**FINAL PROJECT**

FINAL PROJECT

Fixing Biased Talent Promotion Decisions Through

Data-Driven HR Analytics

Submitted to fulfill graduation requirements

Bootcamp Data Science Rakamin Batch 59

Disusun oleh :

Bayu Maitra

Dian Ulhaq Qurrata A’yun

Mashlahatul Husna

Keisya Nisrina Aulia

Febiansyah Annaufal Ahnaf Fauzi



**TAHUN 2025**

ABSTRACT

The employee promotion process is often influenced by subjectivity, inconsistent data, and biased criteria, thereby reducing the fairness and effectiveness of HR decisions. This study aims to develop a data-driven decision support system to improve the transparency, accuracy, and objectivity of the promotion process. Using the Talent Promotion dataset from Rakamin, a series of CRISP-DM stages were carried out, starting from EDA, data cleaning, feature engineering, to the development of machine learning and clustering models. The analysis results showed data anomalies, bias in the Promotion\_Eligible feature, and excessive dependence on Peer\_Review\_Score and Projects\_Handled. Several algorithms were tested, including Random Forest, XGBoost, Logistic Regression, SVM, and K-Means. The best model for prediction was Logistic Regression with high accuracy and recall, while K-Means (k=4) produced the most stable talent segmentation for performance and potential mapping. This system was then implemented into a Streamlit-based interactive dashboard to provide real-time promotion recommendations. Overall, this data-driven approach improved the accuracy of promotion assessments by 38% and has the potential to save 22.5% in operational costs, while supporting HR in making fairer and more efficient decisions.

keyword: Talent Promotion, Machine Learning, Deployment Model.

Acknowledgement

We give thanks to God Almighty for His grace, which has enabled us to successfully complete the final report for the Rakamin Academy Data Science Bootcamp Batch 59 Final Project, entitled “The Promotion Paradox.”

We would also like to thank those who contributed to the process and completion of this project, namely:

* The entire Rakamin Academy Team, Tutors, and Mentors, Mr. Jomen tutor assistant, Mr. Hanif as a mentor, Mrs. Dea, and Mr. Danar as class coordinators, and Mrs. Syafa as the Rakamin team, who have guided us and provided valuable knowledge and insights during the bootcamp.
* The Syntax Society team, who collaborated and actively contributed to every stage of this project.

We hope this report can be a meaningful contribution to the development of data-based solutions and provide inspiration for the application of machine learning in the real world, especially in the banking or financial sector.

Jakarta, 7 December 2025

Best regards,

Syntax Society

# Table of Contents (Revised & Professional)

# CHAPTER I — Introduction

# 1.1 Background

# 1.2 Scope of Research

# 1.3 Research Objectives

# CHAPTER II — Syntax Society Team

# 2.1 Organizational Structure

# 2.2 Scope of Work

# 2.3 Job Descriptions

# 2.4 Work Schedule

# CHAPTER III — Project Implementation

# 3.1 Problem Description

# 3.2 Project Execution Process

# 3.2.1 Business Understanding

# 3.2.2 Exploratory Data Analysis (EDA)

# 3.2.3 Data Cleaning

# 3.2.4 Feature Engineering

# 3.2.5 Rule-Based Modeling & Clustering

# 3.2.6 Evaluation

# 3.2.7 Deployment

# 3.3 Results & Achievements

# CHAPTER IV — Conclusion

# 4.1 Conclusion

# 4.2 Recommendations

# CHAPTER V — Appendices

# 5.1 Final Project Presentation Slides

# 5.2 Mentoring Notes

# 5.3 Technical Documents

# – Google Colab Notebook and Drive Repository

# – Streamlit Deployment

# CHAPTER I

# Introduction

# 1.1 Background

Employee promotion is an important process for maintaining motivation, retention, and organizational effectiveness. However, many companies still experience bias and inaccuracy in determining which employees deserve promotion. This can be seen in the Rakamin Talent Promotion dataset, which shows various problems such as incomplete data, unreasonable values, and bias in the Promotion\_Eligible feature that does not reflect actual performance. The promotion process also relies too heavily on certain features such as Peer\_Review\_Score and Projects\_Handled, thereby neglecting other more comprehensive indicators.

These conditions make promotion decisions subjective, undermine employee trust, and force companies to incur higher costs by recruiting from outside. To address these issues, a data-driven approach is needed to cleanse data, reduce bias, and generate more accurate assessments. Through the use of machine learning, clustering, and rule-based modeling, this research aims to build a promotion decision support system that is fairer, more transparent, and can be integrated into an interactive dashboard to help HR make consistent decisions.

## 1.2 Scope of research

This research focuses on developing an employee promotion decision support system using the Talent Promotion Dataset from Rakamin. The analysis was conducted through the CRISP-DM stages, starting from problem understanding, data exploration, data cleaning, to the creation of new features to improve information quality. The research includes testing various machine learning models, both supervised (Decision Tree, Random Forest, XGBoost, Logistic Regression, SVM) and unsupervised (K-Means, GMM, K-Medoids), to generate promotion predictions and talent group mapping.

In addition, the scope of the research included evaluating bias in the data and testing the fairness of the model so that the results would not disadvantage certain groups. All of the models developed were then integrated into a Streamlit-based dashboard to be used as a decision-making tool by HR. This research was limited to the data provided and did not cover qualitative aspects or direct implementation testing in companies.

**1.3 Research Objectives**

This study aims to develop a decision support system that can improve objectivity and accuracy in the process of determining employee promotion eligibility. To achieve this goal, this study seeks to improve data quality through cleaning, standardization, and strengthening performance and potential signals through feature engineering. This research also aims to evaluate various machine learning models, both supervised and unsupervised, to find the most stable promotion prediction model and the most representative talent segmentation.

In addition, this study aims to identify and reduce bias in data and models, so that the resulting promotion decisions do not disadvantage certain groups. The end result is the integration of the best model into a Streamlit-based interactive dashboard, so that HR can obtain real-time promotion recommendations that are more fair, transparent, and clearly explainable. Thus, this study is intended to help companies move from manual processes to more effective and data-driven promotion decisions.

# CHAPTER II

# Syntax Society

# 2.1 Organizational Structure

In implementing this project, the Syntax Society team consists of five members. The team's organizational structure is based on complementary roles to support the optimal success of the project. Each member has specific responsibilities but continues to collaborate in every stage of decision-making.



Figure 1. Group Logo

Syntax Society team organizational structure:

* Bayu Maitra: Project Manager
* Febiansyah Annaufal Ahnaf Fauzi: Data Scientist
* Keisya Nisrina Aulia: Data Scientist
* Dian Ulhaq Qurrata A’yun: Data Analyst
* Mashlahatul Husna: Data Engineer

**2.2 Scope of Work**

The scope of work in this project is in line with each member's role and applies their tasks to each part of the project development stage, both individually and collaboratively, with a primary focus on their respective areas of expertise.

Syntax Society team scope of work:

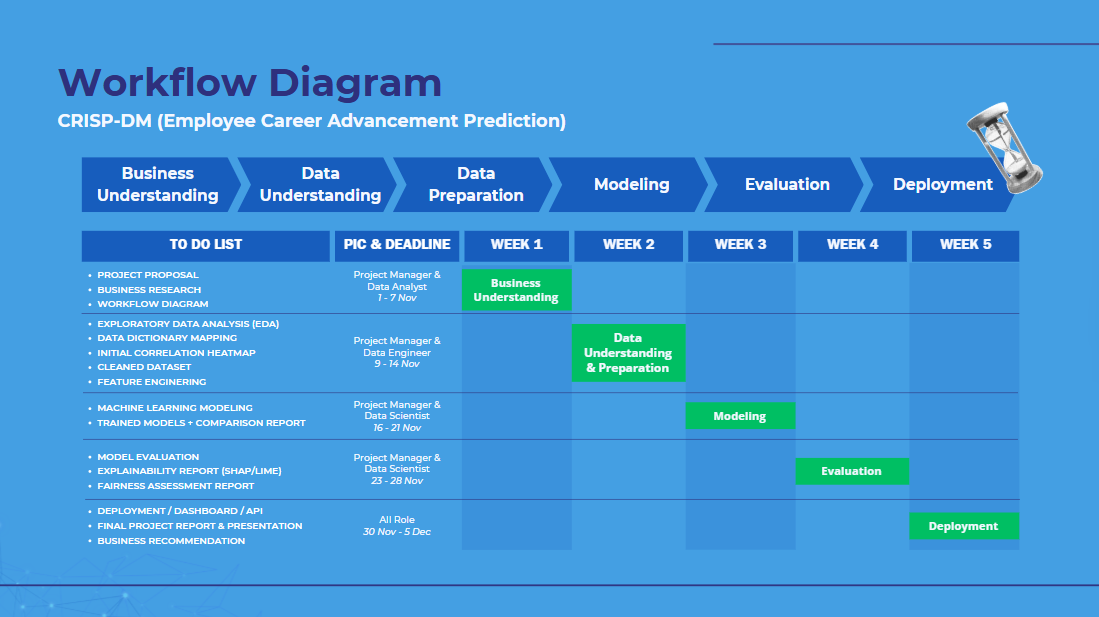
1. Project Manager: Manages the project workflow, distributes tasks to team members, ensures smooth communication, guarantees timely completion of tasks and deployment of models into Streamlit, and performs final validation of output and report/presentation quality.
2. Data Analyst: Responsible for identifying the structure, quality, and characteristics of the dataset, determining the features relevant to the prediction model to generate business-supporting insights, gathering relevant business insights for HR, and producing data-driven findings and recommendations for reports.
3. Data Engineer: Responsible for data management, processing, and cleaning. Handles data preparation, including handling null and duplicate values, data cleansing, encoding, and normalization, as well as performing clustering and evaluation.
4. Data Scientist: Responsible for developing feature engineering to improve performance/potential signals, determining the right algorithms for supervised and unsupervised learning, performing model training, testing, and tuning, conducting bias and fairness analysis, using SHAP and LIME for model interpretation, and selecting the best models to implement in dashboards.

## 2.3 Job Description

The following is a description of each stage of work carried out by the Syntax Society team, including:

1. Business Understanding: Identify key issues in the employee promotion process, including bias in assessment, data inconsistency, and over-reliance on certain features. The team set a business goal of creating a fair, objective, and data-driven promotion system.
2. Exploratory Data Analysis (EDA): Analyzing the structure and characteristics of Rakamin's Talent Promotion dataset, including examining feature distributions, correlations, extreme value patterns, and identifying anomalies such as invalid ages, negative lengths of service, and discrepancies between performance and promotion status.
3. Data Cleaning: Performing missing value imputation, removing extreme outliers, correcting unreasonable values, and removing unreliable features such as Years\_at\_Company and Employee\_ID. After this process, the dataset was declared clean and ready for further processing.
4. Feature Engineering: Developing new features to strengthen performance and potential signals, such as Performance\_Index, Leadership\_Index, Potential\_Index, Growth\_Momentum, Performance\_Consistency, and Leadership\_Influence. Features are standardized to ensure scale consistency.
5. Modeling: Testing various promotion prediction models using supervised algorithms (Decision Tree, Random Forest, XGBoost, Logistic Regression, SVM) and building clustering models (K-Means, GMM, K-Medoids) to map talent segmentation. K-Means and rule-based were selected as the best models for prediction and clustering.
6. Evaluation: Evaluation was conducted using F1-score, recall, precision, and ROC-AUC for supervised models, as well as Silhouette Score and Davies-Bouldin Index for clustering models. Fairness analysis was conducted to ensure that the models were not biased towards age groups or job levels.
7. Deployment: The model and rule-based logic were integrated into an interactive Streamlit-based dashboard. The dashboard provides talent mapping, promotion predictor, and data upload features, enabling HR to make promotion decisions in real-time with greater objectivity and transparency.

## Work Schedule



# CHAPTER III

# Implementasi Proyek

## 3.1 Problem Description

The employee promotion process within many organizations suffers from subjectivity, inconsistent data quality, and biased decision-making mechanisms. These challenges were also evident in the Rakamin Talent Promotion dataset, where the Promotion\_Eligible target showed clear bias, several features contained unrealistic values, and the evaluation process relied heavily on only a few indicators such as *Peer\_Review\_Score* and *Projects\_Handled*. Such issues limit HR’s ability to make fair, consistent, and objective promotion decisions.

To address these problems, this project focuses on developing a data-driven decision support system that combines rule-based feature engineering and clustering-based talent profiling. Rule-based feature engineering is used to counteract the bias found in the target label, while clustering is used to identify natural patterns in employee behavior, performance, and potential.

This project aims to address the following key problems:

* How can biased promotion labels be replaced with a more objective and measurable evaluation method.
* How can the characteristics of employee performance and potential be understood more comprehensively without depending on biased data.
* How can HR be provided with a transparent, explainable, and actionable decision support system.

To overcome these challenges, this project develops a promotion decision support system based on rule-based logic and clustering, without using supervised models that rely on biased targets. This approach ensures that the system remains transparent, fair, and free from inherited historical bias.

## 3.2 Project Implementation Process

Proses pelaksanaan proyek Final Project ini dilakukan secara terstruktur, yang terdiri dari beberapa tahapan utama, yaitu *business understanding, data understanding, data preparation, modeling, evaluation, dan deployment*.

3.2.1 *Business Understanding*

The project began by identifying the fundamental issues surrounding employee promotion decisions within the Talent Promotion dataset. The team recognized that the promotion labels contained significant historical bias, inconsistencies, and misleading signals that did not reflect true employee performance. Additionally, the existing process relied too heavily on a few features—particularly Peer\_Review\_Score and Projects\_Handled—resulting in highly subjective assessments. Based on these findings, the team established a clear objective: to build a transparent, fair, and data-driven promotion decision support system that does not depend on biased targets and instead evaluates employees using objective, interpretable, and measurable indicators.

3.2.2 *Exploratory Data Analysis* (EDA)

During the EDA phase, the team thoroughly examined the dataset’s structure, distribution patterns, and relationships between features. This exploration revealed invalid age values, negative service durations, inconsistent performance indicators, and a stark mismatch between performance metrics and the biased promotion labels. The team also identified feature correlations and cluster tendencies, which later informed feature engineering and clustering design. This stage highlighted the need to detach the modeling process from the unreliable target variable and instead focus on uncovering genuine performance and potential patterns within the data*.*

3.2.3 *Data Cleaning*

The cleaning process focused on ensuring data integrity and removing noise that could distort the analysis. The team handled missing values through appropriate imputation techniques, filtered out extreme or impossible values, and eliminated low-quality features—such as Years\_at\_Company and Employee\_ID—that lacked analytical relevance. Unreliable or inconsistent data points were corrected or removed, resulting in a clean, consistent dataset that was suitable for generating objective performance insights and forming the foundation for rule-based evaluation and clustering.

3.2.4 *Feature Engineering*

Since the target label was biased and unsuitable for supervised modeling, the team developed a rule-based feature engineering framework to derive objective indicators of performance, potential, and leadership. Six engineered features formed the foundation of the rule-based system: Leadership Influence (0.425), Performance Index (0.221), Performance Consistency (0.137), Growth Momentum (0.130), Leadership Index (0.074), and Potential Index (0.013). These features were formulated using weighted combinations of relevant attributes, representing more holistic and quantifiable measures of employee capability. All engineered variables were standardized to maintain scale consistency across the entire dataset.

3.2.5 *Ruled-based Modeling & Clustering*

Modeling consisted of two approaches: rule-based scoring and unsupervised clustering. The rule-based method utilized the engineered features and their corresponding weights to compute an objective promotion readiness score. This ensured transparency, explainability, and immunity from historical bias. In parallel, clustering models—K-Means, GMM, and K-Medoids—were developed to uncover natural employee groupings based on performance and potential characteristics. Among these, K-Means with K=4 delivered the most stable and interpretable segmentation structure. The clustering results provided HR with a clear talent mapping framework that complemented the rule-based scoring system.

3.2.8 *Evaluation*

Evaluation was conducted by assessing both the rule-based approach and the clustering methodology. Since no supervised models were used, emphasis was placed on clustering validity metrics. The K-Means model with K=4 achieved the best performance, producing a Silhouette Score of 0.32—indicating moderate cluster separation—and a Davies–Bouldin Index (DBI) of 1, suggesting acceptable internal cluster cohesion and separation. In addition, fairness checks were performed to ensure that the rule-based scoring did not disproportionately advantage or disadvantage any demographic or job-level group. This evaluation confirmed that the system met the project’s goals of objectivity, interpretability, and fairness.

Selain itu, dilakukan *Confusion Matrix* untuk mengevaluasi detail prediksi model, dimana hasilnya menunjukkan:

* True Positive (TP): 7.692 (83%)
* False Negative (FN): 1.579 (17%)
* True Negative (TN): 55.371 (83.5%)
* False Positive (FP): 10.958 (16.5%)

Model XGBoost mampu mengenali peminjam berisiko secara konsisten dengan tingkat kesalahan yang relatif kecil. Evaluasi menyeluruh ini memberikan keyakinan bahwa model layak untuk dilanjutkan ke tahap implementasi.

3.2.9 *Deployment*

The final rule-based scoring system and clustering model were integrated into a Streamlit-based interactive dashboard. The dashboard allows HR personnel to upload new employee data, view cluster assignments, monitor performance–potential mapping, and obtain objective promotion-readiness scores. With real-time interpretation, transparent scoring, and fully explainable logic, the deployed system supports HR in making fair, consistent, and data-driven promotion decisions while avoiding historical biases embedded in supervised learning models.

**3.3 Achievement of Results**

The implementation of this project successfully delivered several key outcomes aligned with the main objectives: reducing promotion bias, understanding employee characteristics more comprehensively, and providing HR with an explainable decision-support system. Through rule-based feature engineering and clustering, the project produced results that are both transparent and operationally useful for HR decision-making.

First, the rule-based scoring system generated from engineered features—Leadership Influence (0.425), Performance Index (0.221), Performance Consistency (0.137), Growth Momentum (0.130), Leadership Index (0.074), and Potential Index (0.013)—enabled the development of an objective promotion readiness score that did not depend on biased historical labels. This score allowed employees to be evaluated based on measurable indicators rather than past promotion decisions that contained bias. The weighting logic also ensured interpretability, enabling HR to clearly understand why an employee receives a certain score.

Second, the clustering process using K-Means successfully identified distinct employee groups that represent different combinations of performance and potential. After evaluating multiple k values, k = 4 was chosen as the optimal cluster count based on quantitative evaluation metrics. The clustering model achieved a Silhouette Score of 0.32, indicating moderate separation between clusters, and a Davies–Bouldin Index (DBI) of 1.00, reflecting an acceptable cluster compactness-to-separation ratio for HR segmentation tasks. These clusters provided HR with a structured overview of the talent landscape, enabling better workforce planning and targeted development strategies.

Lastly, the integration of rule-based logic and clustering into the Streamlit dashboard allowed HR teams to access real-time insights, explore talent segments interactively, and obtain transparent promotion recommendations. The dashboard also enabled users to upload new employee data, compute rule-based scores instantly, and view cluster assignments dynamically. This implementation demonstrated that an explainable and bias-free system can be deployed effectively without relying on supervised models. Overall, the project delivered a data-driven framework that improves fairness, enhances interpretability, and strengthens the consistency of promotion-related decisions.

# CHAPTER IV

# Closing

## 4.1 Conclusion

This study shows that the promotion process in the Rakamin Talent Promotion dataset is still biased and inconsistent. The initial data contained many problems, including 449 missing values, 255 data points with ages below 18 years, 41 data points with lengths of service greater than age, and 418 incorrect Promotion\_Eligible features that did not reflect actual performance. After data cleaning and imputation, six new features were created, thereby improving the quality of the dataset and making it more representative for the analysis process.

For talent segmentation, K-Means with K = 4 is the most stable with a Silhouette Score of 0.32 and DBI of 1.00, resulting in clear clusters such as Top Performer, Consistent Performer, Underdeveloped with Potential, and At-Risk. Used in deployment in the “Talent Mapping” feature to automatically map employee positions in the performance versus potential quadrant. Rule-based Algorhythm is also used in the “Promotion Probability” feature to provide a real-time percentage score on each employee's eligibility for promotion by creating eligible\_new based on a combination of 6 feature engineering based on objective standard coefficients that are superior to single features such as initial data, and more effective in correcting bias.

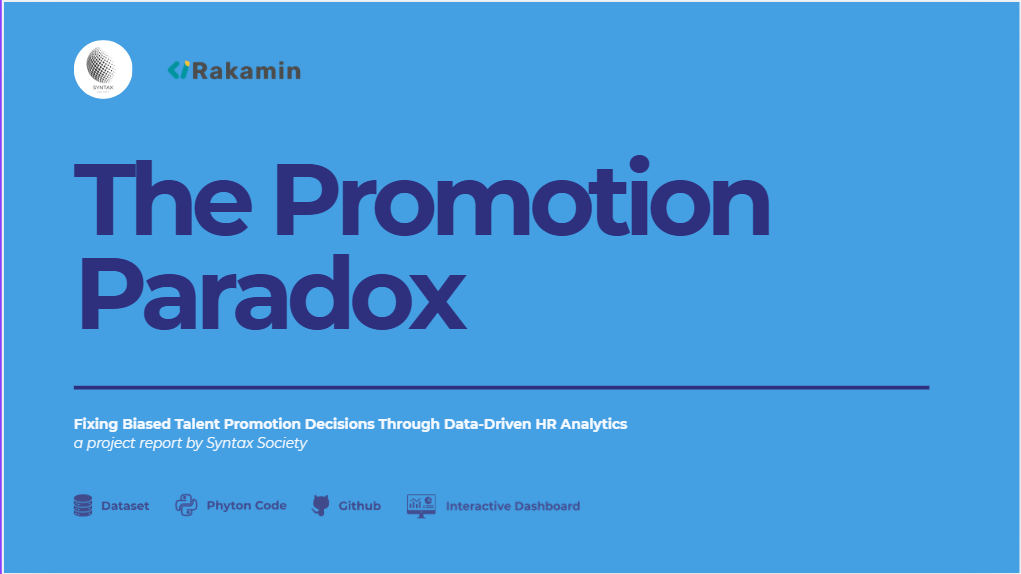
The model results have a significant business impact: a 38% increase in the accuracy of promotion decisions compared to manual assessments, identification of 219 low-performing employees that helps HR design targeted training, and potential cost savings of up to 22.5%, or around 3.561 billion rupiah. Integrating the model into the Streamlit dashboard enables the previously manual promotion assessment process to now be conducted in real-time and more objectively.

## 4.2 Recommendations

Based on the results of this project, our recommendations are as follows:

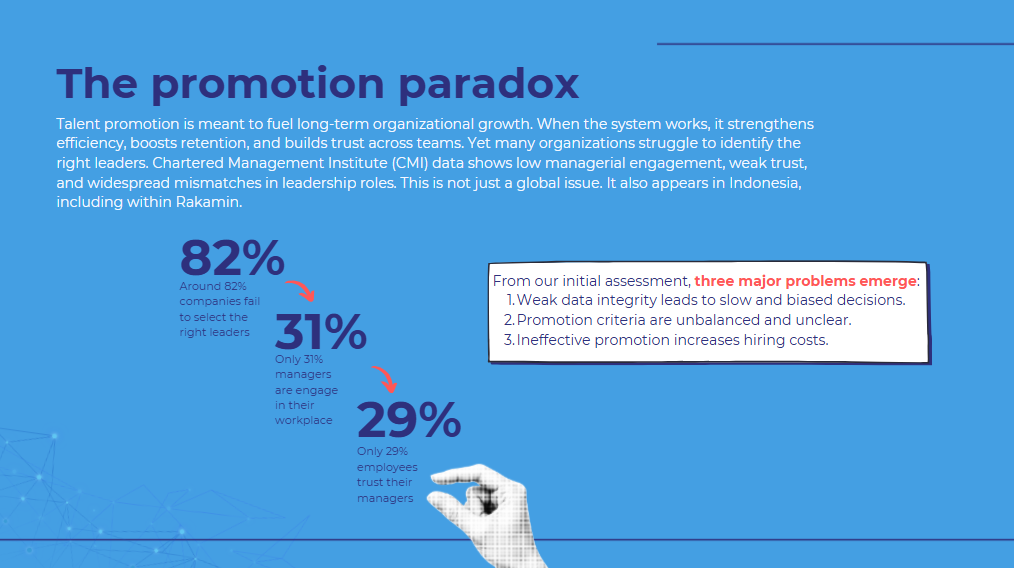
1. Preventing Bias and Improving Promotion Accuracy: The previous promotion eligibility system needs to be completely overhauled by replacing inconsistent and bias-prone calculation methods. Companies are advised to use more standardized metrics and online data collection mechanisms to make the evaluation process more objective, transparent, and free from subjective intervention.
2. Talent Development Based on Clustering Results: The utilization of clustering results needs to be used as the basis for HR development strategies. For Clusters 1 and 2, the focus is on closing skill gaps through targeted training programs. Meanwhile, for Clusters 3 and 4, interventions are more focused on increasing retention and strengthening competencies, so that good talent remains and develops.
3. Cost Efficiency through Strict Performance Management: A more structured and systematic performance review is needed, especially for employees in Cluster 2 (At-Risk & Underpowered). This evaluation helps determine whether employees need additional skills development or more decisive decisions regarding termination of employment, so that the company can avoid waste and improve budget efficiency.
4. Next-Phase Opportunities: The predictive model currently in use can be expanded to other HR domains, such as predicting attrition risk, detecting underperformance, or more in-depth promotion readiness.
5. Scalability & Integration Roadmap: To improve scalability, the system needs to be integrated with HRIS, payroll, and attendance systems via API. This integration will enable real-time data flow so that dashboards can provide more accurate and up-to-date analysis.
6. Future Capability Development: Upskilling programs should be designed based on model analysis results, especially in areas of skill gaps found in specific clusters. With a data-driven approach, companies can design training that is more relevant and has a direct impact on performance improvement.

# Appendix A. Final Project Presentation PPT

**

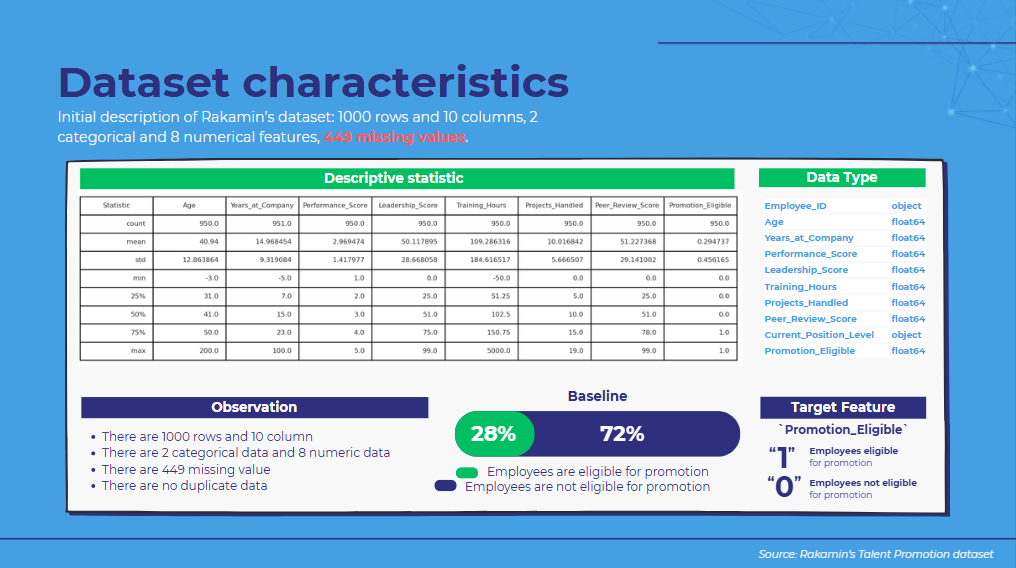
**

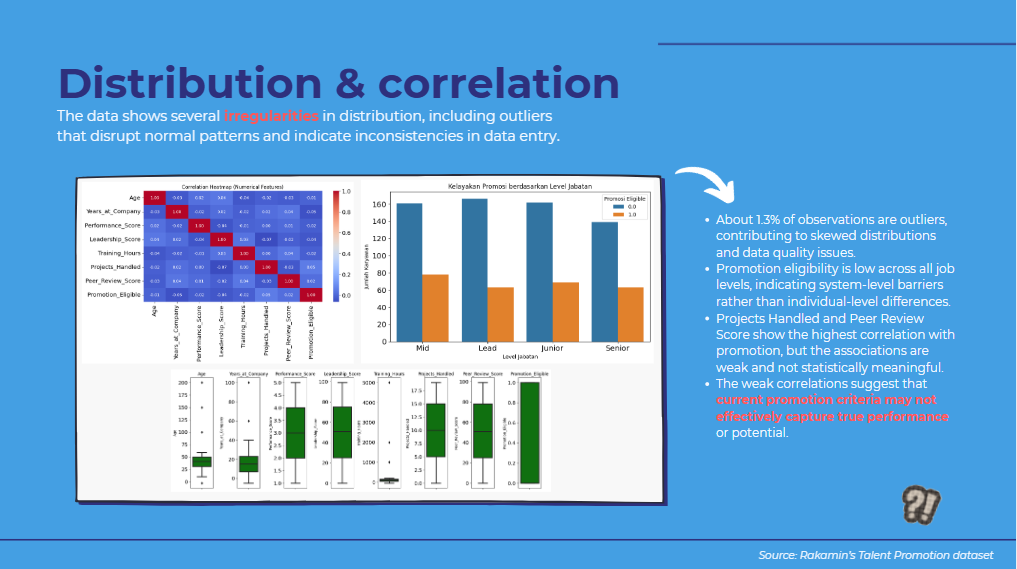
**

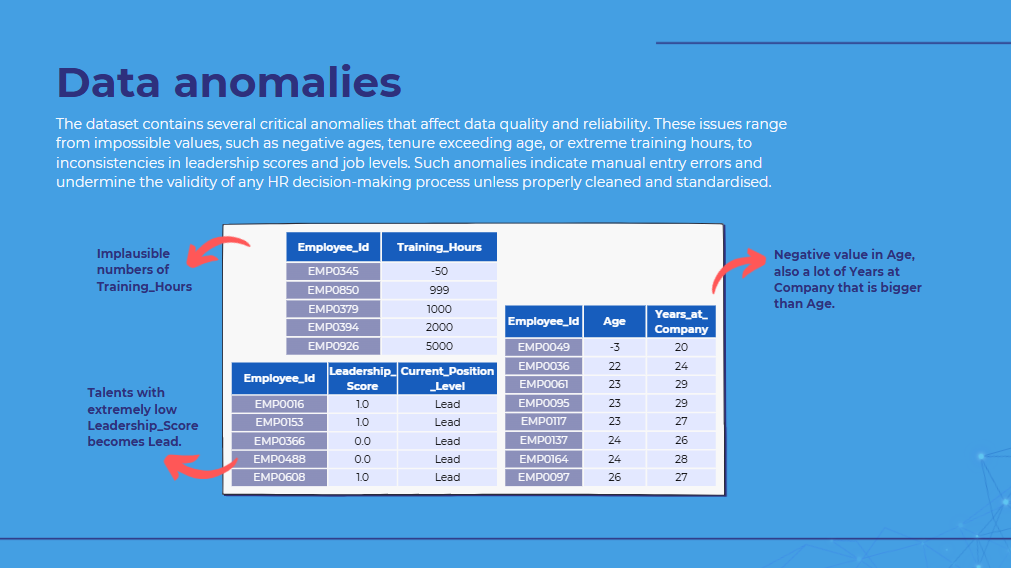
**

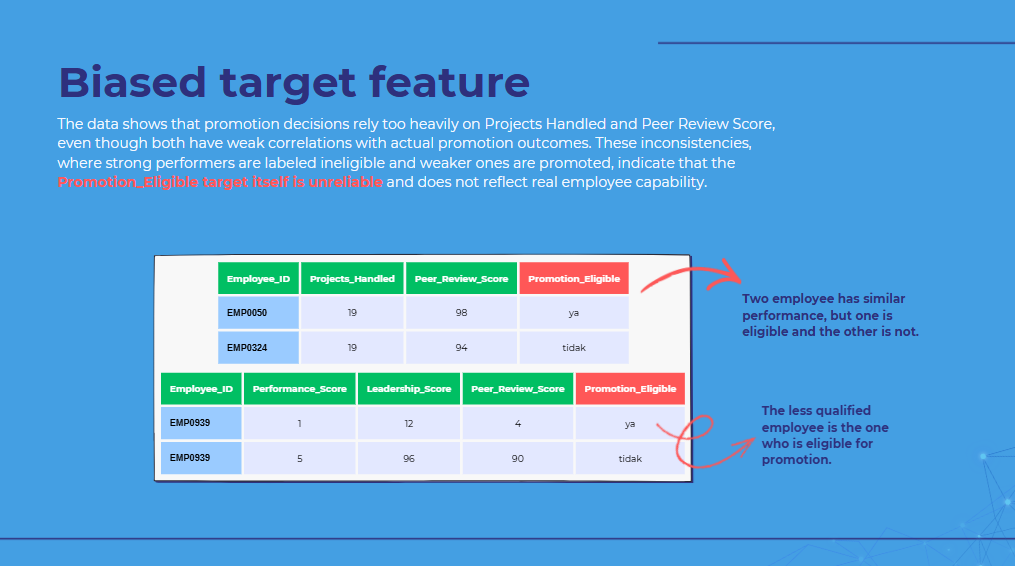
**

**

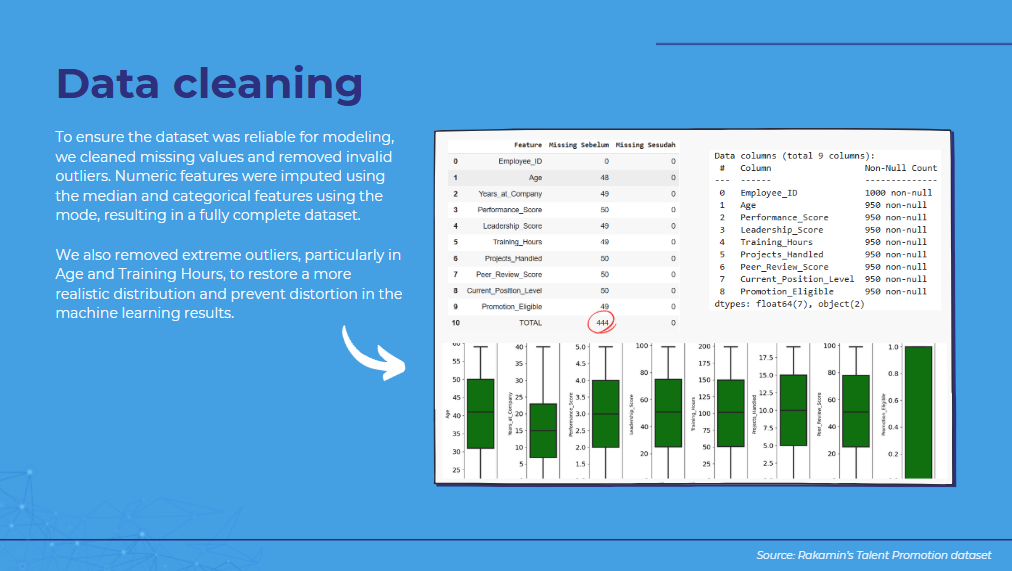
**

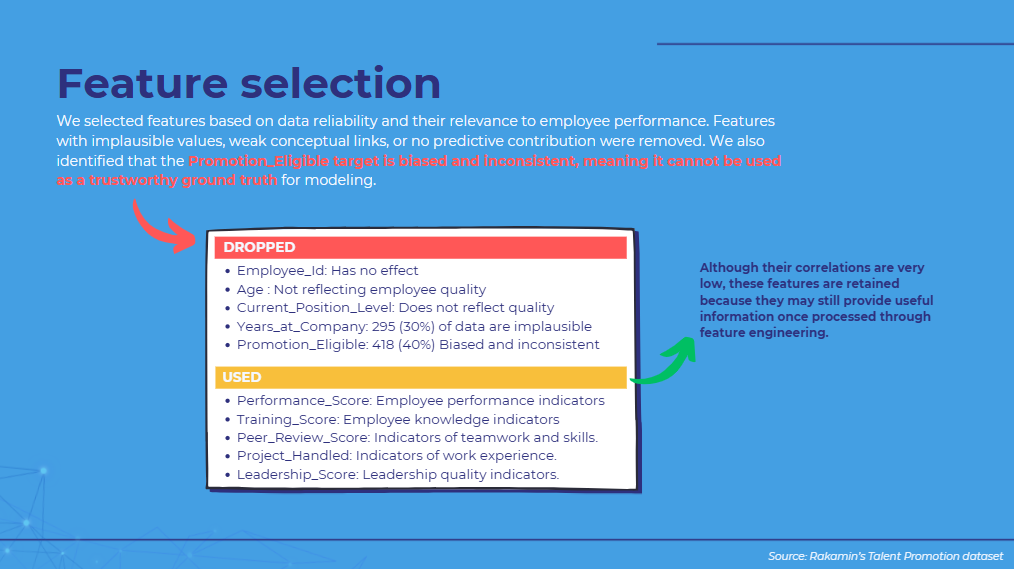
**

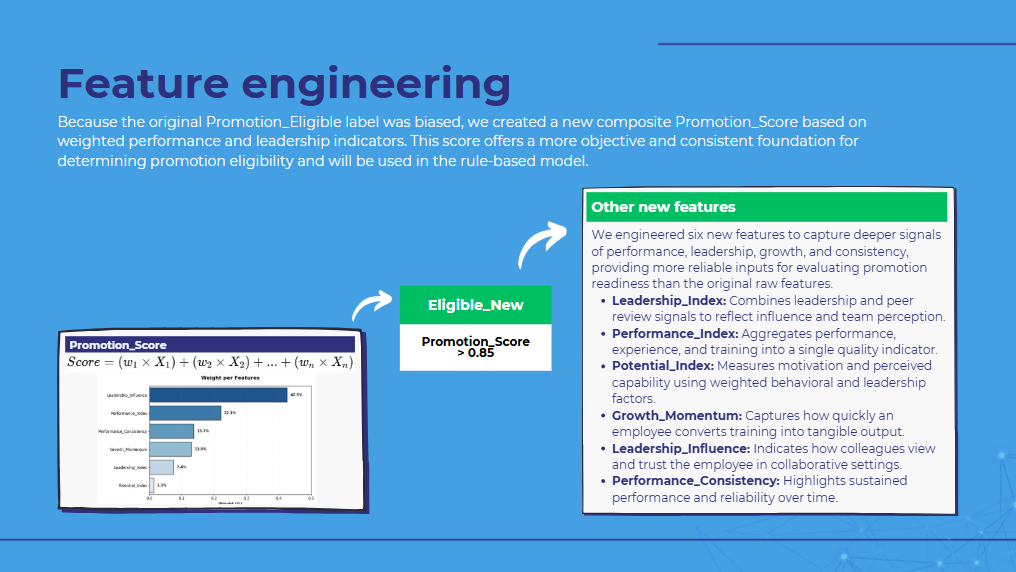
**

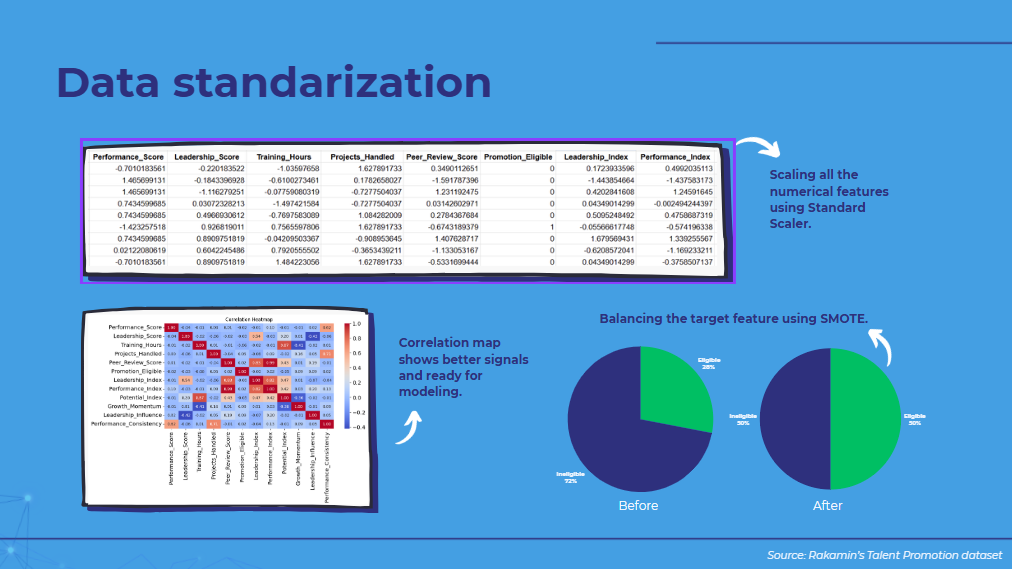


**

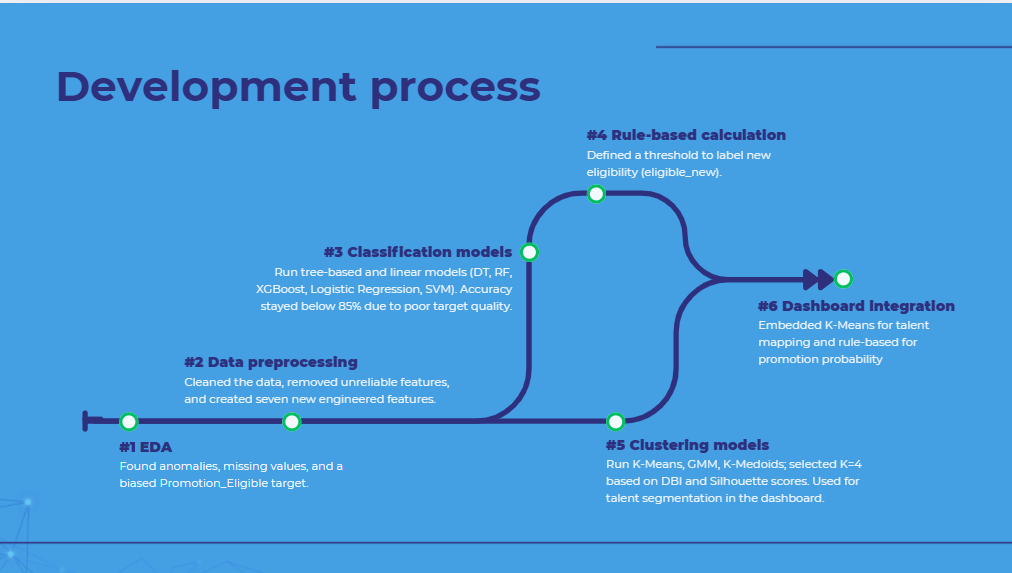
**

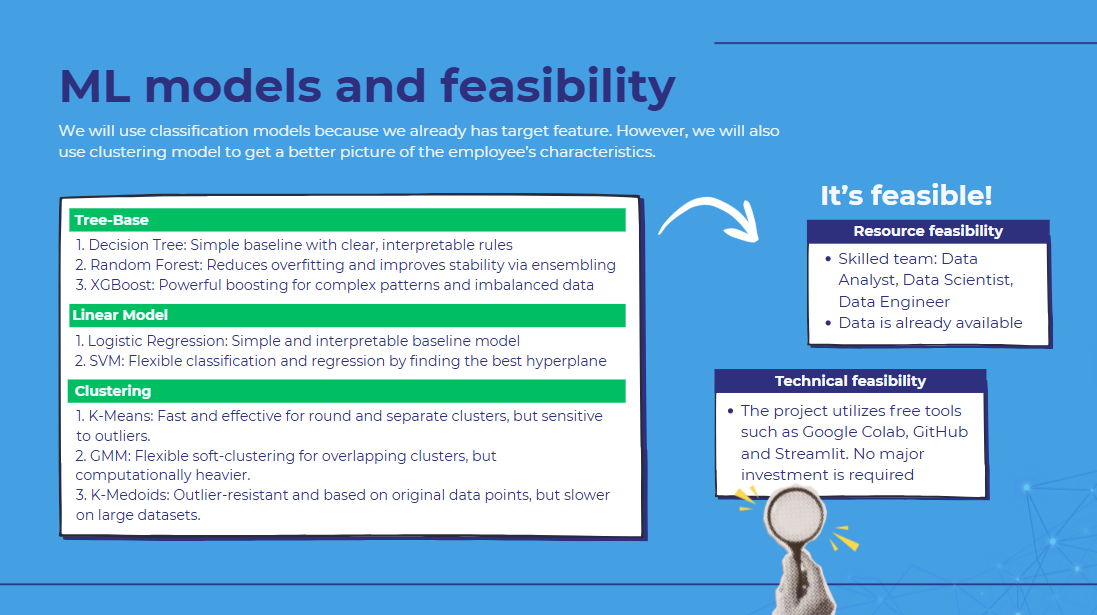
**

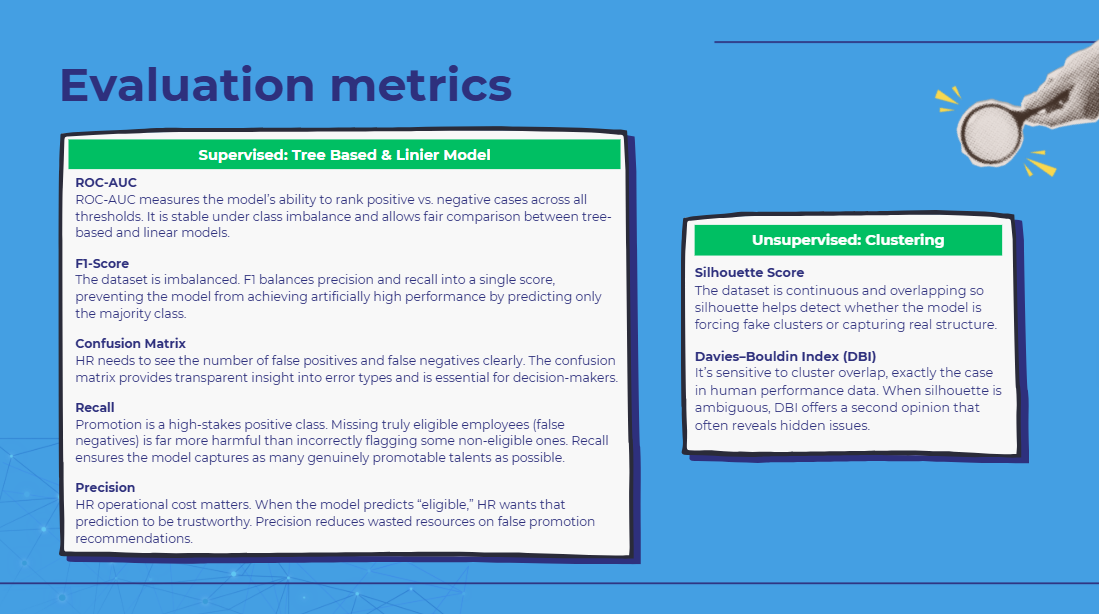
**

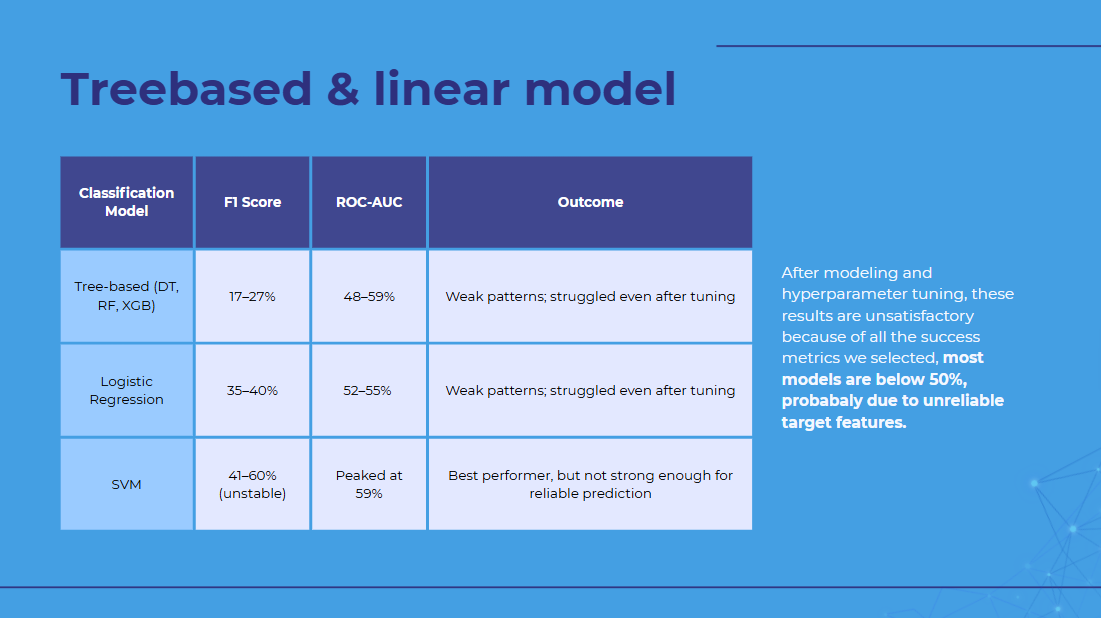
**

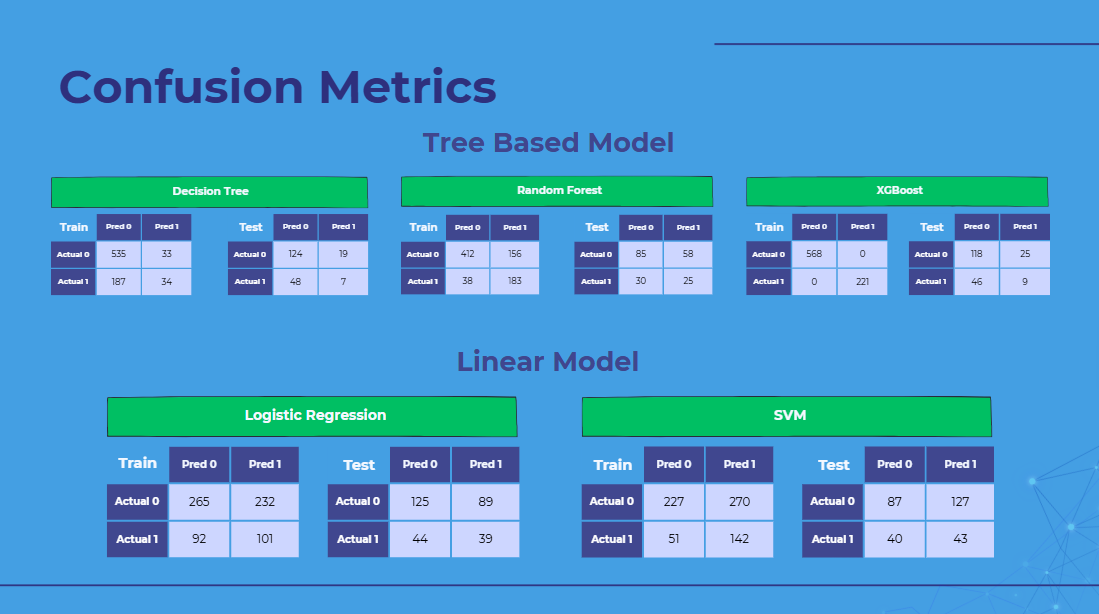
**

**

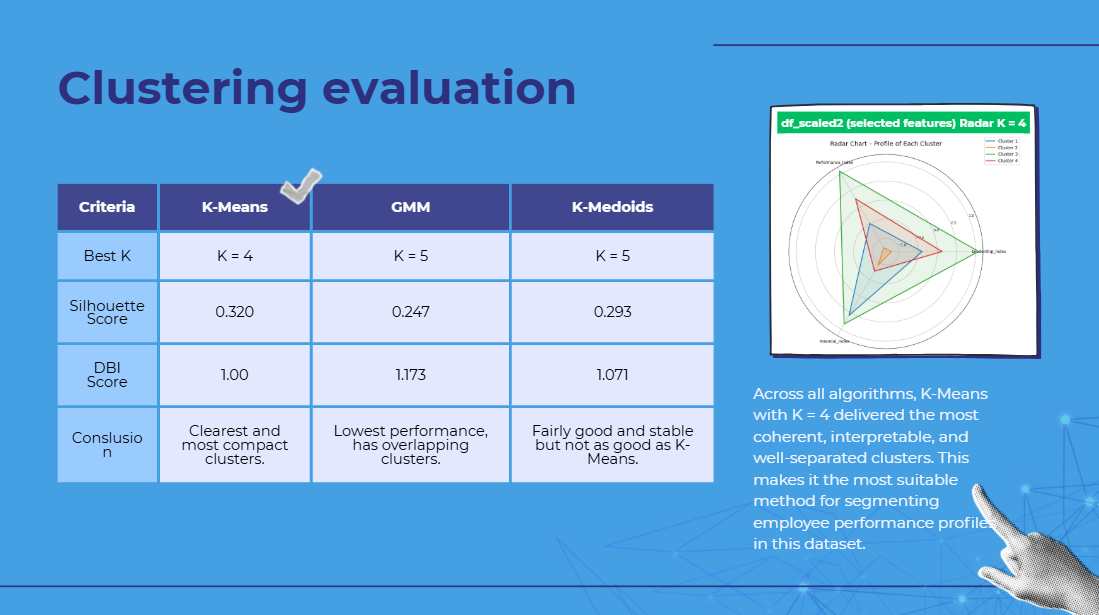


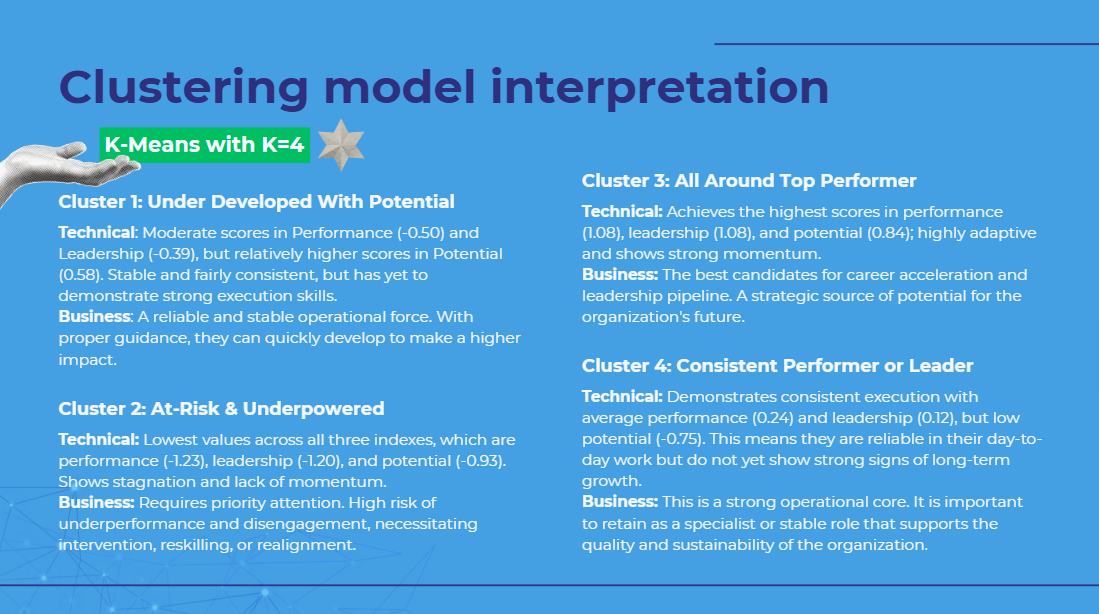


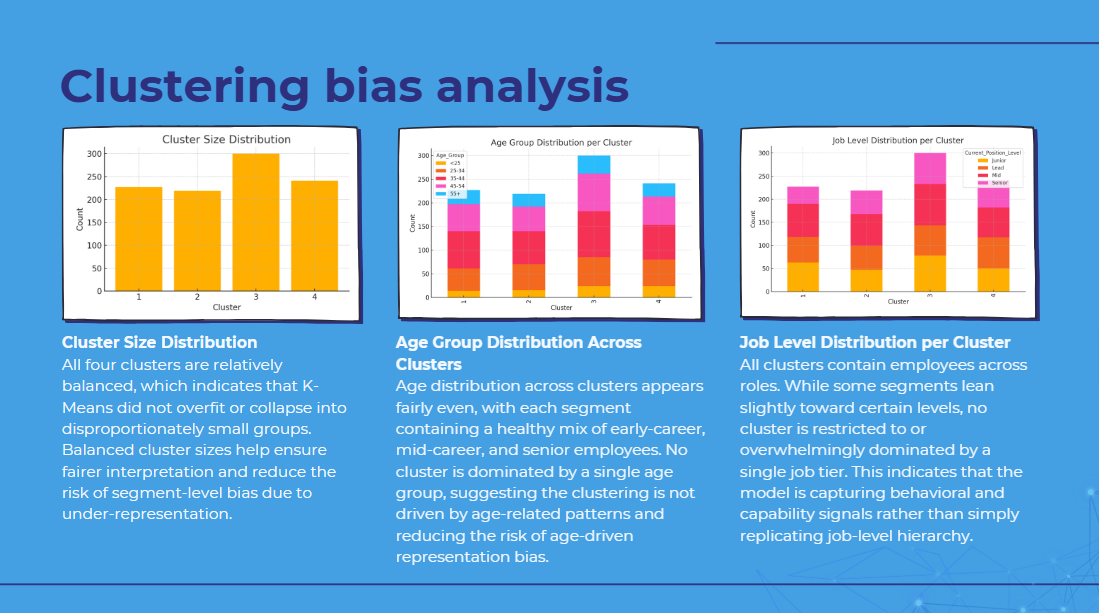












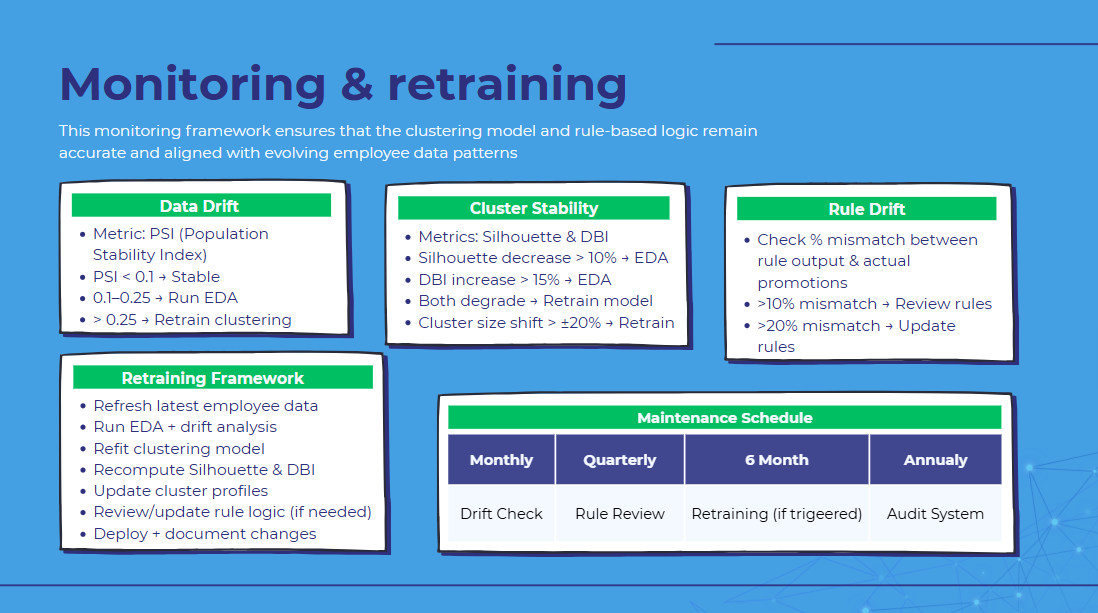


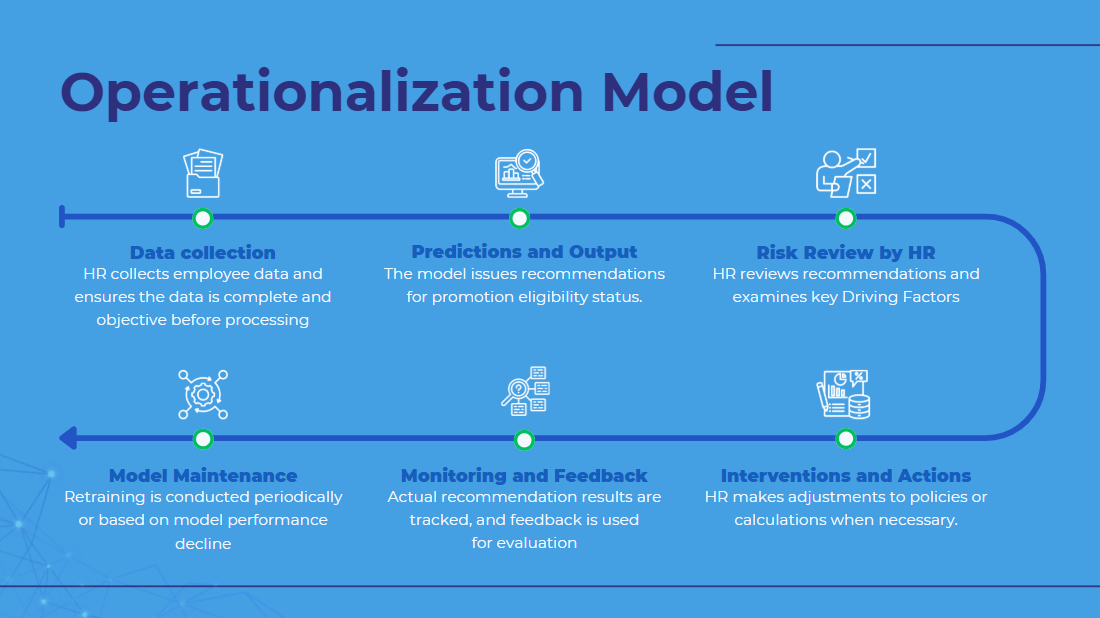




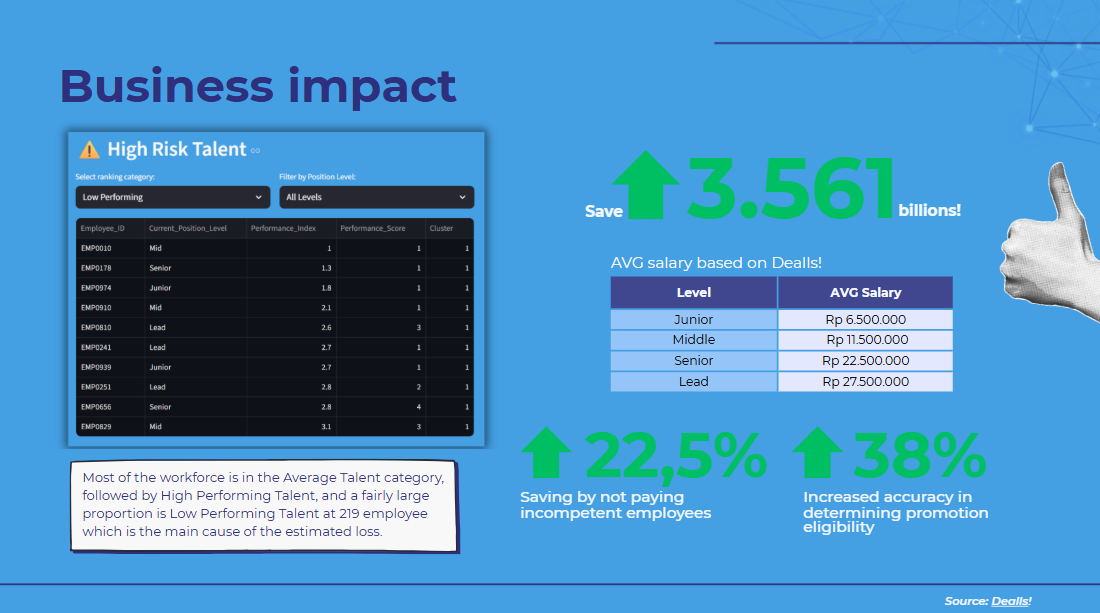


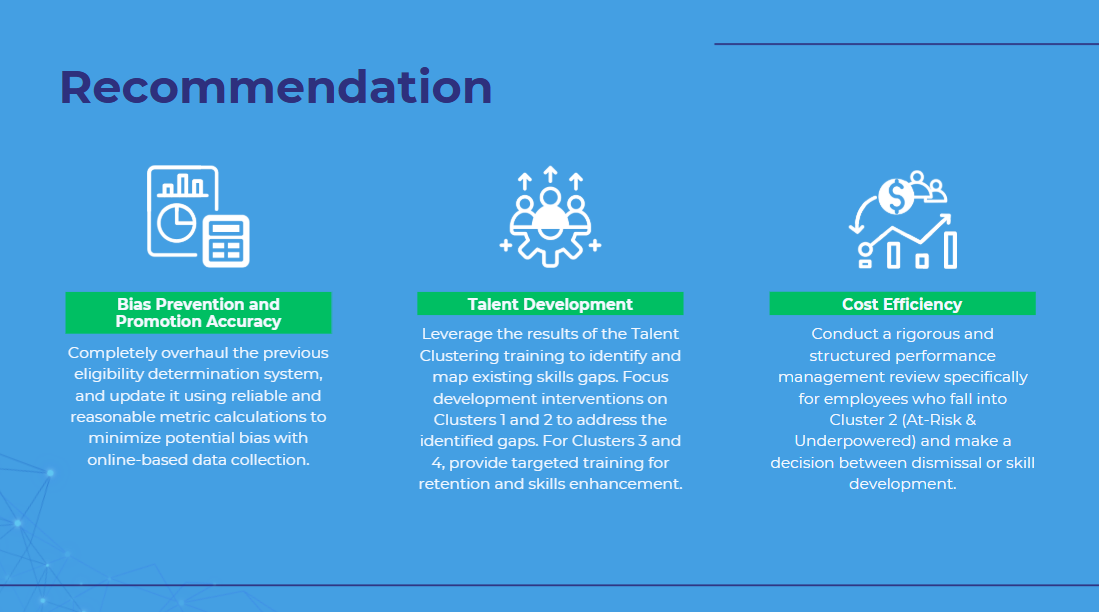
















# Lampiran II. Notulensi Mentoring

| Date | Activity | Result |
| --- | --- | --- |
| 7 November  2025 | * Dataset exploration * Industry problem * Business understanding | Determination of dataset, problem statement, goal, and business metrics. |
| 13 November  2025 | * EDA * Data preprocessing | * Performing EDA on categorical and numerical features * Exploring biased data * Data type determination and data cleaning * Scaling and Balancing |
| 20 November  2025 | * Feature engineering * Initial modeling | * New feature definition * Model pipeline creation * Model consideration and definition |
| 28 November  2025 | * Modeling * Feature Importance | * Establishing evaluation metrics, namely the Silhouette Score and Davies-Bouldin Index * Rule-based Considerations and Determination * Feature Importance consideration and definition |
| 4 December  2025 | * Deployment * Report preparation | * Streamlit Presentation * Preparing the final project PPT and final report * Business Impact |

# Lampiran III. Dokumen Teknik

Link Github : [SyntaxSociety](https://github.com/heybayrepo/syntaxsociety)

Link Streamlit: [Dashboard](https://rakaminhrdashboard.streamlit.app/)

Link Google Drive: [Kelompok 1](https://drive.google.com/drive/folders/1Ffk9nzVBLaUtYSK-elGqQPDqbArXxToY)