**Author Guidelines for the Preparation of Contributions to Springer Proceedings – FOCUS2024**

\*Author11, Author2 1, Author31, and Author42

1Address 1

1\*Email id

2 Address 2

2Email id

\* Corresponding Author

**Abstract –** The Human Resource (HR) department plays a significant role in managing workforce dynamics. This includes identifying potential employees for job promotion based on their abilities and performances. However, in the Malaysian job market, this process is often affected by manual intervention and biases persisting between people with different ethnicities, religions, and genders. This research aims to develop a classification model with Python to automate job promotion and enhance the transparency and efficiency of talent management processes. Additionally, a dashboard will be generated to facilitate visualization of the company's employee performance. By analyzing various factors such as employee performances and training records, the model will reduce the subjective human bias traditionally present in promotion decisions. The insights generated from this research offer valuable guidance for HR professionals seeking to adopt an automated merit-based advancement system for promotion practices within their organizations.

**Keywords –** Human resource, job promotion, automation, classification, dashboar

**1 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 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 Objectives**

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

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

**4 Methodology**

**4.1 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.

**4.2 Data Cleaning**

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.

**4.3 Data Visualisation**

We utilized PowerBI to visualize employee performance namely. 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 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.

**4.4 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.

**4.5 Data Modelling**

For model training, we implemented 4 different classification algorithm models using Python for each dataframe which are "Logistic Regression", "Decision Tree", "Random Forest" and "K-nearest Neighbors". 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.

**4.5 Model Evaluation**

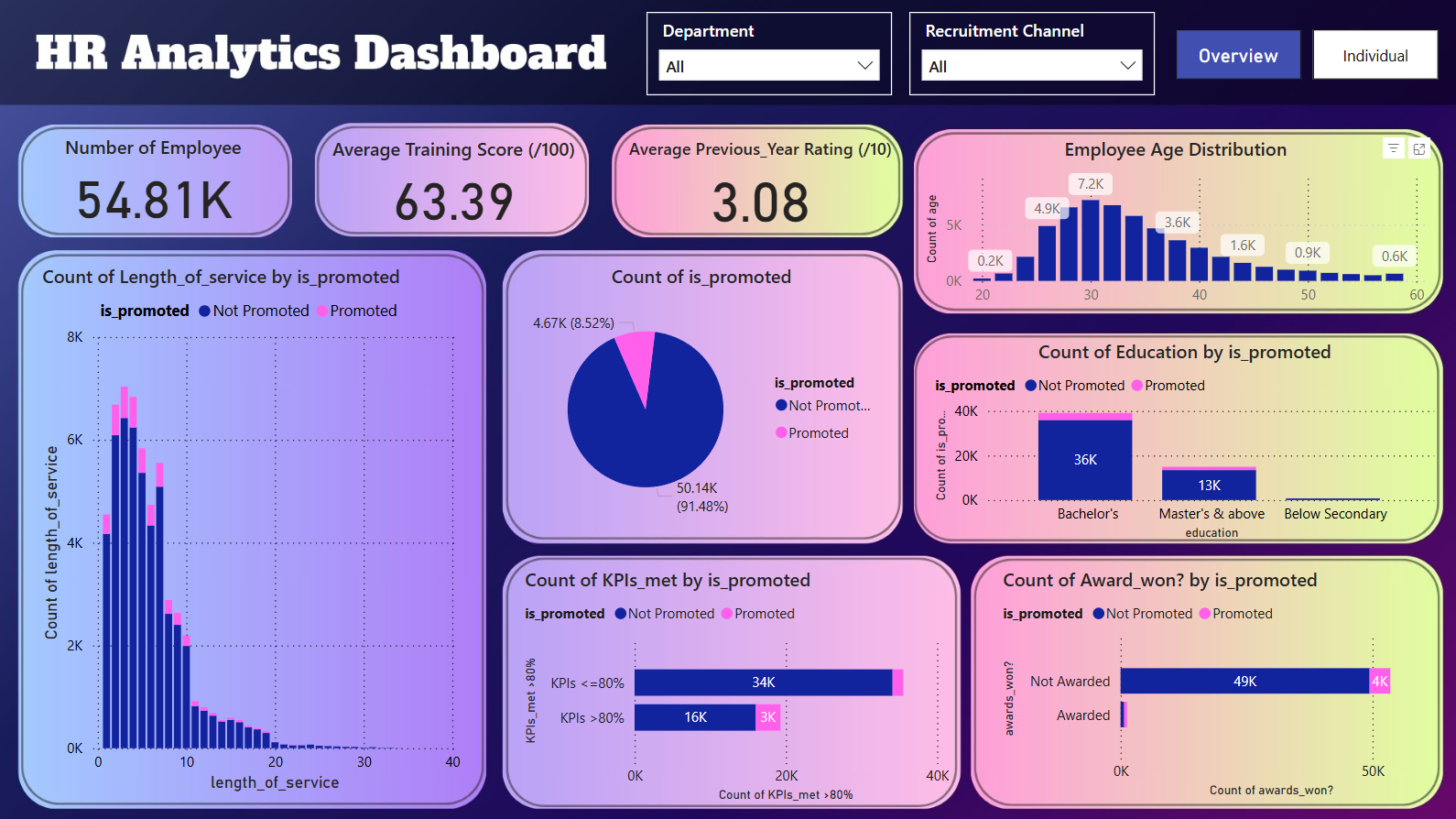
For each model trained on both dataframes, we evaluated their performance on testing data using key evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics were organized into a table format to facilitate easy comparison across models and dataframes.

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

**5 Results and Discussion**

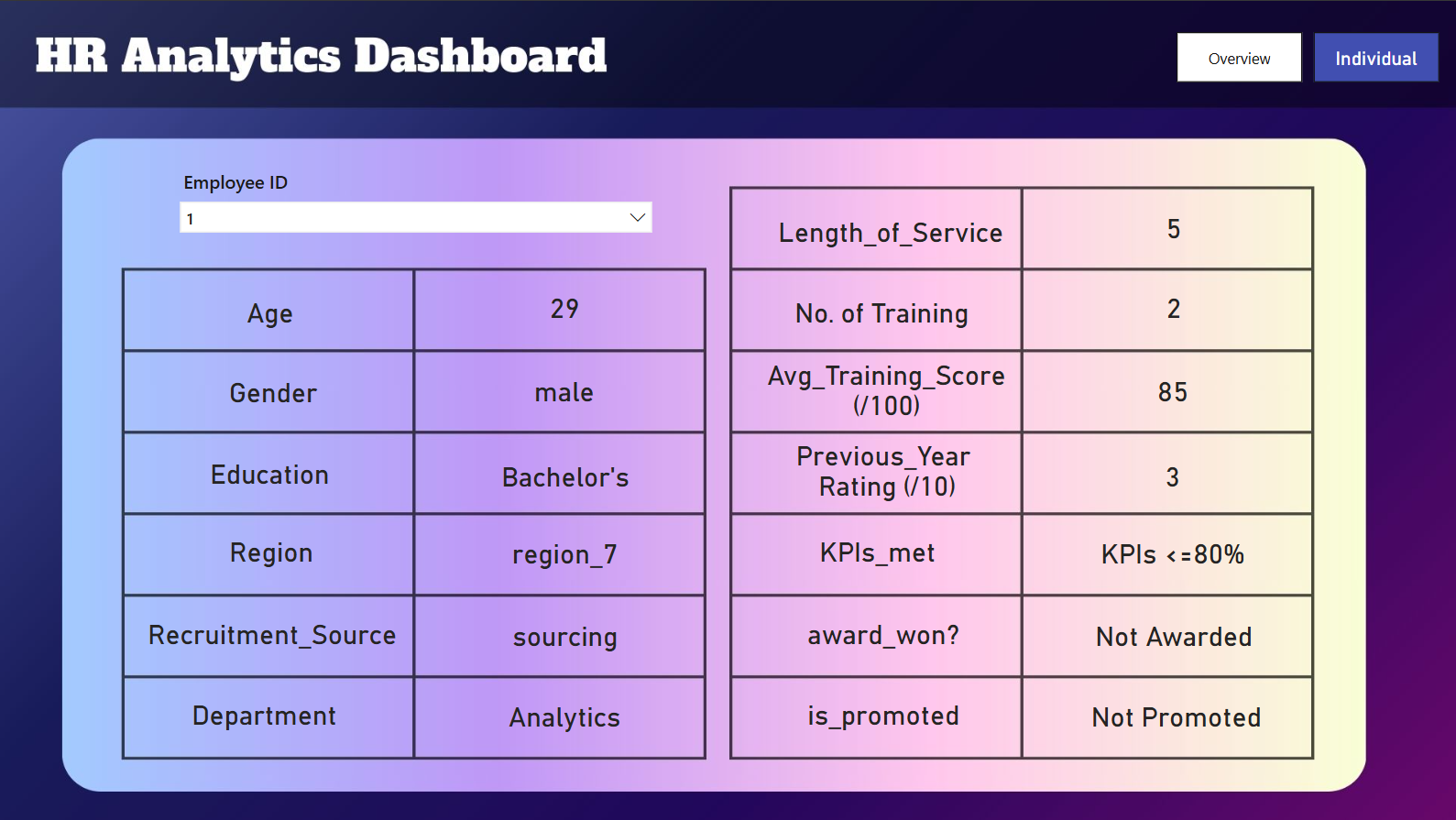
**5.1 Visualisation on Employee Performance**

In this project, PowerBI is ultilised for visualizing and analyzing employee performance. Its intuitive drag-and-drop interface made it accessible to users across varying technical skill levels to create insightful dashboards. Leveraging real-time data connections and dynamic visualizations, PowerBI provided immediate insights into workforce dynamics, enabling HR teams to monitor performance metrics continuously. Moreover, PowerBI offers the flexibility of cloud-based access, allowing multiple users to access and interact with the dashboard simultaneously. Additionally, PowerBI provides robust capabilities for data transformation and manipulation. With its comprehensive suite of tools, HR professionals can easily clean, filter, and format data to meet their specific requirements. This flexibility enables HR teams to tailor the dashboard to their unique needs.

****

**Fig. 1.** PowerBI Dashboard for Overview Performance

Based on the visualization in Fig. 1, it's evident that 8.52% of the 54.81K employees across all departments and channels have received promotions. The average training score is 63.39 out of 100, indicating moderate effectiveness of training programs for the company. The average previous year rating is 3.08 out of 10, which is relatively low, suggesting room for improvement compared to the recent training score. The majority of employees are in the 30-40 age range, with fewer employees in the 20-30 and 50+ age ranges. Most employees have a length of service between 0-10 years, and the count of promotions is also higher within this range. This infers that new employees are more easily promoted. Additionally, employees with higher education levels (Bachelor's and Master's) are more likely to be promoted as they likely possess more skills and knowledge in their respective fields. Moreover, it's very clear that employees who meet 80% or more of their KPIs have a higher chance of being promoted, highlighting the importance of KPI performance in promotion decisions. The ratio of awarded employees promoted to awarded employees not promoted is higher than the ratio of non-awarded employees promoted to non-awarded employees not promoted. This suggests that receiving awards is a strong indicator of promotion likelihood. The dashboard can also be filtered by department or recruitment channel using the slicer above.

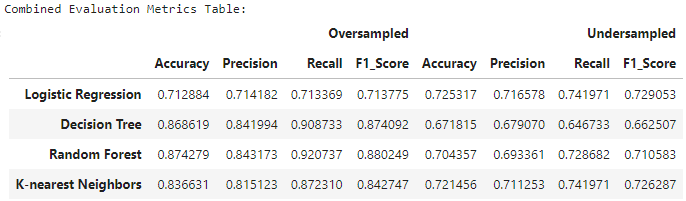


**Fig. 2.** PowerBI Dashboard for Individual Performance

In addition to providing an overview of the performance of all employees, there is also a dashboard interface for HR to view the details and performance of an employee by searching using their unique employee ID, as shown in Fig. 2.

**5.2 Evaluation on Classification Model**

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.



**Fig. 3.** Comparison Result of Evaluation Metrics on Oversampled and Undersampled Data

Based on the evaluation metrics shown in Fig.3., it is evident that the overall model performance is better on the oversampled data compared to the undersampled data when generalised on unseen testing 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 when trained on the oversampled data as it achieves the highest accuracy (0.8724), precision (0.8431), recall (0.9207), and F1 score (0.8802).

**6 Conclusion**

In conclusion, all research objectives have been achieved. PowerBI was utilized for visualizing employee performances. In addressing the class imbalance issue, the up-sampling approach on the minority class has demonstrated superior performance compared to down-sampling the majority class. For our project, the random forest classifier was chosen as the most suitable classification model, as it outperformed other models based on every performance metric evaluated. Regarding factors, the average training score emerges as the most important factor for the company in identifying suitable candidates for position promotion. This research is hoped to provide valuable insights to streamline the workload of human resource management and effectively address bias in position promotion processes.

**References**

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Mor­gan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov