GROUP 3

1.0 Introduction

The Human Resource (HR) department plays a crucial role in managing the workforce of an organization. Typically, HR is involved in various operations such as recruitment, employee training and development, fostering relationships with employees, and conducting performance appraisals. A key aspect of HR's mandate includes overseeing promotions within the company, as promotions and demotions can have a significant impact on morale, retention, and productivity. According to Indeed, an employee who has been promoted and changes jobs may expect an average pay raise between 10% and 20% (Bert, 2024).

Position promotion is an integral part of career development, as it rewards workers with more responsibilities and better pay. Based on organizational policies, these promotion-based decisions are taken on different aspects. These can be the length of service, experience, seniority, performance, etc (Barman, 2024). However, the job environment in Malaysia commonly faces bias and discrimination, which can impede the fairness of promotions. This bias can lead to negative outcomes, including a toxic culture, reduced employee satisfaction, and a high turnover rate. The repercussions of this prejudice are significant, as the issue not only affects the employees themselves but also impacts the entire enterprise.

2.0 Problem Statement

In a competitive business environment, the identification and promotion of talented employees are crucial for organizational success. However, the challenge of identifying high-potential employees based on their performance and abilities is exacerbated by the presence of biases in the Malaysian working environment. This scenario often results in discrimination and prejudice persisting between employees and higher management with different ethnicities, religions, skin color, and gender.

To further support this statement quantitatively, the State of Discrimination Survey Malaysia conducted by Architects of Diversity in 2023 revealed that discriminative experiences among Malaysians ranked second (30%) during the job search process and third, with 29% experiencing discrimination at work. These findings underscore the prevalence of discrimination and its detrimental impact on fostering toxic working environments in our multicultural country (Persatuan Pendidikan Diversiti, 2023).

Through this research, our aim is to enhance the efficiency and transparency of talent management processes by automating the promotion prediction process to identify employees most likely to be promoted. We hope to mitigate the impact of discrimination and biases in the Malaysian workforce, fostering a more inclusive and merit-based system for talent recognition and advancement.

3.0 Objectives

To solve the issue stated above, we had established several research objectives that we aim to achieve by the end of this project:

1. Analyze the performance of employees through various visualisation tools.
2. Develop a classification model to identify potential workers for position promotion based on their abilities and performances.
3. Identify the most important factor that contributes to position promotion for the company.

4.0 Methodology

(i) Data Collection

The dataset was retrieved from a public datahack contest held on the Analytics Vidhya website (**Link**). It contains various features related to employee details, performance, training evaluations and the state of whether the employee is recommended for promotion as the target variable. The dataset has 54808 rows and 13 columns, providing a large sample size for robust statistical analysis and model training.

(ii) Data Cleaning

Before performing data visualization, the dataset was cleaned by checking for null values and any duplicated rows with Python. Two columns were identified with null values: "previous_year_rating" and "education." To handle these, the "previous_year_rating" column was filled with the value 0, indicating that the employee is working for the first year and does not have any previous year rating. For the "education" column, a clustering algorithm was employed. Specifically, Principal Component Analysis (PCA) was applied for dimensionality reduction on all numerical features and then k-means clustering was used to impute the missing values in the "education" column.

(iii) Data Visualisation

We utilized 3 tools to visualize employee performance namely Python, PowerBI, and Tableau. This step included conducting descriptive analysis on summarized employee statistics, distribution of variables, and examining the relationship between features and the target variable "is_promoted" to analyze employee performance within the company. For PowerBI and Tableau, the generated content was visualized into a dashboard report for easier navigation and comprehensive analysis. Through visualization, we had a clearer understanding of the dataset's characteristics which allowed us to identify patterns and trends in employee performance and provide further insights for analytics and model training. Lastly, we conducted a comparative analysis of the pros and cons of the 3 visualization tools used.

(iv) Data Preprocessing

Several preprocessing steps were carried out using Python Library to prepare the dataset for classification tasks. First, extreme outliers were removed using the z-score method to ensure the robustness of subsequent analyses. Following this, label encoding was applied to the "education" column to map categorical values to numerical representations, while one-hot encoding was implemented on the "recruitment_channel" column to transform categorical data into binary vectors. Feature selection was conducted to enhance the predictive performance of the model by dropping columns deemed irrelevant for the classification task, which are "employee_id", "region", "department", "gender" and "age".

To address the issue of class imbalance in the classification task, two separate dataframes were created to compare the performance of different approaches in handling this issue. Firstly, an oversampled dataframe (oversampled_df) was generated by up-sampling the class with lower samples using the SMOTE (Syntheticers to Minority Over-sampling Technique) library, specifically by up-sampling class '1' to match the sample size of class '0' (where "0" ref not promoted and "1" refers to promoted). SMOTE works by selecting minority observations that are similar to each other and drawing a line between the examples in order to create new synthetic samples (Hoffman, 2021). Conversely, an undersampled dataframe (undersampled_df) was created by down-sampling the class with higher samples using the Pandas library, in which class '0' was down-sampled to match the sample size of class '1'. These two approaches allow for a comparative analysis of handling class imbalance and its impact on classification performance. The dataset will then be split into 80:20 ratio for training and testing.

(v) Data Modelling

For model training, we implemented 5 different classification algorithm models using Python for each dataframe which are "Logistic Regression," "Decision Tree," "Random Forest," "K-nearest Neighbors," and "Gradient Boosting." The objective was to evaluate the performance of different models on the same task and identify the most suitable model based on their performance metrics generalized on unseen testing data. Besides comparing internally between different models, we also compared the overall performance of classification models on the oversampled and undersampled data to identify any significant differences in the results.

(vi) Model Evaluation

For each model trained on both dataframes, we evaluated their performance on testing data using key evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics were organized into a table format to facilitate easy comparison across models and dataframes. Additionally, we visualized the confusion matrix for each model, providing insights into the distribution of true positive, true negative, false positive, and false negative predictions. Furthermore, we created ROC-AUC curve graphs to visually represent the trade-off between true positive rate and false positive rate across different classification thresholds.

By comparing the results of each model across both dataframes, we aimed to identify the model with the best performance for further hyperparameter tuning. This comprehensive evaluation process enabled us to make informed decisions regarding the selection of the most effective classifier for the classification task at hand.

(viii) Feature Importance

Finally, we visualized the feature importance of the tuned model to identify the most significant factors contributing to position promotion within the company.

5.0 Results & Discussion

(i) Visualisation of employee performance using Python Libraries (Matplotlib and Seaborn), PowerBI and Tableau. (PowerBi, Python, Tableau)

In this project, we employed several visualization tools, including the Python libraries Matplotlib and Seaborn, as well as PowerBI and Tableau. These tools were important in presenting clear insights and trends regarding employee performance within the dataset. The visualizations generated include the number of employees, average training scores, average previous year ratings, employee age distribution, count of the target variable "is_promoted," and the relationships between length of service, education, KPIs met, and awards won by the count of "is_promoted." For PowerBI and Tableau, all visualizations can be filtered using department and recruitment channel filters for further analysis.

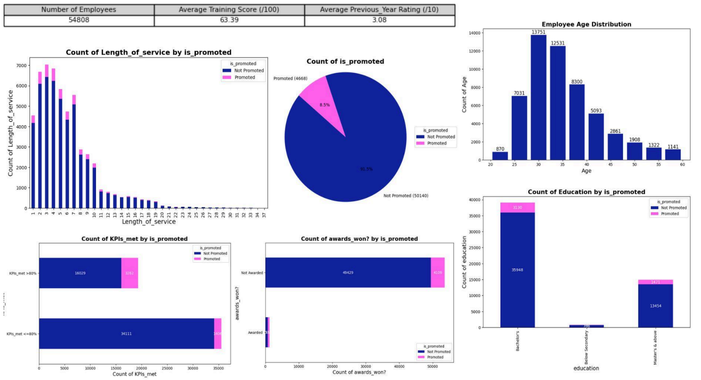


Figure 1: Python Visualisation

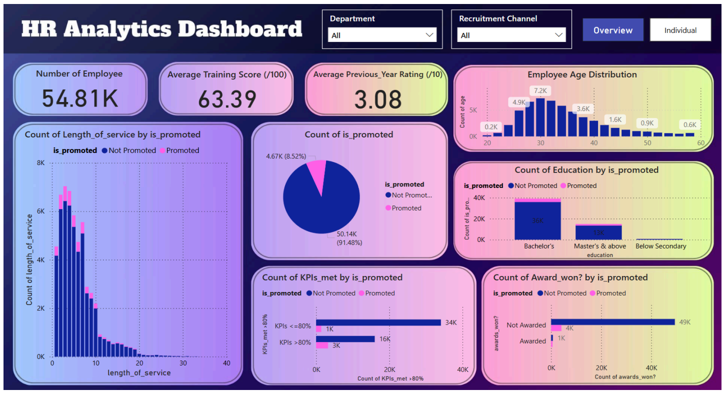

HR Analytics Dashboard
Department
All
Recruitment Channel
All
Overview
Individual
Number of Employee
54.81K
Average Training Score (/100)
63.39
Average Previous_Year Rating (/10)
3.08
Employee Age Distribution
Count of Length_of_service by is_promoted
Count of is_promoted
4.67K (8.52%)
50.14K (91.48%)
Count of Education by is_promoted
Count of KPIs_met by is_promoted
Count of Award_won? by is_promoted
is_promoted
Not Promoted
Promoted
length_of_service

Figure 2: PowerBI Dashboard Visualisation

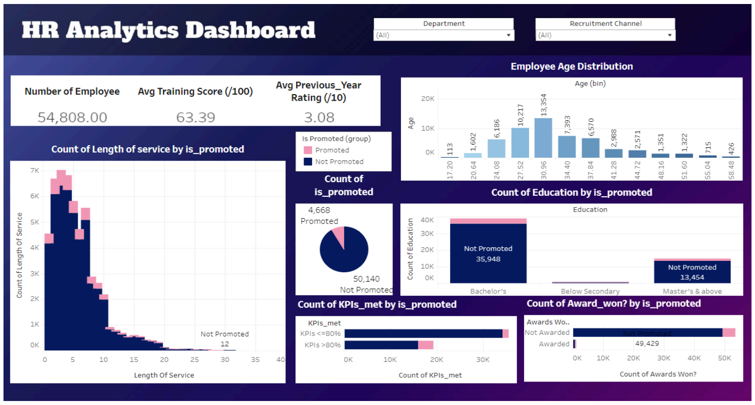


Figure 3: Tableau Dashboard Visualisation

Based on the visualizations, it's evident that 8.5% of the 54,808 employees across all departments and channels have received promotions. Notably, the department with the highest percentage of promotion is technology department leads with 10.76% of its employees being promoted. Within this department, the average training score stands at 79.93%, and promotions typically occur within the first to seven years of service. Furthermore, when examining recruitment channels, it becomes apparent that referrals stand out with the highest promotion rates, particularly striking in the legal department where 42.86% of promotions originate from this source. In contrast,employees in the legal department typically undergo a waiting period of 3 to 6 years before advancement, and those often showcase KPIs exceeding 80%.

We also conducted a comparative analysis based on our user experience with three different tools on their pros and cons.

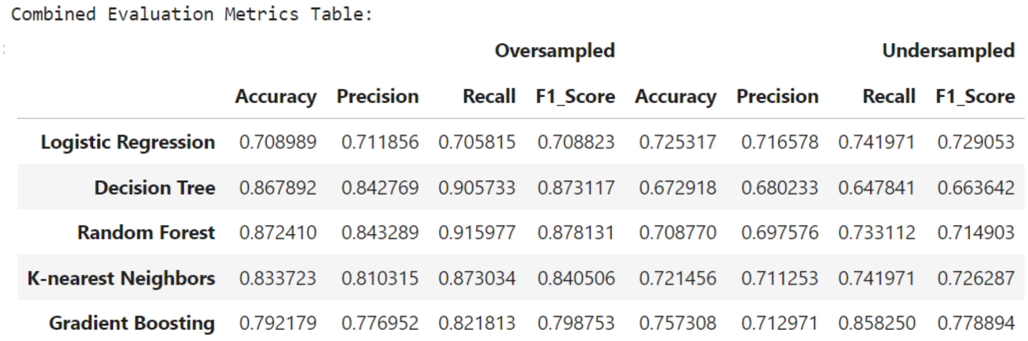
Tools	Python Libraries	PowerBI	Tableau
Pros	1. Offers various customization options for a wide range of plot types. 2. Can be integrated with other Python libraries, such as NumPy and Pandas, for data handling and manipulation.	1. Offers an intuitive drag-and-drop interface, making it accessible for users with little to no coding experience. 2. Supports real-time data connections and dynamic visualizations. 3. Dashboard can be accessed and interacted with by multiple users through the cloud. 4. Data can be transformed and manipulated easily into required format.	1. Offers an intuitive drag-and-drop interface, making it accessible for users with little to no coding experience. 2. Supports real-time data connections and dynamic visualizations. 3. Data can be transformed and manipulated easily into required format. 4. Can handle large dataset better than PowerBI.
Cons	1. Complex and not	1. It offers less flexibility in terms	1. Time taken to mastered the

	beginner-friendly as it requires Python programming basic. 2. Produces only static and non-interactive visualisation. 3. Cannot be visualise into an interactive dashboard.	of visualization customization compared to Python libraries like Matplotlib and Seaborn. 2. Performance can degrade with extremely large datasets or complex transformations.	functionalities in Tableau is higher than using PowerBI. 2. Tableau desktop required paid license which is expensive. 3. Has less data source integration compared to PowerBI.
After comparing the three tools, we concluded that Power BI is the most suitable data visualization tool to be used in our project.			

Table 1: Pros and Cons of Data Visualisation Tools

(ii) Evaluation of Classification Model on Oversampled and Undersampled Data

The initial exploration of the dataset revealed a significant imbalance in the 'is_promoted' column, with promoted employees (indicated by '1') being notably underrepresented. This imbalance necessitates the use of sampling techniques to address the disparity. Both oversampling and undersampling techniques have been applied to the dataset to manage the imbalance. Five machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, and Gradient Boosting—have been evaluated on both the oversampled and undersampled datasets to determine the most effective method for improving prediction accuracy.



Combined Evaluation Metrics Table:

	Oversampled				Undersampled			
	Accuracy	Precision	Recall	F1_Score	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.708989	0.711856	0.705815	0.708823	0.725317	0.716578	0.741971	0.729053
Decision Tree	0.867892	0.842769	0.905733	0.873117	0.672918	0.680233	0.647841	0.663642
Random Forest	0.872410	0.843289	0.915977	0.878131	0.708770	0.697576	0.733112	0.714903
K-nearest Neighbors	0.833723	0.810315	0.873034	0.840506	0.721456	0.711253	0.741971	0.726287
Gradient Boosting	0.792179	0.776952	0.821813	0.798753	0.757308	0.712971	0.858250	0.778894

Figure 4: Comparison result of evaluation metrics on oversampled and undersampled data

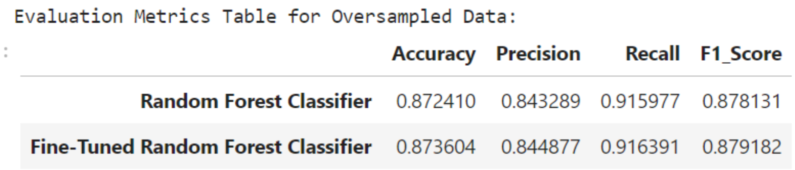
Figure 4 shows that the model performance is better on the oversampled data compared to the undersampled data, based on evaluation metrics such as accuracy, precision, recall, and F1 score, all generalized on the testing data. The results indicate that the overall performance of the model is superior with oversampled data compared to undersampled data.

The main factor contributing to this better performance is that oversampling increases the number of instances in the minority class while preserving all the information in the majority class, leading to a balanced dataset. Moreover, oversampling allows for better generalization by utilizing more data in the model and avoids introducing bias against the minority class, which in this case is the promoted employees. This is essential for achieving fair and accurate predictions, as the model is less likely to overlook or misclassify minority class instances.

In contrast, undersampling reduces the number of instances in the majority class, potentially leading to the loss of valuable information. This reduction can be detrimental because it limits the amount of data the model has to learn from, potentially weakening its ability to accurately predict outcomes for the majority class. Additionally, this reduction can introduce bias against the majority class and result in poor performance on new, unseen data in the imbalanced dataset. Furthermore, undersampling can lead to overfitting if the model fails to generalize well to new instances that reflect the true distribution of the data.

Based on the results above, the best-performing model selected among all classifiers is the Random Forest algorithm when trained on the oversampled data. The Random Forest model on the oversampled data achieves the highest accuracy (0.8724), precision (0.8433), recall (0.9160), and F1 score (0.8781).

Evaluation with correlation matrix and ROC-AUC curve graph for each classification task evaluation can be accessed through this (Link).



```
Evaluation Metrics Table for Oversampled Data:
```

	Accuracy	Precision	Recall	F1_Score
Random Forest Classifier	0.872410	0.843289	0.915977	0.878131
Fine-Tuned Random Forest Classifier	0.873604	0.844877	0.916391	0.879182

Figure 5: Comparison result of evaluation metrics of random forest classifier after tuning

After hyperparameter tuning on the best random forest classifier, as shown in Figure 5, the results improved slightly, with an accuracy of 0.8736, precision of 0.8449, recall of 0.9164, and F1 score of 0.8792.

(iii) The Key Factor Contributing To Position Promotion in the Company

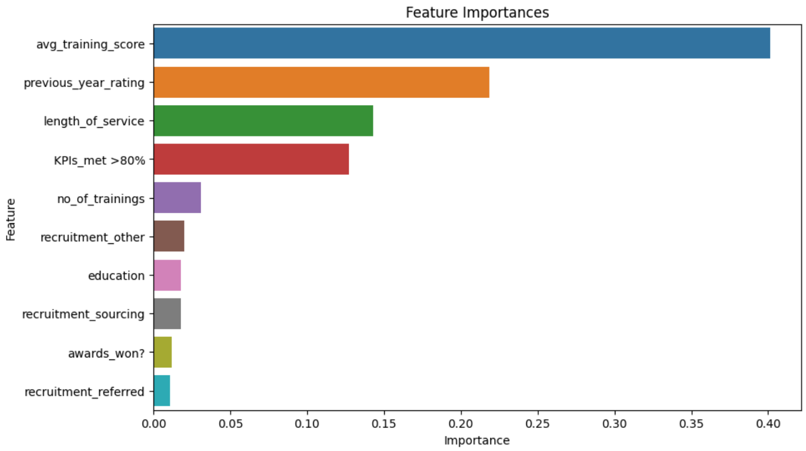


Figure 6: Feature Importance on the fine-tuned random forest classifier

In Figure 6, it is evident that the average training score, with a significance level of up to 40% , is the major factor that HR considers when deciding on employee promotions. This may due to the reason that employees with higher training scores have a higher level of skill development and knowledge enhancement, making them more competent and capable in their roles and deserving of position promotion. The ranking continues with employee ratings from the previous year at 22%, followed by the length of employee service at 14%, and employees with higher KPIs at 13%, in a descending order of significance, prior to considering any other features.

6.0 Conclusion

In conclusion, we have successfully achieved all the objectives stated in 3.0 Objectives. For visualizing employee performances, we utilized three different visualization tools: Python, PowerBI, and Tableau. Based on our user experience, all of us agree that PowerBI is the most convenient and user-friendly tool for data visualization due to its free platform, dynamic interactive cloud dashboard, and user-friendly functionality and interface. For the classification task, in handling the class imbalance issue, the up-sampling approach on the minority class has been proven to have better performance than down-sampling the majority class. For our project, the random forest classifier is chosen as the most suitable classification model as it performs the best based on every performance metric evaluated. Regarding factors, the average training score is the most important factor for the company to identify the most suitable candidate for position promotion. Hence, we hope that this research project will facilitate the workload of human resource management and address the issue of bias in position promotion.

7.0 References

Bert, J. (2024, February 2). *How To Effectively Negotiate a Promotion Salary Increase*. Indeed. https://www.indeed.com/career-advice/pay-salary/negotiate-promotion-salary

Barman, J. P. (2024, April 1). *Employee Promotion: The Types, Benefits, & Whom to Promote*. Vantage Circle. https://blog.vantagecircle.com/employee-promotion/

STATE OF DISCRIMINATION SURVEY MALAYSIA 2023. (2023, September). Persatuan Pendidikan Diversiti. https://www.aodmalaysia.org/_files/ugd/15355c_9d8e35607f5a440995263f255a2b0e93.pdf

Hoffman, K. (2021, February 14). *Machine Learning: How to Handle Class Imbalance.* Medium. https://medium.com/analytics-vidhya/machine-learning-how-to-handle-class-imbalance-920e48c3e970