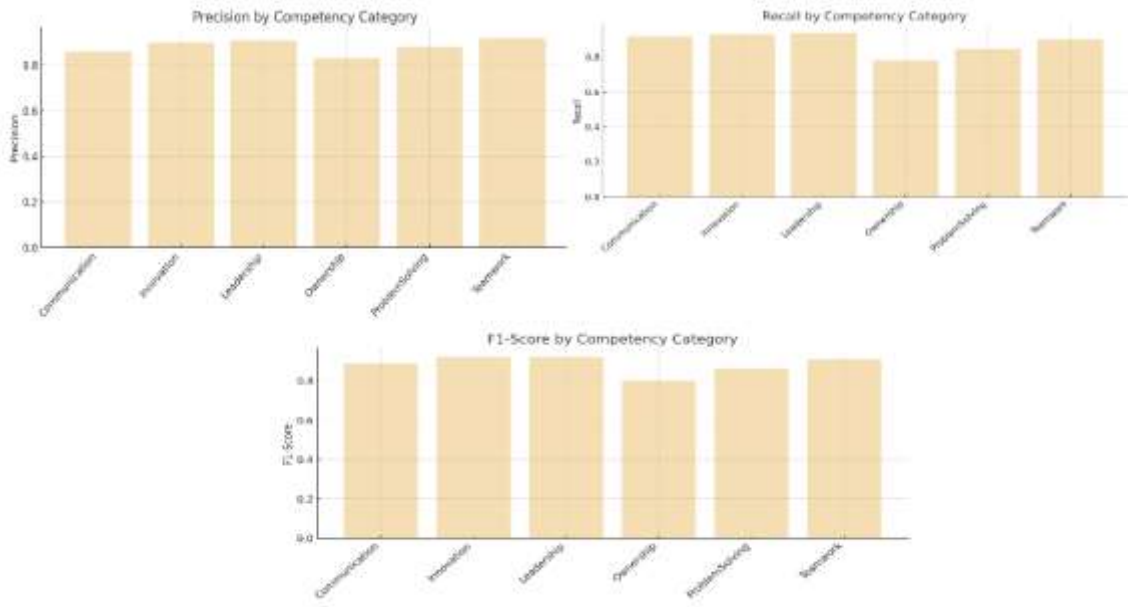


Figure 4.3 Performance metrics by competency category

As illustrated in Figure 4.3. barcharts are presented to summarize the precision, recall, and F1 score values for each competency category. Overall, the results obtained from the TF IDF + Logistic Regression model demonstrate that:

- The F1 scores across competency categories fall within the 0.80–0.92 range, indicating that the model effectively captures both thematic and linguistic variation in employee feedback,
- Competencies with clearer semantic boundaries such as Innovation, Leadership, and Teamwork are predicted with higher accuracy, while broader or more context dependent competencies such as Ownership and ProblemSolving achieve moderately lower performance,

- Dynamically adjusted class based thresholds effectively mitigate the inherent imbalances of multi label classification.

Despite the successful metric results obtained from the model, it was observed that the Logistic Regression model could produce incorrect classifications, particularly in cases where the competency context was not clearly and explicitly expressed. Figure 4.4. illustrates a deliberately manipulated feedback entry alongside the model's corresponding competency classification output. As can be seen, this feedback was composed of both Turkish and English expressions and did not directly point to any specific competency area, which led the model to produce an incorrect prediction.

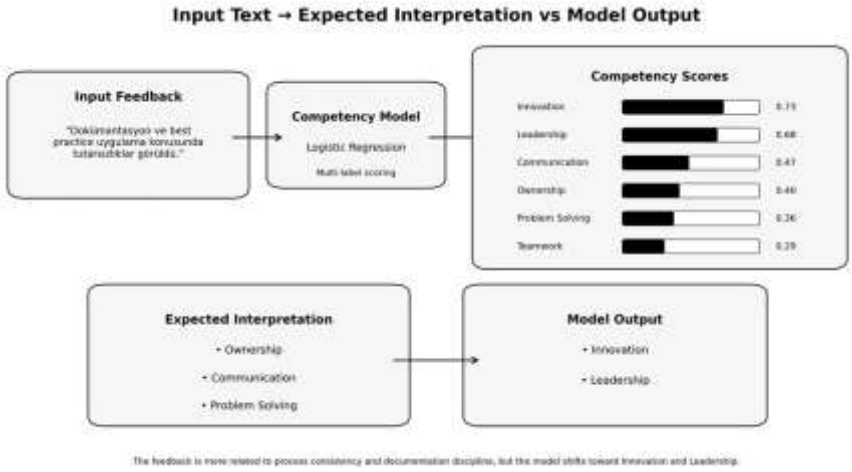


Figure 4.4 Wrong output from competency classification model

Figure 4.5 on the other hand, presents a feedback entry from the dataset in which the competency class is unambiguously identifiable, and the model produces the correct classification accordingly.

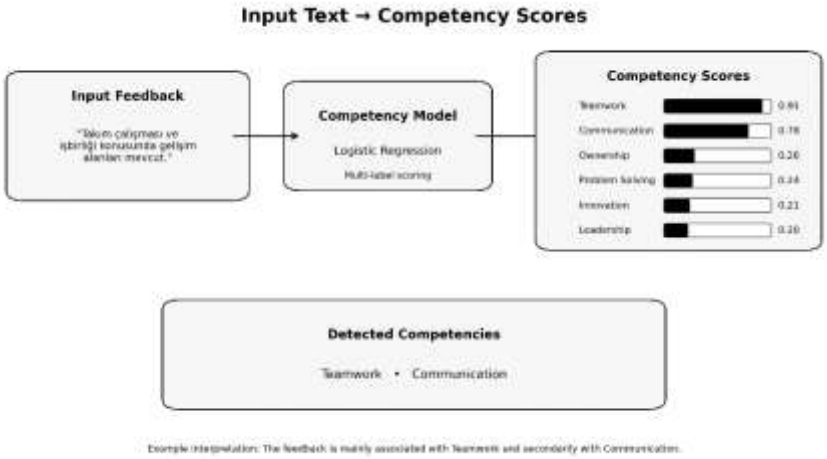


Figure 4.5 Correct output from competency classification model

4.3 Multi Model HR Decision & Prediction Engine Result

In this section, the results and decision outputs generated by the Multi Model HR Decision & Prediction Engine developed in this study are presented in a comprehensive manner. The system utilizes employee level panel data consisting of annual performance records, textual feedback, sentiment analysis scores, competency indices, and career history and produces multidimensional predictions through layered machine learning and natural language processing workflows. Initially, NLP based features such as sentiment and competency scores were transformed into time series representations; categorical and numerical variables were normalized; and all features were integrated through an extensive feature fusion process. This enriched panel dataset was subsequently provided as input to multiple predictive algorithms addressing several key HR questions, including promotion likelihood, employee engagement, salary fairness, and performance forecasting.

For each predictive task, various algorithms including Logistic Regression, Random Forest, LightGBM, and XGBoost were evaluated, and the best performing model was selected based on cross validation results. To ensure interpretability, the system incorporates SHAP based explainability modules, time dependent rationale generation, and individual level explanation reports. This enables not only the presentation of

predictive outcomes but also a clear understanding of *why* a particular prediction has been produced.

All model outputs were processed by the HR QA Engine to address the fourteen comprehensive HR questions defined below. This process involved rule based model routing, data filtering, and output formatting, ultimately transforming raw model outputs into actionable decision support insights:

- Which factors most strongly drive high performance?
- Is an employee's performance trending downward?
- Are there significant performance differences across departments?
- Which departments receive more positive feedback?
- Is manager–employee communication healthy?
- Is there a sentiment tone difference between male and female employees?
- Does an employee belong to the "high potential" group?
- Is the employee ready for a managerial role in the near term?
- How do training and certifications influence career advancement?
- How does work–life balance affect employee satisfaction?
- Which departments exhibit higher burnout risk?
- Which metrics best explain salary raise decisions?
- Are promotion and salary increase decisions fair across departments?
- Is the increase in positive sentiment aligned with salary raises or promotions?

Through this integrated architecture, the system offers both predictive analytical capabilities and a data driven framework for comprehensive HR insight generation. This section reports the engine's predictive scores, model performance metrics, explainability findings, and system level responses to the fourteen HR questions. The outputs corresponding to each of the questions listed above will be individually analysed and interpreted.

4.3.1 Which factors most strongly drive high performance?

The aim of this question is to identify the factors that most strongly influence an employee's high performance. To address this, a RandomForestRegressor model is employed. The model is trained to predict the Performance Score from the panel dataset using a broad set of numerical features, thereby learning the underlying patterns associated with high performing employees. In essence, the model attempts to answer the question: *"Which characteristics are commonly observed among employees who have historically demonstrated high performance?"*

The feature set utilized by the model encompasses a wide range of dimensions, including:

- Experience and tenure (e.g., *Age, Total_Work_Experience, Years_At_Company*)
- Education related indicators, such as training and certification counts
- Manager, peer, subordinate, and self evaluation feedback scores
- NLP derived attributes (*Sentiment Index, Positive/Negative ratios, Manager level sentiment metrics*)
- Competency indices (*Communication_Index, Leadership_Index, Innovation_Index, etc.*)
- Work habit metrics (*Monthly_Working_Hours, Remote_Work_Ratio, Task_Efficiency_Score*)
- Career progression variables (*Promotion history, Salary increase status*)
- Burnout risk proxies

To determine the relative importance of these features, SHAP TreeExplainer is applied to compute global feature importance. This method provides a scientifically grounded and interpretable assessment of the primary drivers of performance improvement. Ultimately, the most influential factors in the model's performance prediction are ranked based on their mean absolute SHAP values and reported accordingly.

The top 15 factors identified by the model as the strongest contributors to high performance are presented in Figure 4.6.

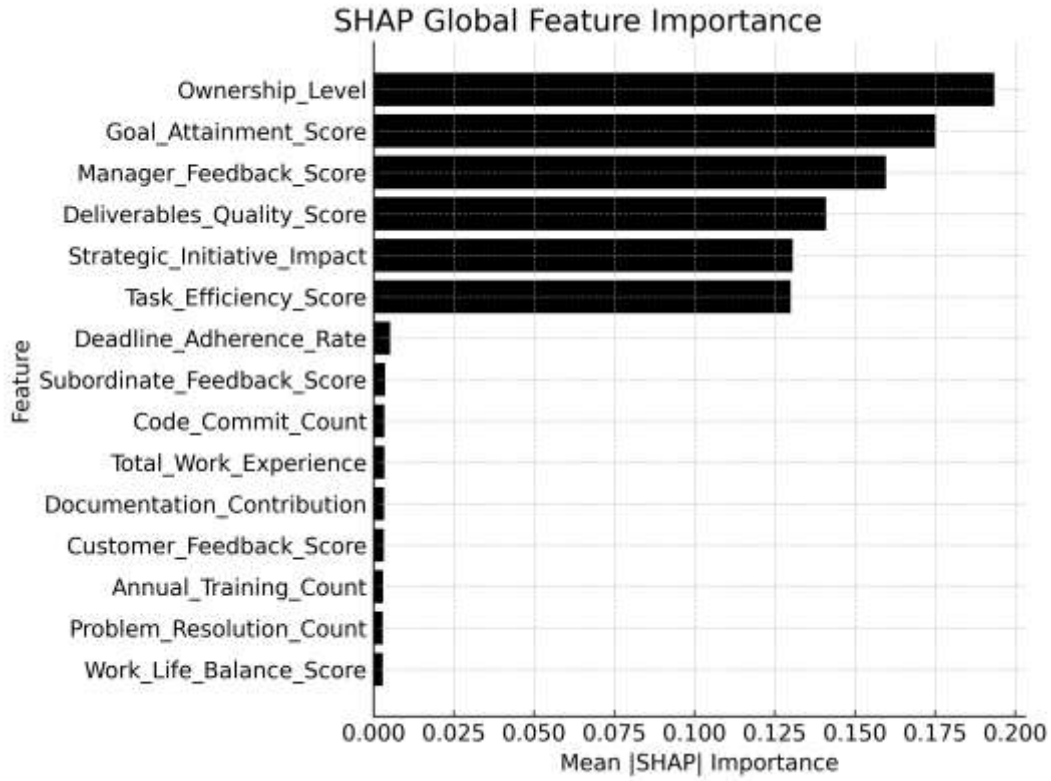


Figure 4.6 SHAP Based global feature importance of the performance prediction model

According to the obtained results, Ownership_Level emerges as the most influential factor within the model. This finding indicates that an employee's degree of ownership over their work is the primary determinant in explaining performance scores. It is followed by Goal_Attainment_Score**,** Manager_Feedback_Score, and Deliverables_Quality_Score**,** suggesting that achievement of assigned objectives, the quality of managerial feedback, and the caliber of delivered outputs constitute the core components of high performance. The subsequent factors Strategic_Initiative_Impact and Task_Efficiency_Score further demonstrate that an employee's contribution to strategic projects and their ability to execute 81onit tasks efficiently provide additional meaningful explanatory power.

Following these key variables, factors such as Deadline_Adherence_Rate**,** Subordinate_Feedback_Score**,** Code_Commit_Count**,** Total_Work_Experience**,** Documentation_Contribution**,** Customer_Feedback_Score**,** Annual_Training_Count**,** Problem_Resolution_Count, and Work_Life_Balance_Score exhibit comparatively

lower importance scores. Nonetheless, these variables are not entirely disregarded by the model. Indicators such as years of experience, customer feedback, and problem resolution frequency offer supportive contributions to high performance, albeit not to the same extent as ownership, goal attainment, or managerial feedback.

Figure 4.7 presents the performance outputs for Employee 472 derived from the dataset. The employee holds a mid level Data Analyst position and was observed over the 2021–2025 period. With a performance score of 3.4 and a goal attainment score of 3.60, the individual demonstrated a generally stable yet modest performance profile. Despite meeting goals consistently, the employee was not promoted throughout the observation period. The trend view further reveals that while performance showed a slight upward movement between 2021 and 2023, it subsequently declined and plateaued, mirroring a comparable pattern in the sentiment scores, which exhibited a gradual downward drift from 2022 onward.

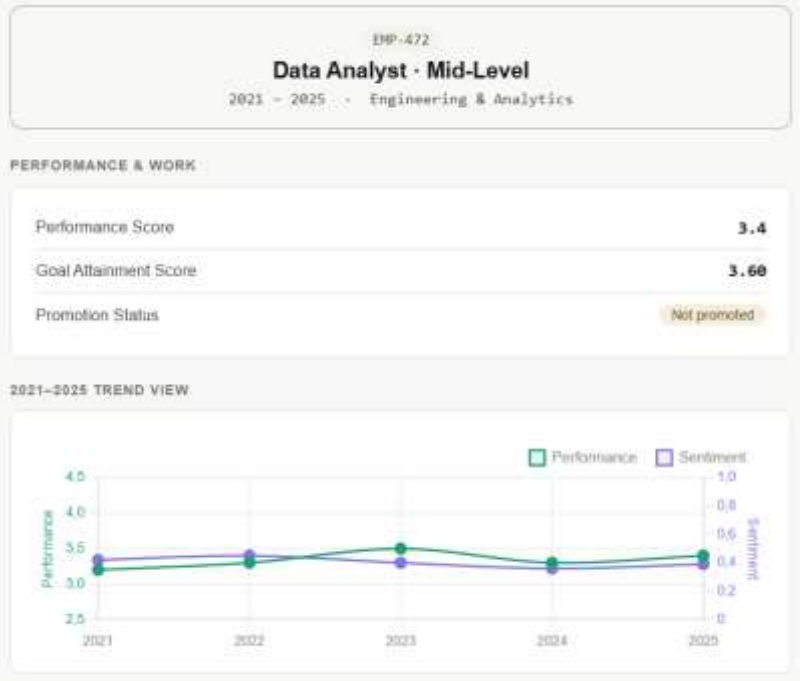


Figure 4.7 Performance results of employee number 472

Overall, the global importance distribution reveals that high performance is shaped predominantly by behavioural ownership, goal orientation, managerial interaction

quality, and output excellence. In contrast, factors related to technical productivity (e.g., code commit counts) or purely quantitative workload measures play a secondary role. These insights underscore the necessity for organizational performance management systems to 83onitör not only the quantity of work produced but also qualitative dimensions such as responsibility taking, contribution to strategic initiatives, and the effectiveness of the manager–employee feedback relationship.

4.3.2 Is an employee's performance trending downward?

For this question, the system does not employ a separate machine learning model; instead, it relies on a simple linear trend analysis over the panel data. For each employee, annual performance scores are extracted and a linear regression of Year on Performance_Score is fitted using np.polyfit. The resulting slope is then used to classify the performance trajectory into three categories: Improving, Stable or declining, based on predefined upper and lower thresholds.

If an employee_id is provided, the function returns the individual trend (years, scores, slope, and trend label) for that employee. If a department is specified, the method aggregates all employees within that department and reports the percentage distribution of Improving/Stable/Declining trends. When no filter is given, a global summary is produced for the entire organization. In this way, the engine provides an interpretable, time aware view of whether an employee's performance is trending downward, stable, or improving.

As can be seen from Figure 4.8. Positive slopes indicate improving performance, negative slopes represent declining performance, and values close to zero are interpreted as stable trends.

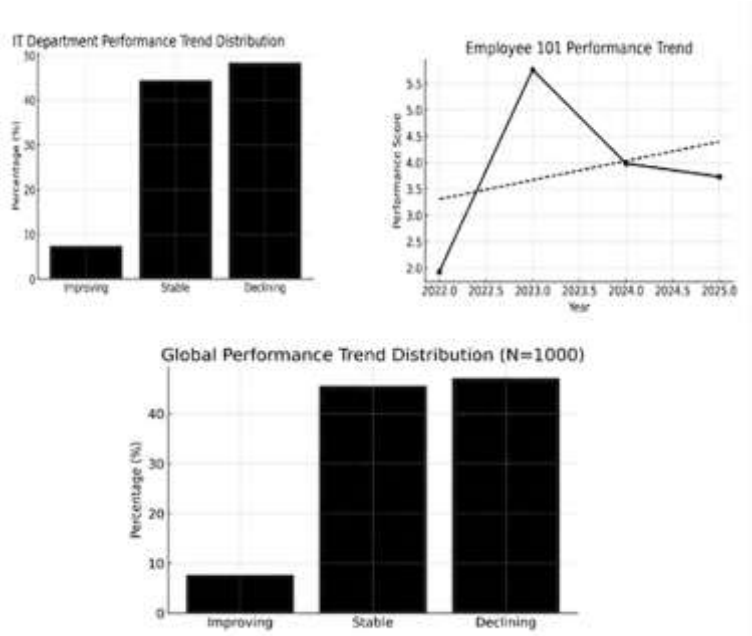


Figure 4.8 Multi Level performance trend analysis across employees and departments

Across 1000 employees, the global trend distribution is that results show that nearly half of the workforce demonstrates a downward performance trajectory, while only a small subset exhibits meaningful improvement. The relatively high proportion of stable trends indicates that a large segment of employees maintains consistent but not increasing performance levels. This finding highlights the need for targeted development and support strategies to prevent performance stagnation.

4.3.3 Are there significant performance differences across departments?

The aim of this question is to examine whether there are meaningful differences in performance levels across departments within the organization. In this context, no machine learning model is employed; rather, department level performance statistics are derived directly from the panel dataset.

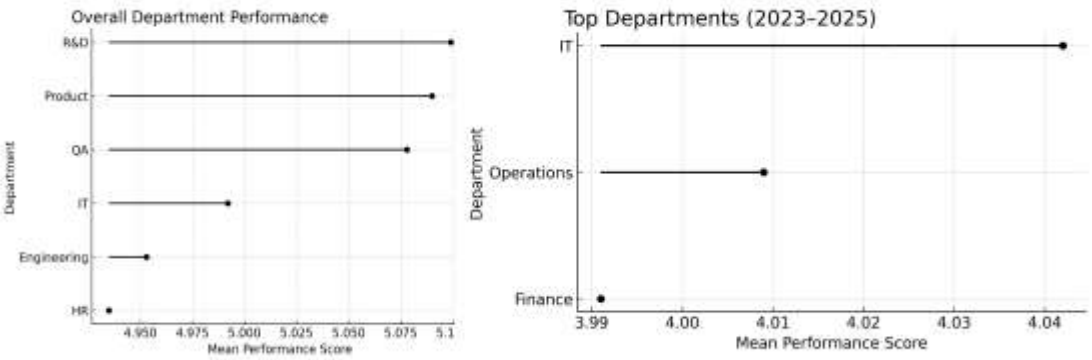
The analysis begins by aggregating the annual performance scores of employees within each department and computing several descriptive statistics based on these values. These include the number of unique employees in the department, the mean performance score,

the standard deviation, and the median performance score. Through these metrics, a systematic overview of each department's performance profile is obtained.

To maintain the robustness of the findings, departments with a number of employees below a predetermined threshold are excluded from the analysis. Subsequently, departments are ranked according to their average performance scores, allowing the identification of those that exhibit the highest overall performance as well as those that fall below the organizational average. This approach makes structural performance differences across units visible and supports the identification of departments' strengths and areas for improvement.

In summary, this method provides an analytical framework that objectively evaluates performance variability at the departmental level and offers valuable insight for human resource management, informing policy development and organizational enhancement efforts.

In Figure 4.9.a presents the graphical outputs related to the departments with the highest and lowest average performance scores. An examination of these charts indicates that, although the organization does not exhibit a sharp performance divide, units that are more closely aligned with the product lifecycle such as R&D and Product demonstrate above average performance levels. Support oriented departments (e.g., HR, IT) appear at relatively lower levels however these differences are not particularly pronounced.



a: Overall Department Performance b: Top Departments(2023 2025)

Figure 4.9 Comparison of Mean Performance Scores across Departments

In Figure 4.9.b graph presents the ranking of departments with the highest performance scores when the analysis is restricted to the years 2023–2025. During this period, the IT department exhibits the highest average performance. This finding suggests that, in recent years, factors such as operational improvements or workload balancing within IT, process optimizations on the Operations side, and the strengthening of a data driven working culture within the Finance department may have contributed positively to overall performance outcomes.

4.3.4 Which departments receive more positive feedback?

The aim of this question is to examine the extent to which different departments within the organization receive positive feedback and to assess how sentiment distributions vary across units. This enables an evaluation of whether organizational indicators such as communication climate, employee satisfaction, and departmental culture differ meaningfully at the unit level. In this analysis, no machine learning model is employed; instead, descriptive statistics are computed using the sentiment scores contained in the panel dataset, which were originally generated through NLP models during Phase 1.

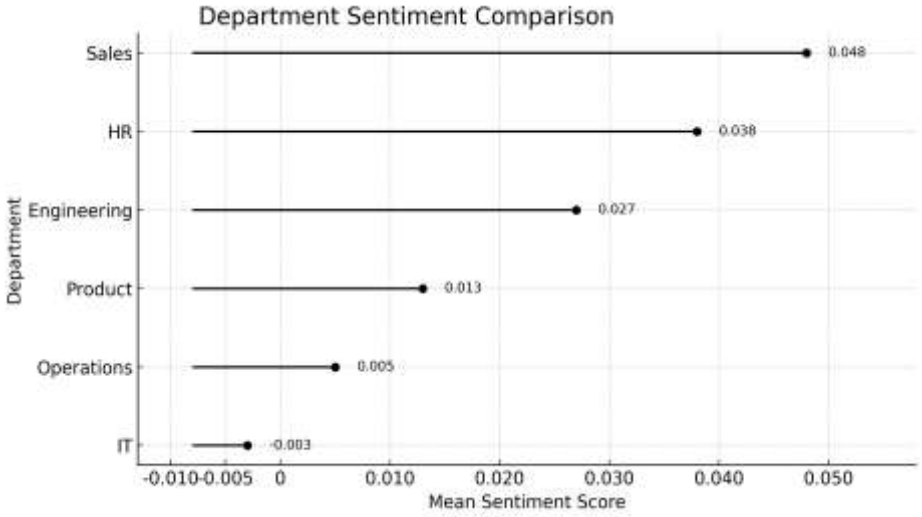


Figure 4.10 Comparison of mean sentiment scores across departments

As shown in Figure 4.10. the Sales, HR, and Engineering departments exhibit the highest average sentiment scores. This finding suggests that customer interaction dynamics,

employee support functions, or cohesive team communication within technical units may contribute to a more positive feedback culture in these departments. In contrast, the IT, Operations, and Product departments display predominantly neutral or slightly negative sentiment levels. Factors such as operational stress, high workload, sudden task demands, or the pressure associated with resolving technical issues may contribute to more negative sentiment trends within IT and Operations.

4.3.5 Is manager–employee communication healthy?

The primary aim of this question is to evaluate the overall health of manager–employee communication within the organization and to examine how communication quality is distributed across departments. In this context, no machine learning model is employed; instead, an employee level classification is performed using communication and feedback indicators extracted from the panel dataset. As an explanation layer, a combination of rule based explanation and panel analytics approaches is adopted.

Communication health is determined using a rule based decision mechanism specifically designed to align with the structure of the dataset. Accordingly:

- If both the communication score and the managerial feedback score exceed predetermined thresholds → the employee is classified as Healthy
- If both scores fall below the thresholds → the employee is classified as Problematic
- Employees who fall between these two extremes → are categorized as At_Risk

This approach offers the advantage of capturing both the quantitative dimension of communication (communication indices) and the qualitative aspect of managerial interaction (manager feedback).

In Figure 4.11 presents the communication health distribution for a total of 1,000 employees and the distribution for the IT group.

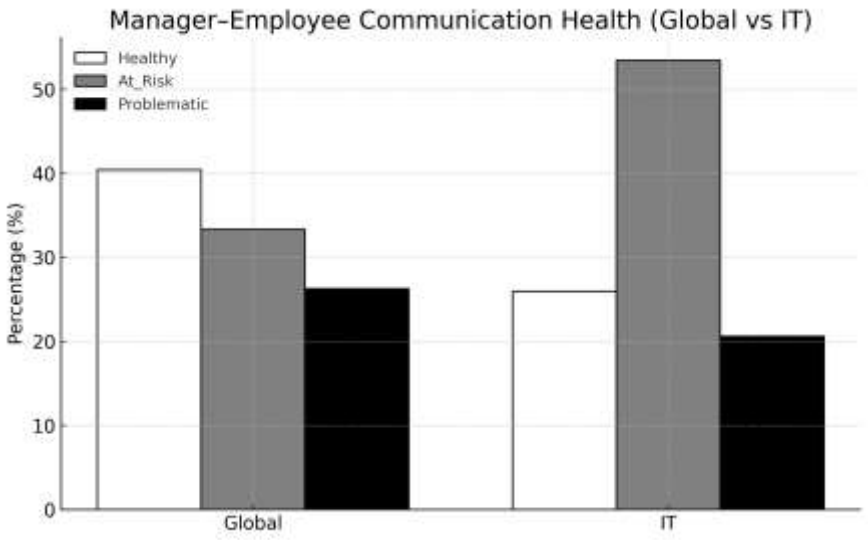


Figure 4.11 Manager–Employee communication health: global vs. IT department

These results indicate that the majority of employees fall within the *Healthy* category, suggesting that overall communication scores and managerial feedback levels exceed the expected thresholds for most individuals.

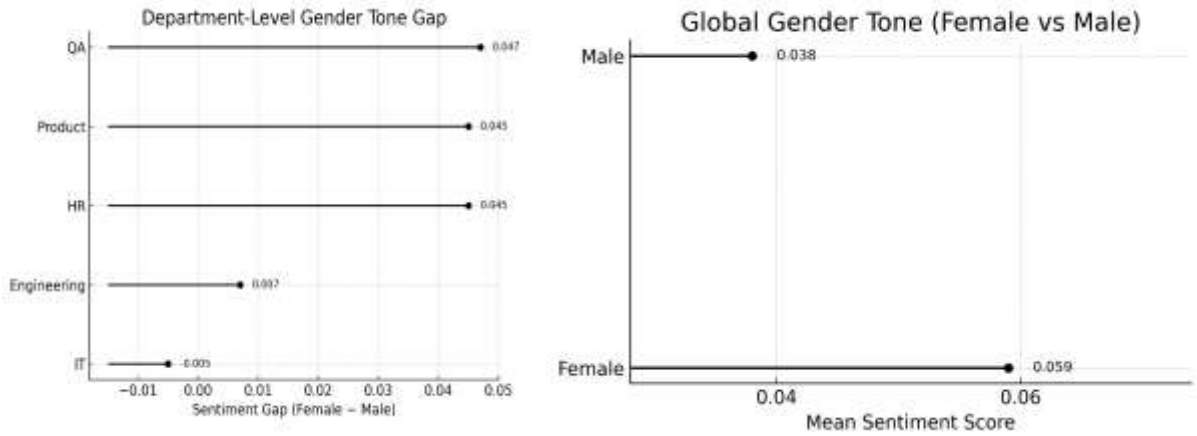
This distribution demonstrates that communication risk is considerably higher in the IT department, with a noticeably smaller proportion of employees classified as *Healthy*. Heavy operational workload, ongoing problem solving pressure, and strict time constraints within IT may be contributing factors to the weaker communication scores observed in this unit.

4.3.6 Is there a sentiment tone difference between male and female employees?

The aim of this question is to examine whether the feedback tone used in employee evaluations differs between male and female employees. No machine learning model is employed for this analysis; instead, descriptive statistics are computed by comparing sentiment scores across gender groups within the panel dataset. The system first automatically identifies the appropriate sentiment column and subsequently calculates the mean, median, and standard deviation of sentiment at both the global and department levels. The gender based difference is interpreted through the “female – male” mean

sentiment gap: a larger gap indicates a more pronounced tone difference, whereas a value close to zero suggests the absence of a meaningful disparity.

As shown in Figure 4.8.b, the tone gap between female and male employees at the organizational level is relatively small; however, the results indicate that female employees tend to use a slightly more positive tone in their feedback compared to their male counterparts.



a: Department Level Gender Tone Gap *b:Global Gender Tone*

Figure 4.12 Gender Based differences in sentiment tone across the organization

Figure 4.8.a presents the department level analysis. In departments such as QA, HR, and Product, the female–male tone gap is noticeably positive, indicating that female employees express more positive sentiment in their feedback within these units. By contrast, in departments such as IT and Engineering, the tone difference between genders is minimal or nearly non existent. The distribution of these disparities across departments suggests that organizational roles and work practices may influence the sentiment expressed in employee feedback.

4.3.7 Does an employee belong to the “high potential” group?

The objective of this question is to determine whether an employee belongs to the high potential (HiPo) group using machine learning–based classification. Unlike the questions

in which no model is employed, this task utilizes a supervised classification approach, where a multidimensional model is trained by incorporating performance metrics, competency indices, feedback indicators, sentiment features, trend variables, and career history derived from the panel dataset. In total, nearly eighty variables are included in the model.

RandomForestClassifier is used as the core predictive model. To address class imbalance, the parameter `class_weight="balanced"` is activated, and additional hyperparameter optimization is applied to improve predictive performance. Random Forest is selected due to its ability to effectively capture nonlinear patterns and complex feature interactions within high dimensional and heterogeneous HR datasets. As the explanation layer, SHAP based global feature importances are employed, enabling a systematic understanding of the most influential factors underlying the model's decisions. Furthermore, when individual level explanation is required, the system switches to "employee mode" and generates personalized decision rationales (top factors) for each employee.

In table 4.6. the global results illustrate which features exert a influence on the high potential classification score. The organization wide HiPo rate is approximately 8%. According to the global feature importance values computed by the model, the most influential predictors of HiPo status are the Performance Score, Manager Feedback Score, Goal Attainment Score, performance trend indicators, and the moving average performance variable. Notably, the high ranking of trend based features such as *Performance_MA2* indicates that not only current performance but also its sustainability plays a critical role in determining high potential employees. Managerial feedback and goal achievement performance appear as supportive factors that meaningfully contribute to the model's predictions.

Table 4.6 Key predictive features for identifying High Potential employees

Rank	Feature	Description	Role in Prediction
1	Performance_Score_Norm	Normalized version of the performance score	Indicates consistent high performance; strongest predictor of high potential status
2	Performance_Score	Raw yearly performance score	Reflects overall performance level
3	Performance_MA2	Two year moving average of performance	Captures performance trend and stability over time
4	Manager_Feedback_Score	Manager's evaluation of the employee	Signals leadership capability and managerial endorsement of potential
5	Goal_Attainment_Score	Ratio of goals successfully achieved	Represents the employee's ability to deliver results and meet organizational expectations

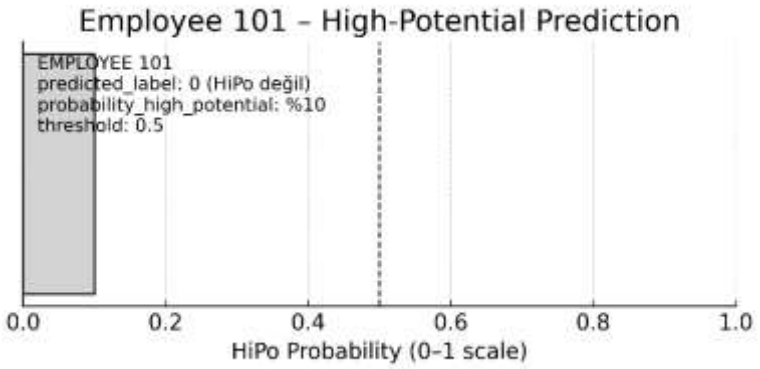


Figure 4.13 Individual High potential probability prediction for employee 101

Figure 4.13 displays the individual level prediction. For Employee 101, the estimated probability of being HiPo is extremely low. The results indicate that the employee's performance scores, trend indicators, and managerial feedback do not align with the profile of a high potential employee. Moreover, the explanation layer (top_factors) does not reveal any strong contributing features, confirming that the model assigns the "Non HiPo" label with high confidence.

4.3.8 Is the employee ready for a managerial role in the near term?

The purpose of this question is to evaluate the output of a machine learning model designed to assess whether employees are ready for a managerial role in the near term. XGBoost model were employed for this task. More than 80 variables including tenure, performance indicators, competency indices, sentiment features, trend metrics, and managerial feedback were incorporated into the training process. The labels were generated through a synthetic rule set aligned with managerial readiness criteria in the literature, jointly considering leadership, ownership, communication skills, teamwork, problem solving ability, and performance stability. These models were selected due to their ability to learn non linear relationships, handle high dimensional HR data effectively, and deliver strong predictive performance for complex targets such as Manager Readiness. SHAP based global feature importance analysis was used to interpret and explain the model's decision making process.

The global feature importance values calculated by the XGBoost based model reveal the variables that most strongly influence the prediction of managerial readiness. The top five contributing factors are as follows.

Total_Work_Experience: Represents the employee's total years of professional experience.

The model uses overall seniority as a strong reference point when estimating managerial readiness. Extensive experience is interpreted as a positive signal, particularly for roles that require leadership and managerial responsibility.

Years_At_Company: Indicates the employee's tenure within the organization.

This reflects the assumption that employees who are familiar with organizational processes, possess institutional knowledge, and are aligned with the organizational culture are more likely to be prepared for managerial positions.

Goal_Attainment_Score: Represents the employee's rate of goal achievement.

Aspects such as results orientation, delivery quality, and success in meeting targets are regarded as critical indicators for managerial roles. The model identifies high goal achievement scores as one of the key components of managerial readiness.

Manager_Feedback_Score: Corresponds to the manager's overall assessment of the employee. Managerial perception acts as a critical signal that encompasses both competence and trust dimensions. High feedback scores indicate that the employee is perceived as reliable not only technically but also in terms of leadership capability and willingness to assume responsibility.

Performance_Score: Reflects the employee's periodic performance score. Consistently high performance levels are treated by the model as one of the fundamental prerequisites for managerial candidacy and therefore contribute strongly to the prediction.

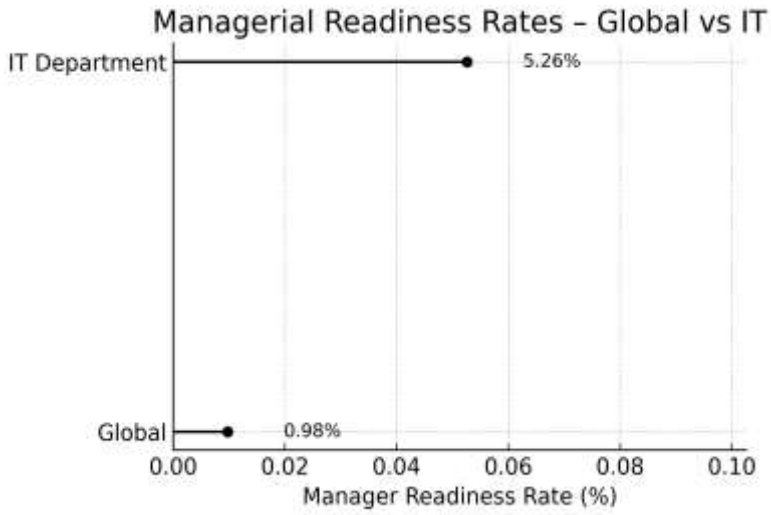


Figure 4.14 Near Term managerial readiness rates: Global vs. IT department

Figure 4.14 presents the outputs obtained from the model. The results indicate that approximately 0.98% of employees across the organization roughly 1% are predicted to be ready for a managerial role. When the analysis is restricted to the IT department,

however, this rate increases to 5.26%**,** suggesting a noticeably higher level of managerial readiness within that unit.

4.3.9 How do training and certifications influence career advancement?

The aim of this question is to examine how education level, annual training participation, and certifications relate to employees' career advancement, operationalized in this study as promotion status. A classification based approach is employed to evaluate whether education related variables contribute to the likelihood of promotion. In addition, correlation analysis and model based feature interpretation are used to assess the relative influence of these variables on promotion outcomes.

In this analysis, the dependent variable is defined as Promotion_Status, representing whether the employee was promoted or not during the relevant period. Independent variables include Education_Level**,** Annual_Training_Count**,** Certification_Count**,** Skill_Development_Index**,** as well as several control variables such as tenure, performance metrics, managerial feedback scores, and competency indices.

The performance metrics of the Logistic Regression model trained using the relevant feature set to answer the question *"How do training and certifications influence career advancement?"* are presented in Table 4.7. Upon examining the metrics, it is evident that the model does not perform well. Its overall performance is only slightly above random guessing (0.50). An F1 macro score of approximately 0.62 indicates that the model struggles to learn both classes (promotion vs. no promotion) in a balanced manner, suggesting a weak underlying signal in the data. The ROC AUC value of 0.626 further demonstrates that although the model can distinguish between the classes, its discriminative power remains limited.

Table 4.7 Performance metrics of the career advancement prediction model

Metrics	Value
Accuracy	0.655
F1 macro	0.622
F1 weighted	0.768
ROC AUC	0.626

In Table 4.8 the model outputs indicate that promotion rates generally exhibit an upward trend as the level of education increases. In particular, there is a meaningful increase from bachelor's degree holders (31.8%) to those with a master's degree (33.1%). Overall, higher education levels have a positive effect on promotion likelihood; however, this effect is not strictly linear.

Table 4.8 Promotion rates across education levels

Education Level Number	Promotion Rate
Associate's degree	26.7%
Bachelor’s Degree	31.8%
Master’s Degree	33.1%
Doctorate	31.3%

In this model, no direct relationship was identified between the number of certifications or training programs completed during the relevant period and promotion outcomes. This is likely due to the model’s limited predictive capacity. However, it should also be noted that in real world settings, the likelihood of a one to one causal relationship between the number of certifications/trainings and promotion decisions is generally considered to be quite low.

4.3.10 How does work–life balance affects employee satisfaction?

The purpose of this question is to examine how employees’ work–life balance influences their overall satisfaction and to identify potential areas for organizational improvement. No machine learning model is employed for this analysis; instead, the relationship

between the Work_Life_Balance_Score and the Employee_Satisfaction_Score is assessed using a rule based and descriptive analytics approach.

Since the panel dataset does not contain a direct "satisfaction" variable for employees, a satisfaction proxy was constructed using the normalized inverse of *the* Burnout Risk Score.

BRSN: Burnout Risk Score Norm

$$\text{Satisfaction}_{\text{Proxy}} = 1 - \text{BRSN} \quad (4.1)$$

The Pearson correlation between the Work_Life_Balance_Score and the Satisfaction_Proxy is:

$$r = 0.0535$$

This value indicates a very weak but positive relationship between the two variables. In other words:

- As work–life balance increases, satisfaction tends to increase as well,
- however, the effect size is quite small,
- and therefore, work–life balance alone is not a decisive factor in shaping satisfaction.

This finding suggests that the concept of satisfaction is too multidimensional to be explained solely by burnout or work–life balance. The literature widely emphasizes that factors such as organizational culture, managerial feedback, workload, career advancement opportunities, and perceptions of pay fairness also play significant roles.

In the first stage of the analysis, employees' work–life balance scores were categorized into three levels:

- Low Balance
- Moderate Balance
- High Balance

Table 4.9 Distribution of Work–Life balance categories across departments

Balance Category	Number of Person	Department
Low Balance	580	IT, Engineering
Moderate Balance	259	Finance
High Balance	161	HR, QA

Table 4.9 presents the category results both at the departmental level and across the organization.

At the company wide level, however, this relationship is generally weak for the majority of employees.

4.3.11 Which departments exhibit higher burnout risk?

The objective of this question is to identify the levels of burnout risk across different departments within the organization and to determine which units exhibit elevated risk based on a data driven approach. Rather than relying solely on simple descriptive statistics, this analysis employs a machine learning–based risk prediction model that holistically evaluates behavioural, performance oriented, and sentiment driven indicators of employee well being.

Two classification models are used to estimate burnout risk:

- XGBoost Classifier
- Random Forest Classifier

To enhance interpretability, SHAP based global feature importance analysis is applied, enabling a clear understanding of which factors the models rely upon when predicting burnout risk. The SHAP results indicate that high workload (e.g., extended working hours), negative sentiment patterns, and unstable or unsustainable performance trajectories are among the most influential contributors to increased burnout risk.

The XGBoost model was selected as it exhibited superior classification performance. The evaluation metrics for the model are presented in the Table 4.10 below.

Table 4.10 Performance metrics of the burnout risk classification model

Metrics	Value
Accuracy	0.804
F1 macro	0.799
F1 weighted	0.805
ROC AUC	0.862

Table 4.11 Department Level burnout risk rates and average probabilities

Department	High Risk Rate
Engineering	0.452
QA	0.449
Finance	0.445
IT	0.444
HR	0.441
Sales	0.425
Product	0.425
Operations	0.424
R&D	0.397

Based on the model outputs, the departmental distribution of burnout risk across all employees is presented in Table 4.11. The top three high risk departments identified by the model are, Engineering – 45.2% high risk, QA – 44.9% high risk, Finance – 44.5% high risk. These findings suggest that factors such as high technical intensity, strong delivery pressure, continuous problem solving expectations and elevated exposure to customer/product issues substantially contribute to the increased burnout risk in these departments.

The burnout model indicates that the distribution of risk across departments is not homogeneous, revealing substantially higher burnout levels particularly within units that involve technical roles. Therefore, the model's findings provide an evidence based foundation for developing preventive strategies such as workload rebalancing, strengthening feedback culture, implementing employee support programs, and introducing flexible work planning particularly in units where burnout risk is found to be high.

4.3.12 Which metrics best explain salary raise decisions?

The objective of this question is to identify the metrics that best explain salary raise decisions within the organization and to uncover the measurable criteria underlying these decisions through a data driven approach. Rather than relying solely on correlation analysis or simple descriptive statistics, this study employs an XGBoost Classifier model that holistically evaluates employees' performance indicators, competencies, work output, feedback scores, and behavioural signals. To interpret the model's outputs, SHAP based global feature importance analyses are conducted, enabling a systematic examination of the variables that exert the strongest influence on salary raise decisions.

The table 4.12 below presents the metrics that the model identifies as being correlated with salary raise decisions. An examination of these metrics shows that raise decisions are largely aligned with performance and output oriented indicators.

Table 4.12 Key metrics explaining salary raise decisions

Parameter	Percentage of Raise
Performance Score	0.66
Strategic Initiative Impact	0.53
Deadline Adherence Rate	0.42
Goal Attainment Score	0.30
Manager Feedback Score	0.28

In other words, the table indicates that employees with higher performance scores tend to receive greater salary increases, and similarly, those with a high level of strategic initiative impact also exhibit a higher likelihood of receiving raises.

4.3.13 Are promotion and salary increase decisions fair across departments?

The aim of this question is to assess whether promotion and salary increase decisions are distributed fairly across departments within the organization and to detect potential structural disparities using data driven methods. For this purpose, the fairness analysis is

conducted through both descriptive comparisons at the department level and machine learning–based evaluation techniques.

In the first stage of the analysis, the outputs of logistic regression models trained to predict promotion and salary increase decisions are grouped by department. Subsequently, promotion_rate**,** salary_increase_rate, and model_predicted_probability values for each department are compared. This approach enables the examination of whether departments operate under similar organizational conditions and whether there is any systematic deviation (bias) in decision distributions.

In the second stage, SHAP based explanation techniques are employed to determine whether the model's decisions remain consistent across departments. SHAP values provide insight into which features drive promotion or salary increase predictions and whether these decision making patterns differ between departments.

The individual explanation mechanism allows for micro level fairness assessment by revealing which factors influence the decision for each employee. For example, if two employees with comparable performance, tenure, and feedback profiles receive different promotion eligibility outcomes, the system can clearly identify the factors underlying this discrepancy through their SHAP attributions. This enables the detection of potential department based bias and ensures transparent evaluation of decision processes.

When employees who received a promotion are compared with those who did not, it becomes evident that goal attainment and performance indicators are strongly aligned with promotion decisions. As shown in Table 4.13 summary statistics reveal the following patterns:

Table 4.13 Fairness analysis of promotion decisions based on key performance metrics

	Promoted	Non promoted	Diff	Explanation
Goal_Attainment_Score	4.65	4.15	0.50	Difference indicates that goal completion success plays a significant distinguishing role in promotion outcomes
Performance_Score	4.50	4.20	0.30	Difference shows that the general performance level of promoted employees is meaningfully higher
Manager_Feedback_Score	4.55	4.30	0.25	It suggests that managerial feedback serves as a secondary yet consistent signal contributing to promotion decisions.
Regarding Total_Work_Experience	4.20	4.15	0.05	Indicates that the organization does not follow a strictly seniority based promotion practice.

Overall, these results demonstrate that promotion decisions within the organization are primarily driven by measurable performance outcomes, complemented by managerial evaluations.

When gender based promotion rates are examined, the disparate impact (DI) value calculated by the system is found to be approximately 0.889. DI is computed as the ratio of the promotion rate of the group with the lower rate to that of the group with the higher rate. According to the widely used ‘80% rule’ in the fairness literature, a DI value below 0.80 is typically interpreted as an indicator of meaningful adverse impact. The DI value of 0.889 obtained in this study suggests that, although the promotion rates between gender

groups are not perfectly symmetrical, the difference remains above the critical threshold of 0.80. Therefore promotion rates are not entirely equal across genders; however, in the simulated scenario, no substantial pattern of discrimination is observed. The disparity remains limited and within an acceptable range. This finding indicates that the system does not generate systematic gender based disadvantage, though there remains room for future improvement to support more balanced policy design.

When departmental averages of Promotion_Status_Num are examined, it becomes evident that promotion rates differ meaningfully across units:

- Sales exhibits the highest promotion rate at approximately 42.0%.
- This is followed by Finance (~37.0%) and technical/support units such as IT and Engineering (~33–34%).
- In departments such as R&D and Product, promotion rates decline to the 25–27% range.

This distribution can be interpreted from two complementary perspectives. In target and revenue driven departments such as Sales, performance and goal achievement are more visible and objectively measurable. Therefore, higher promotion rates in such units may be a natural outcome of their operational structure. Lower promotion rates in departments like R&D and Product may indicate that promotion mechanisms struggle to fully capture performance signals in long term, project based, or less visibly quantifiable work domains. This points to a potential policy improvement area regarding how performance and contribution are assessed in these functions.

Overall, the differences in promotion rates across departments should not be interpreted solely as indicators of “unfairness.” Instead, they must be understood in relation to the nature of work, performance measurement methods, and role expectations. Nevertheless, conducting a more detailed examination of departments with notably lower promotion rates appears critical for strengthening organizational justice perception.

4.3.14 Is the increase in positive sentiment aligned with salary raises or promotions?

The aim of this question is to examine the extent to which increases in positive sentiment within employee feedback align with salary raise or promotion decisions. No machine learning model is used for this analysis; instead, a trend alignment approach is applied by evaluating employees' sentiment trends alongside their career progression indicators over time.

Since this question is entirely based on data analysis and statistical interpretation, fairness detection was applied using a combination of salary related variables (Prev_Salary, Current_Salary, Salary_Raise_Rate), sentiment based attributes generated by the sentiment model (Sentiment_Index, Sentiment_Delta, Sentiment_Prev), and career related indicators (Promotion_Status and Salary_Increase_Status).

Table 4.14 presents the corresponding values for a selected employee. An examination of these values shows that the sentiment scores reflected in the employee's feedback do not exhibit a meaningful relationship with the employee's promotion or salary increase outcomes. Even when the sentiment remained neutral, the employee was observed to receive both a promotion and a salary raise.

Table 4.14 Individual Level alignment between sentiment change and career advancement decisions

Employee_ID	1454
Year	2021
Current_Salary	79479 TL
Prev_Salary	71300
Salary_Raise_Rate	1147 TL
Sentiment_Index	0.0
Sentiment_Prev	0.0
Sentiment_Delta'	0.0
Promotion_Status	Promoted
Salary_Increase_Status'	Yes

Similarly, when the company wide distribution of sentiment categories (neutral, negative, positive) is analyzed as shown in the table 4.15. below and compared with global

promotion and salary increase statuses, no statistically meaningful alignment is observed. In other words, at the organizational level, positive sentiment trends do not systematically predict or correspond to promotion or salary increase decisions

Table 4.15 Alignment between Sentiment Categories and Career Advancement Outcomes

Category of Sentiment	Salary Increase Status	Promotion_Status
Negative sentiment	No	Promoted
Neutral sentiment	Yes	Promoted
Positive sentiment	Yes	Not Promoted

4.4 Discussion

This study presents a multi layered data science framework designed to enhance decision making processes in human resources management. By integrating NLP based text analytics, machine learning–driven predictive modelling, individual level explainability techniques, fairness diagnostics, and time series trend assessments under a unified analytical architecture, the research demonstrates how complex HR problems can be addressed with a holistic, data centric approach. The two foundational NLP components of the thesis Competency Classification and Sentiment Analysis played a critical role throughout the entire system, evolving from a single year (2025) structure into a multi year panel framework (2021–2025). These modules enriched the analytical pipeline by providing behavioural and linguistic signals that fed into advanced prediction and diagnostic mechanisms.

From a data centric perspective, the synthetic panel dataset played a critical role in enabling controlled experimentation across multiple HR analytics tasks. By design, the dataset preserved realistic distributional properties such as skewed departmental sizes, hierarchical role pyramids, and temporally evolving performance patterns while avoiding deterministic relationships between input features and outcome variables. This balance allowed the models to learn meaningful associations without exploiting artificial shortcuts inherent to overly structured synthetic data.

Between 2021 and 2025, a certain amount of noise was added to the performance scores and goal attainment scores, thereby creating variations across the years. The calculations related to these features are detailed in the data library section. The method used for adding noise is uniform distribution. Uniform distribution provides a clearly interpretable and bounded noise mechanism, ensuring that no single annual evaluation deviates beyond a predefined range. An alternative approach would have been to model ε as a Gaussian random variable, which is the most commonly adopted noise model in performance simulation literature. The Gaussian distribution offers the theoretical advantage of reflecting the Central Limit Theorem. Another alternative approach would have been Heavy tailed distributions such as the Student t. While Student t noise would produce more realistic outlier behaviour, it also risks inflating variance in a synthetic dataset where ground truth labels are already well defined, potentially obscuring the signal to noise ratio that downstream classification models depend upon.

The competency model, developed using TF IDF and Logistic Regression, successfully extracted behavioural and technical competency indicators from employee feedback. Indices such as Communication, Ownership, Leadership, Teamwork, and Innovation were embedded into the panel dataset and later served as essential features in higher level analytic tasks including High Potential prediction, managerial readiness, salary increase modelling, promotion fairness evaluation, and burnout risk estimation.

Similarly, the sentiment analysis module, built on an XLM R architecture enhanced with weak supervision, context support, label smoothing, temperature scaling, and neutral margin methods, enabled accurate classification of employee feedback into positive, neutral, and negative tones. Although early versions of the model exhibited class imbalance and negative bias tendencies, iterative improvements significantly enhanced contextual reliability. Sentiment signals became central inputs for analyses such as gender tone gap, department level sentiment differences, sentiment trend alignment, and burnout risk modelling.

Collectively, the NLP modules served not only as independent analytical tools but also as the behavioural signal backbone of the entire HR decision engine. The multi model

HR Decision Engine demonstrates how organizations can augment their appraisal processes with data driven insights:

- **Managerial decision support:** Promotion and readiness predictions offer structured guidance for succession planning.
- **Organizational health monitoring:** Sentiment and competency drifts help detect emerging dissatisfaction or engagement declines.
- **Workforce planning:** Multi year trends support proactive talent development strategies.
- **Performance calibration:** NLP indicators help validate the consistency of qualitative evaluations with quantitative performance metrics.

By adopting such a system, organizations can reduce reliance on subjective judgments, improve transparency, and enhance alignment between employee development and organizational objectives.

This study goes beyond a single classification task; it is designed as an end to end performance evaluation system that seeks to answer diverse analytical questions such as high potential identification, burnout risk, managerial readiness, fairness of salary increases, and gender based differences in sentiment tone. Each module is modelled using a different supervised learning approach, depending on the nature of the problem. If unsupervised learning had been adopted as the primary methodology in this system, both the benefits and limitations would vary significantly across modules.

High potential classification, burnout risk prediction, and managerial readiness modules are designed as supervised classification tasks using Random Forest, XGBoost, along with an explainability layer based on SHAP. In these modules, labels form the backbone of the system classifying an employee as “high potential” or “at risk of burnout” relies on a ground truth derived from historical HR decisions or expert evaluations.

Table 4.16 presents each question appropriate techniques.

Table 4.16 Summary of analytical questions and corresponding methodological approaches

Question Category	Methodological Approach
High Potential Classification	Random Forest + SHAP
Burnout Risk	XGBoost / RF hybrid + SHAP
Manager Readiness	XGBoost + Individual level explanation reports
Salary Raise Drivers	XGBoost Classifier + SHAP
Fairness Across Departments	Logistic Regression + SHAP + disparity indices
Sentiment–Promotion/Salary Alignment	Trend slope + alignment matrix
Gender Tone Differences	Sentiment comparative analysis
Department level Positive Feedback	Sentiment statistics
Performance trend, departmental differences, sentiment distributions	Descriptive panel analytics

When examining the methodological approach and model outcomes for each question, it is observed that the High Potential and Manager Readiness models exhibit the highest feature importance scores. This finding aligns closely with the leadership potential and managerial readiness criteria widely discussed in the literature. The model results further confirm that positive behavioural indicators such as ownership, task efficiency, and strategic initiative together with managerial feedback scores, play a complementary role in identifying HiPo profiles.

In the Manager Readiness model, the prominence of seniority related variables (e.g., Total_Work_Experience, Years_At_Company) indicates that managerial preparedness is influenced not only by performance based factors but also by accumulated experience and organizational contextual knowledge. This outcome is consistent with the concept of "managerial maturity" which is frequently emphasized in academic studies.

Fairness analysis constitutes a critical component of this thesis, as it clarifies whether departmental differences in promotion and salary increase decisions stem from workload, role specific responsibilities, and KPI variability, or from potential structural bias. The analyses reveal that although departments exhibit varying promotion_rate and salary_increase_rate values, SHAP based individual explanations indicate that the model's decisions remain consistent. Specifically, employees with similar profiles and

performance levels receive comparable predicted probabilities, regardless of their department. This finding suggests that the decision making mechanism does not exhibit "department level bias," and that the observed differences are attributable to the organization's operational structure rather than discriminatory effects.

High working hours, negative sentiment trends, declining performance trajectories, and low task efficiency values were found to strongly influence burnout risk. This model not only identifies employees with currently elevated burnout levels but also detects those who may become at risk in the future, thereby enabling proactive intervention by HR. The observation of higher burnout risk in certain departments particularly Operations, QA, and IT may be associated with workload imbalance, delivery pressure, and a culture of continuous problem solving.

The analysis of sentiment trend alignment showed that increases in positive sentiment were not systematically aligned with promotion or salary raise decisions at the organizational level. Although individual cases of alignment were observed, the overall results did not indicate a statistically meaningful relationship between sentiment improvement and career advancement outcomes. This suggests that sentiment signals may provide complementary insight into employee experience, but they do not function as direct determinants of promotion or salary increase decisions in the current framework. These findings suggest that the organization could enhance its decision making mechanisms by more comprehensively incorporating employee experience signals.

If an unsupervised learning approach had been adopted in the study instead of supervised learning models; these modules could be redesigned using clustering techniques such as K Means, DBSCAN, or hierarchical clustering. The model would group employees based on similarity in features or semantic proximity without using labels. However, in this case, it would not be possible to directly interpret whether a cluster represents "high potential" or "low engagement"; assigning meaning to clusters would still require human interpretation from an HR expert.

More critically, it would become impossible to answer questions like “Why is this employee considered high potential?” through SHAP values whereas this explainability layer lies at the core of the system’s organizational value.

In the salary increase drivers and inter departmental fairness modules, Logistic Regression and XGBoost Classifier are used together with SHAP. This setup enables explaining which features drive decisions at both the model level and the individual level. It also serves a critical function in terms of organizational accountability.

An unsupervised approach would be particularly insufficient for these modules. To evaluate whether salary increase decisions are fair, it is essential to jointly model the actual decisions (labels) and the features influencing those decisions. A clustering based approach could group employees into certain profile segments; however, it would not be able to verify whether these groups are meaningfully associated with salary increases. Disparity indices could not be computed, and statistical significance could not be evaluated. Therefore, in these modules, unsupervised learning is not a holistic alternative in this system but can be considered a selective complementary approach. In classification and fairness modules that require labels, the supervised approach is indispensable both as a technical necessity and in terms of organizational defensibility. However, in future work, a hybrid design can be adopted: unsupervised clustering can be used as a preliminary step to discover employee profiles, and these profiles can then be fed into supervised models as inputs. This approach has the potential to combine data driven hypothesis generation with measurable predictive performance under a unified framework.

This study was designed as an end to end performance evaluation system encompassing high potential identification, burnout risk detection, and managerial readiness assessment constructed upon a synthetic HR dataset generated for 1,000 employees. Had the dataset been scaled to 20,000 or more employees, the system would not have merely grown in size; rather, methodological choices, model architectures, computational infrastructure, and the reliability of outputs would have undergone fundamental transformation.

With a dataset of 1,000 instances, decision boundaries remain inherently fragile; the misclassification of even a few examples in the test set can shift macro F1 scores by a meaningful margin. Scaling to a 20,000 employee dataset would largely eliminate this fragility. Ensemble methods such as Random Forest and XGBoost improve in direct proportion to data volume a greater number of instances yields deeper and more stable trees, thereby reducing false negative rates in both high potential classification and burnout risk prediction. For transformer based models such as XLM RoBERTa, a dataset of 20,000 instances provides a genuine foundation for fine tuning. At the 1,000 instance scale, such models are largely constrained to rely on their pre trained weights; at the 20,000 instance scale, however, domain specific patterns including organization specific language, terminology, and appraisal conventions become learnable and can be effectively encoded into the model.

In the current 1,000 employee dataset, class imbalance has posed a serious methodological challenge, particularly in categories such as Innovation and Manager Readiness. This necessitated the adoption of specialized loss functions such as Focal Loss, as well as class weighting strategies. In a 20,000 employee dataset, this imbalance would naturally diminish. As the number of instances belonging to rare categories increases, the need for compensatory mechanisms such as Focal Loss weakens accordingly, enabling models to learn minority classes in a more organic manner.

An increase in data volume brings not only opportunities but also proportionally growing risks. In an organization of 20,000 employees, inter rater inconsistency inevitably intensifies. Texts produced by different managers, across different time periods and within departments characterized by varying evaluation cultures, may express the same underlying construct in substantially divergent ways. This noise elevates data cleaning and standardization to a critical preprocessing stage.

Conceptual drift also constitutes a significant risk. The evolution of competency definitions over the course of an organization's growth may result in data from different periods carrying inconsistent content under the same label. This presents a particularly

serious methodological problem in longitudinal modules specifically in performance trend analysis and sentiment promotion alignment.

As the dataset grows in scale, larger model variants such as XLM RoBERTa Large may become preferable. Naturally, as both the dataset expands and the model architectures evolve, the associated data volume and hardware requirements change considerably. An approximate estimation of these requirements is presented in Table 4.17:

Table 4.17 Summary of hardware requirements & data size

Hardware Requirements & Data Size	
GPU	NVIDIA A100 (40–80 GB VRAM)
RAM	64–128 GB
Storage Capacity	100–250 GB
Total feedback texts	400,000 number of texts
Excel file size	20 MB

The primary bottleneck in scaling this system is not the structured tabular data but the embedding matrix. Loading the embeddings of 400,000 texts into memory simultaneously would not be feasible in a 20,000 employee scenario. This constraint introduces the following practical infrastructure requirements:

- Texts must be processed in sequential batches of 500–1,000 samples, with the resulting embeddings written to disk incrementally rather than held in memory.
- Dedicated storage must be allocated for embeddings.
- System RAM must be upgraded.
- GPU VRAM requirements increase.

A dataset of 20,000 employees would elevate this system to a qualitatively distinct platform both methodologically and organizationally. Model reliability would increase, new analytical possibilities would emerge, and findings would become generalizable to a broader employee population. However, this transition simultaneously demands a substantial degree of organizational maturity in terms of data governance, computational infrastructure, and ethical AI oversight. Specifically, such a deployment would exceed

the operational boundaries of standard workstation environments or cloud based notebook platforms, necessitating a dedicated machine learning infrastructure whether on premise high performance computing or a managed cloud ML service along with systematic ML Ops practices to ensure reproducibility, monitoring, and periodic model retraining. The methodological framework established in this study using a 1,000 employee dataset nonetheless provides both the technical and conceptual foundation for a transition to this large scale scenario.

Table 4.18 Space&Time Complexity Big O notation

Models	Space Complexity (Big O Notation)	Space requirement when 1000 to 20000 employees	Time Complexity (Big O Notation)	Time requirement when 1000 to 20000 employees
Logistic Regression	$O(d)$	Constant →	$O(n.d.iter)$	20×↑
Random Forest Classifier	$O(T.n)$	20×↑	$O(T.n.d.\log_2 n)$	28×↑
XGBOOST Classifier	$O(T.depth.d)$	Constant →	$O(T.n.d.\log_2 n)$	28×↑
XLM Roberta	$O(L.T^2 + P)$	Constant →	$O(n.L.T^2.d)$	20×↑

In Table 4.18, the notation terms are defined as follows:

n: Sample Size,

d: Feature Dimension,

T:Number of Trees,

L: Number of Layers,

P:Number of Parameters,

Although the present study was conducted on a dateset of 1,000 employee records, it is worth examining how the computational requirements of the four selected models would scale under a hypothetical expansion to 20,000 samples. Table 19 provides a comparative overview of space and time complexity for each model using Big O notation, illustrating the theoretical growth in memory and processing demands under this scenario.

5. CONCLUSION

This thesis presents a comprehensive and integrated analytical framework aimed at strengthening artificial intelligence–supported decision making mechanisms in human resources processes. The proposed framework was implemented on a realistic yet privacy preserving synthetic HR panel dataset, enabling methodological consistency and reproducibility while overcoming common constraints related to data accessibility, ethics, and confidentiality. By jointly leveraging natural language processing (NLP) and machine learning (ML) techniques, the study addresses critical HR questions such as performance evaluation, managerial readiness, promotion and salary increase decisions, and inter departmental fairness from a multidimensional perspective.

One of the key contributions of this study is the successful development of a multi layered and explainable analytical architecture validated on a synthetic dataset constructed through a rule based and probabilistic data generation protocol. The dataset preserves realistic organizational patterns, including departmental distributions, hierarchical role structures, temporally evolving performance trends, and behavioural signals, while deliberately avoiding deterministic relationships between input features and outcome variables. This design choice ensured that the developed models learned meaningful and generalizable patterns rather than exploiting artificial shortcuts inherent to overly structured synthetic data.

In particular, the integration of competency classification and sentiment analysis outputs derived from synthetically generated employee feedback texts together with structured performance indicators allowed linguistic signals to be positioned as complementary behavioural features rather than direct proxies for decision variables. This approach enhanced the interpretability of NLP based features while minimizing the risks of label leakage and circular inference.

Furthermore, the controlled experimental environment provided by the synthetic dataset enabled a robust evaluation of inter departmental fairness. Through SHAP based explanations, model decisions were interpreted at both global and individual levels,

allowing observed differences to be more clearly attributed to structural organizational factors rather than potential algorithmic bias.

The analysis of sentiment trends and career related outcomes within the multi year panel structure demonstrated that increasing negative sentiment, declining performance trajectories, and elevated workload signals may be associated with adverse future outcomes. These findings indicate that the synthetic dataset not only supported model development and evaluation but also functioned as a simulation environment for testing early warning mechanisms and proactive HR interventions.

This thesis contributes to the literature in the following ways:

1. An integrated NLP–ML–based HR analytics framework validated on a realistic and privacy preserving synthetic dataset.
2. A domain oriented competency classification scheme developed and evaluated on rule based synthetic feedback texts.
3. A context aware sentiment analysis approach adapted to organizational communication settings.
4. A scalable and reproducible multi year synthetic HR panel dataset enabling research in domains with restricted access to real world data.
5. Empirical evidence demonstrating how the joint use of linguistic and numerical indicators provides complementary insights for HR decision making processes.

Despite these contributions, certain limitations should be acknowledged. Due to the nature of synthetic data, the dataset cannot fully capture complex interpersonal dynamics, informal communication patterns, or organizational politics present in real world environments. Additionally, transformer based sentiment models were observed to exhibit sensitivity to ambiguous or neutral expressions, highlighting the need for domain specific fine tuning and data enrichment in future work. These limitations provide important context for interpreting the findings and point to directions for further research.

Overall, this thesis demonstrates that NLP and ML based analytical systems supported by deliberately designed synthetic datasets offer a strong foundation for developing fair, transparent, and strategic AI driven decision support mechanisms in human resources management.

REFERENCES

Alon Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making:“automation bias” and “selective adherence” to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153 169.

Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF IDF approach. In 2016 International conference on electrical, electronics, and optimization techniques (ICEEOT) (pp. 61 66). IEEE.

Bracken, D.W., Rose, D.S. & Church, A.H. (2016). The evolution and devolution of 360° feedback. Industrial and Organizational Psychology, 9(4), 761–794.

Budhwar, P., Malik, A., De Silva, M. T., & Thevisuthan, P. (2022). Artificial intelligence–challenges and opportunities for international HRM: a review and research agenda. The InTernaTIonal Journal of human resource managemenT, 33(6), 1065 1097.

Carter, L., & Liu, D. (2025). How was my performance? Exploring the role of anchoring bias in AI assisted decision making. International Journal of Information Management, 82, 102875.

Cha, Y. (2025). Employees’ reactions to algorithmic performance evaluation: threat of evaluation bias and objectivity. Journal of Information Systems, 39(2), 1 20.

Chowdhury, S., Dey, P., Joel Edgar, S., Bhattacharya, S., Rodriguez Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. Human resource management review, 33(1), 100899.

Chuang, S., Shahhosseini, M., Javaid, M., & Wang, G. G. (2024). Machine learning and AI technology induced skill gaps and opportunities for continuous development of middle skilled employees. Journal of Work Applied Management.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM computing surveys, 55(9), 1 33.

Ekuma, K. (2024). Artificial intelligence and automation in human resource development: A systematic review. Human Resource Development Review, 23(2), 199 229.

Jafari, M., Bourouni, A., & Amiri, R. H. (2009). A new framework for selection of the best performance appraisal method. European Journal of Social Sciences, 7(3), 92 100.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020, October). RFRSF: Employee turnover prediction based on random forests and survival analysis. In International Conference on Web Information Systems Engineering (pp. 503 515). Cham: Springer International Publishing.

Juhdi, N., Pa'wan, F., & Hansaram, R. M. S. K. (2012). EXAMINING CHARACTERISTICS OF HIGH POTENTIAL EMPLOYEES FROM

EMPLOYEES'PERSPECTIVE. International Journal of Arts & Sciences, 5(7), 175.

Kalff, Y., & Simbeck, K. (2025). Explained, yet misunderstood: How AI Literacy shapes HR Managers' interpretation of User Interfaces in Recruiting Recommender Systems. arXiv preprint arXiv:2509.06475.

Lukaszewski, K. M., & Stone, D. L. (2024). Will the use of AI in human resources create a digital Frankenstein?. Organizational Dynamics, 53(1), 101033.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1 35.

Moldoveanu, M., & Narayandas, D. (2019). The future of leadership development. Harvard business review, 97(2), 40 48.

Pan, Y., Froese, F. J., & Xue, S. (2025). The Role of AI in Performance Appraisal: A Mixed-Method Study of Employee Experience Through a Relational Lens. Human Resource Management.

Pessach, D., Singer, G., Avrahami, D., Ben Gal, H. C., Shmueli, E., & Ben Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Decision support systems, 134, 113290.

Ponnuru, S. A., Merugumala, G., Padigala, S., Vanga, R., & Kantapalli, B. (2020). Employee attrition prediction using logistic regression. Int. J. Res. Appl. Sci. Eng. Technol, 8(5), 2871 2875.

Qin, S., Jia, N., Luo, X., Liao, C., & Huang, Z. (2023). Perceived fairness of human managers compared with artificial intelligence in employee performance evaluation. Journal of Management Information Systems, 40(4), 1039 1070.

Reimers, N., & Gurevych, I. (2019, November). Sentence bert: Sentence embeddings using siamese bert networks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP IJCNLP) (pp. 3982 3992).

Regulation, P. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. Regulation (eu), 679(2016), 10 3.

Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., & Zschech, P. (2025). CareerBERT: Matching resumes to ESCO jobs in a shared embedding space for generic job recommendations. Expert Systems with Applications, 275, 127043.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), 2400304.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., ... & Perez, E. (2025). Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. arXiv preprint arXiv:2501.18837.

Shi, R., Wang, Y., Du, M., Shen, X., Chang, Y., & Wang, X. (2025). A comprehensive survey of synthetic tabular data generation. arXiv preprint arXiv:2504.16506.

Stahl, G. K., Björkman, I., Farndale, E., Morris, S. S., Paauwe, J., Stiles, P., ... & Wright, P. (2011). Six principles of effective global talent management. MIT Sloan management review.

Thirunagalingam, A., Addanki, S., Vemula, V. R., & Selvakumar, P. (2025). AI in performance management: Data driven approaches. In Navigating organizational behavior in the digital age with AI (pp. 101 126). IGI Global Scientific Publishing.

Yanamala, K. K. R. (2022). Integrating machine learning and human feedback for employee performance evaluation. Journal of Advanced Computing Systems, 2(1), 1 10.

CURRICULUM VITAE

Name and surname : Nağme CÖMERTLER

Foreign language :

Education:

Bachelor of Science (BSc): Karadeniz Technical University Department of Electrical and Electronics Engineering (2018)

Master of Science (MSc): Ankara University, Graduate School of Natural and Applied Sciences, Department of Artificial Intelligence Technology 2026