Human Resources Analytic Based on Data Mining Methods

Asmaa Elmaghraby, Doaa Hamed, Esraa Negm, Mohamed Essawy, Rawan Hisham, Rawan Elframawy, Mona Arafa

Artificial Intelligence program, ITCS, Nile University
Cairo, Egypt.

[1]a.elmaghraby@nu.edu.eg
[2]d.hamed@nu.edu.eg
[3]e.negm@nu.edu.eg
[4]m.abdelmaged@nu.edu.eg
[5]r.hishma@nu.edu.eg
[6]r.elframawy@nu.edu.eg

***Abstract*— This research paper focuses on Human Resources Analytics based on Data Mining methods to build a predictive model for determining an employee's eligibility for promotion. Promotion within a workplace offers numerous advantages such as increased prestige, respect, and wages/salary, along with greater freedom and agency. Consequently, employees at various levels in an organization strive to climb the corporate ladder to enjoy these benefits. Organizations attach significant importance to promotion decisions, as the actions of individuals in key positions have a direct impact on overall organizational welfare. So, businesses dedicate resources to the selection process of employees to be promoted. The study employs the Naive Bayes, Decision Tree, and K-Nearest Neighbors (KNN) models to analyze data. Subsequently, Association rules can be used in Human Resources Analytics to identify patterns and relationships relevant between different factors that contribute to employee promotion or non-promotion, discover co-occurrence patterns, set promotion criteria, employee profiling and career planning, and rule-based decision support. By analyzing historical promotion data and associated employee attributes, association rule mining can provide valuable insights and aid decision-making processes.**

***Keywords*— Employee Promotion, Human Resources Analytics, Support, Confidence, K-Nearest Neighbors (KNN), Association Rules, Pattern Recognition, Data Analysis Predictive Analytics**

I. Introduction

Promotion in the workplace brings about numerous advantages that are highly sought after by employees. It entails an increase in prestige, respect, and wages/salary, along with greater freedom and agency within the business environment. Consequently, employees at different organizational levels strive to climb the corporate ladder to reap these benefits. In light of this, organizations pay careful attention to the promotion process and its implications for the business. The decisions made regarding promotions have a significant impact on the overall welfare and success of the organization. This research paper aims to contribute to the field of Human Resources Analytics by leveraging Data Mining methods to build a predictive model for assessing an employee's eligibility for promotion. The study will utilize the Naive Bayes, Decision Tree, and K-Nearest Neighbours (KNN) models to analyse relevant data and identify patterns that can help predict promotion outcomes. Additionally, the paper will explore the application of association rules to further enhance the accuracy and effectiveness of the predictive model. By investigating the relationship between various factors and promotion decisions, this research offers valuable insights that can benefit both employers and employees. The findings of this case study can be applied to other organizations seeking to optimize their promotion processes and enhance decision-making regarding employee advancement.

To achieve this, we use publicly available data from 2020, about 54,800 employees on Kaggle. This dataset provides a rich set of attributes that can be used to analyze and model employee attrition and identify important factors influencing employee churn. It offers an opportunity to explore relationships, patterns, and trends within HR analytics, particularly in the context of employee retention and turnover. Researchers and analysts can utilize this dataset to develop predictive models, understand the impact of various factors on employee attrition, and propose strategies for reducing turnover rates in organizations.

Our approach is unique in that it combines two widely used techniques in data mining, prediction models, and Association Rule Mining, to create a more effective and targeted prediction system. We will try to explore and understand the relationships between different attributes and promotional outcomes. By leveraging association rule mining techniques, organizations can gain valuable insights to optimize their promotion processes, align employee development efforts, and enhance decision-making related to promotions. We believe that data mining has a profound influence on an organization’s success by influencing and shaping the HR strategy.

II. Literature Review

Human Resource Decision-Making System. With the advent of knowledge economy, developing countries are facing the dual challenges of completing industrialization and accepting informatization and intellectualization. From now on, enterprises in developing countries, especially large and medium-sized enterprises, must further strengthen their awareness of scientific and technological innovation, improve the capital and product structure, rebuild enterprise organizations, adjust marketing strategies, and change management methods. However, in the final analysis, the challenge brought by the knowledge economy is the competition of talents. Strengthening enterprise human resource management is of great significance to promote the modernization of enterprises. Human resource is the main strategic resource of an enterprise in business decision-making and human resource [6]. As shown in Figure 1, human resources are the strategy of the enterprise.

Several methods have been proposed in Human Resources Analysis. Aziz [2] In this study, three algorithms models which are logistic regression, KNN, and decision tree has been tested to predict the which employees will get a promotion based on certain attributes. shows that logistic regression model performed better than other algorithms with the highest accuracy rate of 93.4%. In contrast, the decision tree shows an 89.5% accuracy score and a huge difference in precision score with only 37.5%. Chien and Chen [3] in their paper describes about data mining techniques as discovery driven rather than assumption driven. They have opined that these techniques used in HR related areas are very rare. They proposed decision tree analysis, clustering, association techniques. In details, they proposed various algorithms for decision tree like CART, CHAID, ID3, and C4.5 and compared them. From the study, using the CHAID algorithm in the sample organisation, they were able to design various.

Jantan et al [4] describes an architecture for talent forecasting by compiling various factors and attributes from various reviews done. They also suggested ANN, Decision trees, rough set theory as well as support vector machines techniques to be the most efficient in sequence of accuracy. Al-Radaideh et al [1] uses various data mining techniques like to predict performance of employees of basically using decision tree algorithms. the generated model was then validated by experimenting it with other organisations. They also opine d that Decision tree classifiers are popular techniques because the construction of tree does not require.

Karande et al [5] uses diverse machine learning algorithms on dataset to predict the employee turnover. Based on the performance of the individual classification, voting is taken, using which the final classification is done. The researcher uses multi-layer neural networks, SVM, voting and logistic regression. Tang [7] his paper takes data mining and random forest algorithm as the starting point, establishes the model of it, further focuses on the famous data mining platform, expands its open-source interface, and builds the management system model. Random forest algorithm is used to build a brain drain prediction model. Using the literature research method, data analysis method, and qualitative and quantitative combined analysis method to analyse the reasons for the brain drain, on this basis, put forward the main strategies to prevent the brain drain. that further medical examination personally to confirm their absence related to health grounds are decreased apart from the interpretation done by the fuzzy score.

Zhao et al [10] used machine learning algorithms starting from decision tree methods to neural networks to predict employee turnover. They conclude after their experiment that, ―If there are more HR datasets available, extreme gradient boosting is recommended to use as the most reliable algorithm. It requires minimal data pre-processing, has decent predictive power, and ranks the feature importance automatically and reliably. However, due to the complexity of employee turnover prediction, one should try to find the classifier that best fits the underlying data before taking this approach.

Sameer [8] in the project describes why an employee leaves an organization using data mining techniques. The researcher developed four models on the attributes given by domain experts. The models were designed using random forest tree, logistic regression, SVM, Gradient boosting machine with Bayesian optimization. Random forest tree giving the highest accuracy rate of about 85%. Samaila et al [9] uses and suggests fuzzy logic model for selection of employees to make the selection biased free and judged. Though they say that further medical examination personally to confirm their absence related to health grounds are decreased apart from the interpretation done by the fuzzy score., extreme gradient boosting is recommended to use as the most reliable algorithm. It requires minimal data pre-processing, has decent predictive power, and ranks the feature importance automatically and reliably.

Khera and Divya [11] in their paper used machine learning algorithms to predict employee turnover among IT professionals. They tag SVM to be an efficient learned algorithm to design the predictive model for employee turnover. They found the accuracy level to be 85 % for the Indian Its professionals retention prediction. They also opined that SVM should be given interest by the researchers as Neural networks to build predictive models. Several methods According to previous studies, no researchers have combines two widely used techniques in data mining, prediction models, and Association Rule Mining, to create a more effective and targeted prediction system. Therefore, in this paper We will try to explore and understand the relationships between different attributes and promotional outcomes. By leveraging association rule mining techniques, organizations can gain valuable insights to optimize their promotion processes.

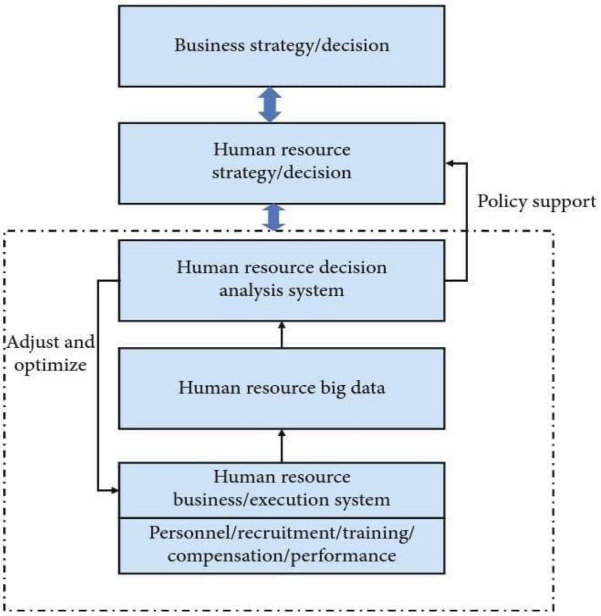

Business strategy/decision
Human resource strategy/decision
Policy support
Human resource decision analysis system
Adjust and optimize
Human resource big data
Human resource business/execution system
Personnel/recruitment/training/ compensation/performance

Fig. 1 Business demand model diagram for enterprise human resources.

III. METHODOLOGY

A. DATASET

The dataset is a thorough compilation of employee data collected from Kaggle that forecasts whether a potential promotion at a checkpoint in the test set will be promoted or not after the evaluation process. The dataset has 54,808 records and 14 features of employee data, including characteristics like department, age, average training hours, education level. This dataset is a helpful tool for HR analysts, researchers, and practitioners who are interested in understanding employee retention and estimating the likelihood that an employee will leave their position. The dataset's various features can be analysed by researchers to find patterns and trends that can be used to create machine learning models that can forecast employee churn.

B. Data Pre-processing

1. Missing value Imputation

Handling missing values is a critical step in data. pre-processing that might affect the analysis's correctness and dependability. By providing a detailed explanation of how missing values were handled. We can see from the dataset that two columns have missing values: education has (2409 in train dataset, 1034 in test dataset) and previous_year_rating has (4124 in tain dataset, 1812 in test dataset) As shown in Figure 2, *distribution of null values in train & test dataset* .

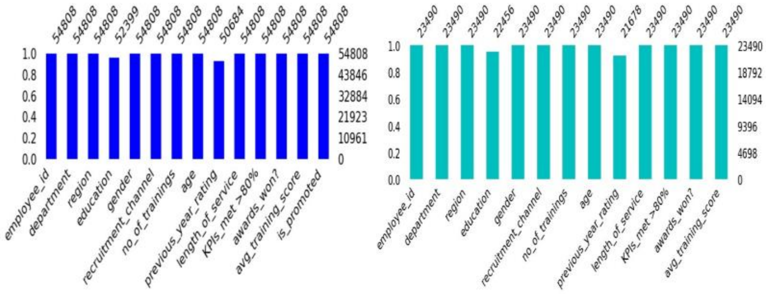


Fig 2, *distribution of null values in train & test dataset*

2. Detect outlier

data points that are unusually high or low compared to the other observations. Outliers can occur for various reasons, such as measurement errors, data entry mistakes, natural variations, or even being indicative of rare events or anomalies in the data. after checking the numerical columns: avg_training_score and length_of_service. As other variables have a limited number of values. As shown in Figure 3, outlier detection.

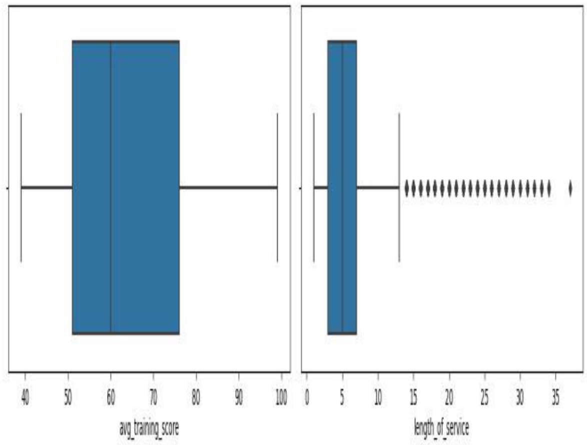
3. Label encoding

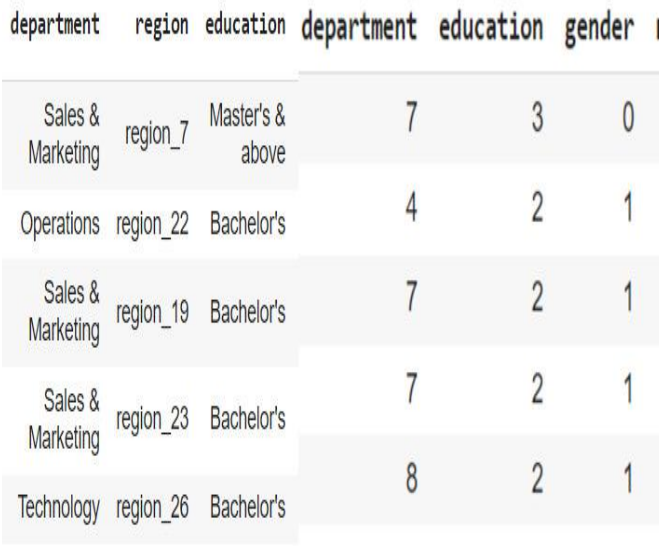
While examining the dataset, we found that some categorical columns must be converted to numerical columns as it will affect the accuracy, so we chose the label encoding method to convert categorical columns into a format that the machine learning model can understand. The categorical columns are education, department, and gender.as shown in Figure 4, Before and After Encoding.

4. Feature Engineering

In order to develop new features that can improve the performance of a machine learning model, we create a new column called 'total_score' in the 'train' DataFrame by multiplying the 'avg_training_score' column with the 'no_of_trainings' column. It calculates the total score for each employee in the training dataset. Similarly, it creates the 'total_score' column in the 'test' DataFrame by multiplying the respective column .

Fig.3 Outlier Detection





department	region	education	department	education	gender
Sales & Marketing	region_7	Master's & above	7	3	0
Operations	region_22	Bachelor's	4	2	1
Sales & Marketing	region_19	Bachelor's	7	2	1
Sales & Marketing	region_23	Bachelor's	7	2	1
Technology	region_26	Bachelor's	8	2	1

Fig. 4 Before and After Encoding

5. Handling Unbalanced Data

As shown in figure 5, *data misbalancing* Therefore, under sampling was the method we used to handle the data misbalancing since it is a rapid and efficient strategy to manage class imbalance in a dataset. We divided the data into classes according to the majority and minority. then sample the majority class less than necessary and combine it with the minority class.as shown in Figure 6,7 the result before and after handling unbalanced data.

Fig.6 Before handling unbalanced data

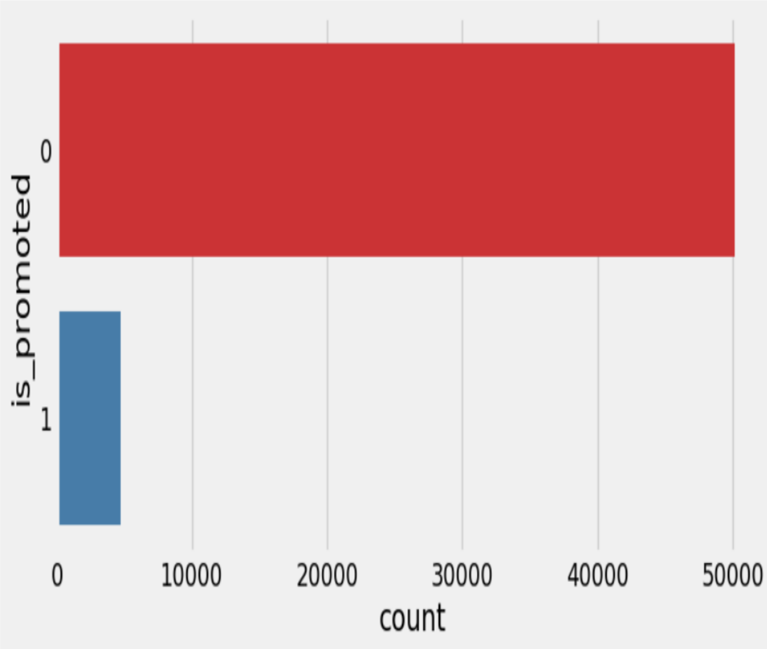


Fig .6 Before handling unbalanced data.

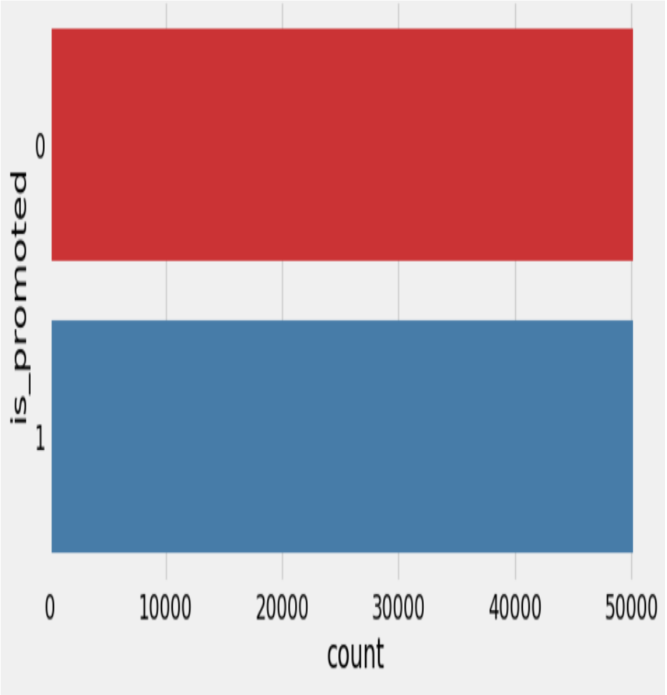


Fig.7 After handling missing data

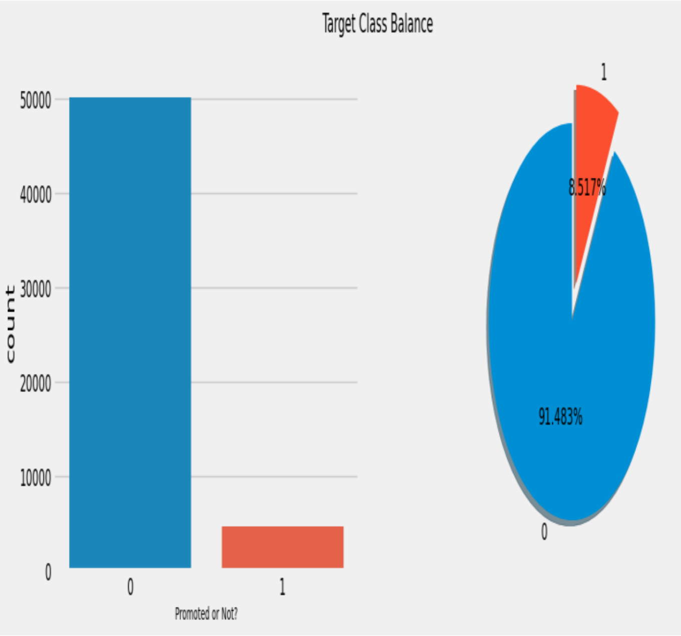


Fig.5 misbalanced data.

C. Data Visualization

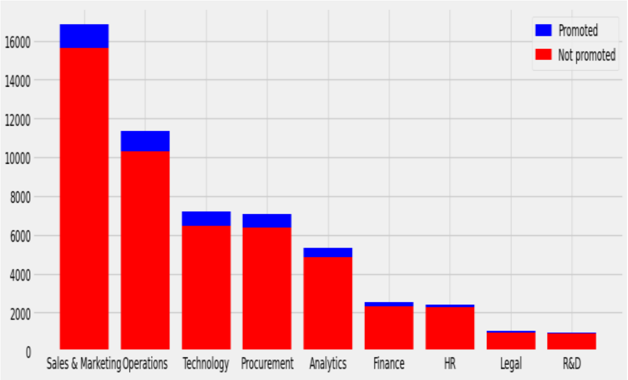


Fig .8 Department column

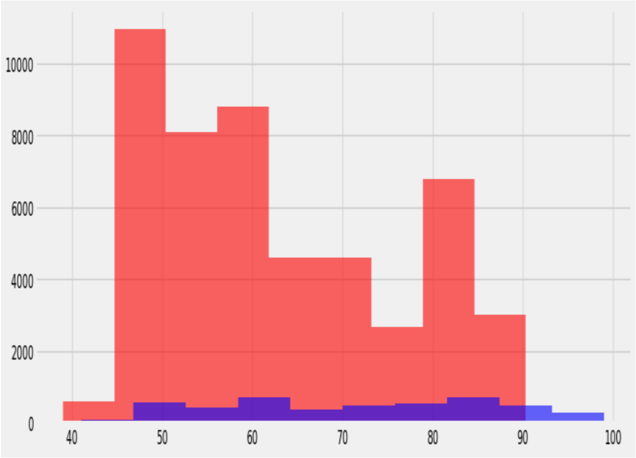


Fig.9 average training score

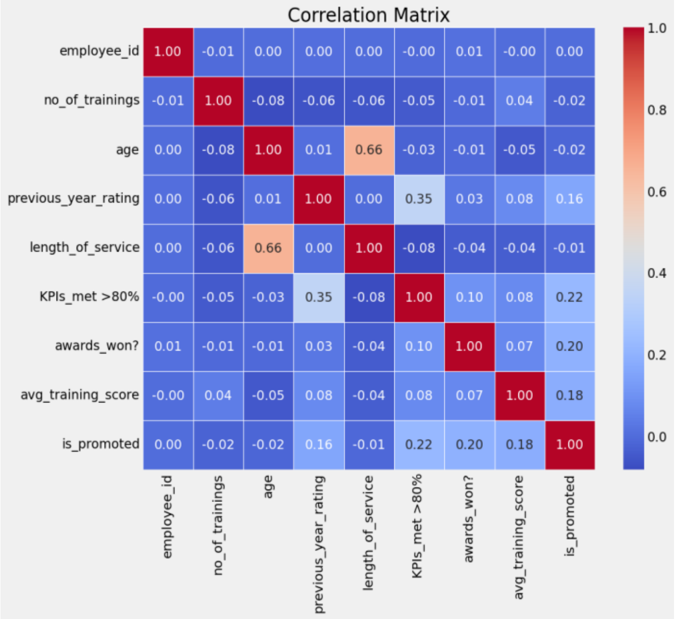


Fig.10 Heatmap and correlation matrix

IV. RESULTS

The evaluation of three models, namely K-Nearest Neighbors (KNN), Decision Tree, and Naïve Bayes, for predicting employee promotion in the Human Resources domain yielded insightful findings. The initial data analysis revealed a severe class imbalance, with 91.48% of employees not being promoted and only 8.51% receiving promotions. To address this issue, three methods were employed: oversampling using KNN, under sampling, and evaluation metrics based on recall and precision. When evaluating the models on the original imbalanced data, accuracy proved to be an inadequate measure due to the skewed class distribution. Consequently, the Receiver Operating Characteristic (ROC) curve analysis was employed to assess accuracy. As shown in fig 11.12.and 13.

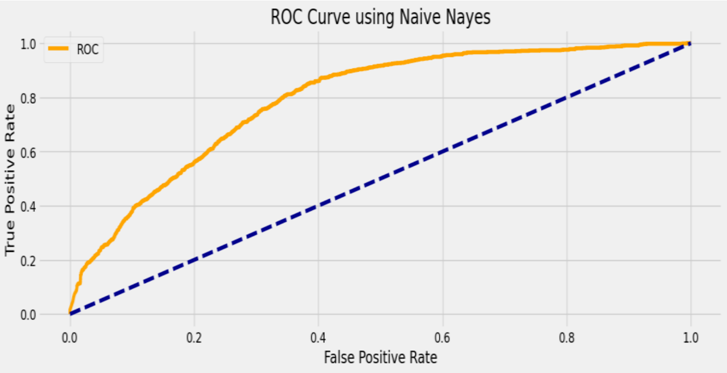


Fig.11 ROC curve using naïve nayes

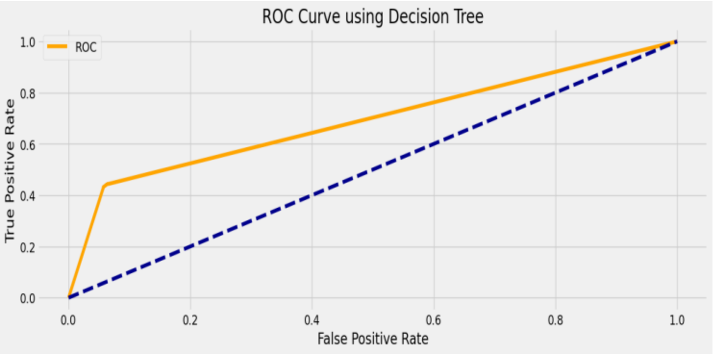


Fig.12 ROC curve using Decition Tree

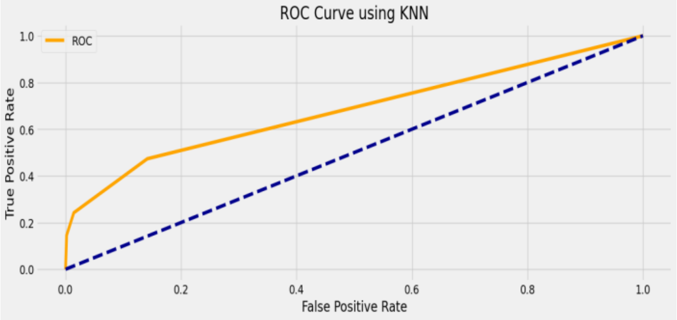
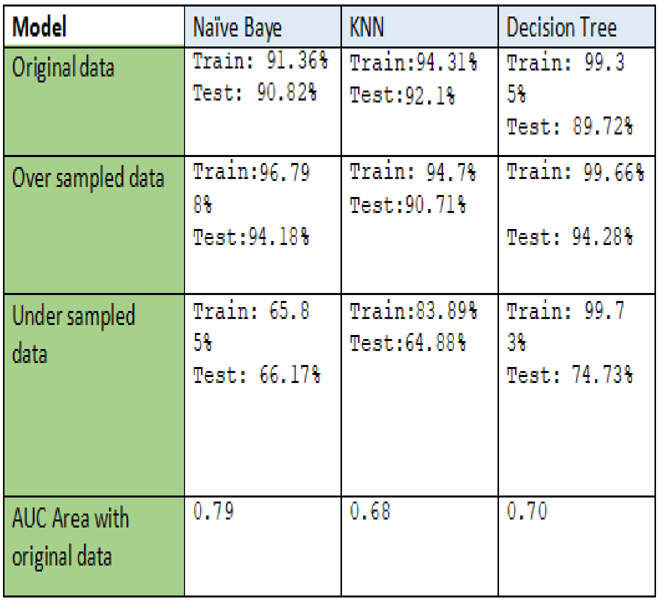


Fig.12 ROC curve using KNN

The results indicated that accuracy alone is not an effective metric for evaluating model performance in imbalanced data scenarios. Upon applying oversampling using KNN, the number of data records increased, thereby equalizing the representation of the two target classes. Subsequently, the Decision Tree model outperformed Naïve Bayes, with KNN performing least favourably in terms of predictive accuracy. Surprisingly, under sampling, intended to reduce the majority class records to balance the dataset, yielded unsatisfactory results. This approach led to a significant loss of valuable information, resulting in a decline in model performance.

To comprehensively evaluate the models, evaluation metrics based on recall and precision were employed. Precision, measuring the proportion of correctly predicted promotions out of all predicted promotions, yielded a test accuracy rate of 93.461% for KNN. This indicates a high level of accuracy in identifying promoted employees. Additionally, the recall metric, which assesses the model's ability to identify all actual promoted employees, achieved a test accuracy rate of 89.0 for KNN.as shown in table 1.



Model	Naïve Baye	KNN	Decision Tree
Original data	Train: 91.36% Test: 90.82%	Train:94.31% Test:92.1%	Train: 99.35% Test: 89.72%
Over sampled data	Train:96.798% Test:94.18%	Train: 94.7% Test:90.71%	Train: 99.66% Test: 94.28%
Under sampled data	Train: 65.85% Test: 66.17%	Train:83.89% Test:64.88%	Train: 99.73% Test: 74.73%
AUC Area with original data	0.79	0.68	0.70

Table.1 Models Results

V. Conclusion

The widespread use of human resources statistical management systems in government departments and enterprises has led to the accumulation of large amounts of data on businesses and talent. Analysing this data can reveal patterns and trends in the distribution of enterprises and talent within regions, which can provide valuable insights for policymaking. This research paper uses data mining and Decision Tree, KNN, and Naïve Bayes algorithms to create a model and insights to optimize their promotion processes, align employee development efforts, and enhance decision-making related to promotions. We believe that data mining has a profound influence on an organization's success by influencing and shaping the HR strategysystem. The model predicts brain drain and identifies strategies to prevent it by analysing literature, data, and qualitative and quantitative methods to understand the reasons behind brain drain.

the accuracy of correctly predicted promotions out of all predicted promotions and the K-Nearest Neighbours (KNN) algorithm achieved a precision test accuracy rate of 93.461%, indicating that it is highly accurate in identifying promoted employees. Overall, these results suggest that the KNN model performed well in predicting employee promotions with high precision and reasonable recall rates. It's important to note that depending on the specific context and objective.

VI. References

[1] Al-Radeidah, Q.A., Nagi,E.A.(2012).Predicting Faculty Development Trainings And Performance Using RuleBased Classification Algorithm. International Journal of Advanced Computer Science and Applications.Vol.3 No.2.pp 144-151.

[2] Azar.A, Sebt,M.A. , Ahmadi, P., Rajaeian,A.(2013).A model for personnel selection with a data mining approach: A case study in a commercial bank. SA Journal of HRM.Vol.11 no.1. doi:10.4102/sajhrm.v11i1.449.

[3] Chien,C.F. Chen,L.F.(2008). Data Mining to improve personnel selection and enhance human capital: A case study in High Technology Industry. Expert Systems with Application. Volume 34. pp 280-290. doi:10.1016%2Fj.eswa.2006.09.003.

[4] Jantan,H. Razak,A. Othman,Z.A.(2011). Towards applying Data Mining Techniques for Talent Management. International Conference on Computer Engine

[5] Kalaivani,V. Elamparithi ,M. (2014). An Efficient Classification Algorithm for Employee Performance Prediction. International Journal of Research in Advent Technology, Vol.2, No.9.pp27-32

[6] Kumar, S., Gupta V. (2013).Impact Of Performance Appraisal Justice On Employee Engagement: A Study Of Indian Professionals. Emerald Group Publishing Limited. Vol. 35(1). pp.61-78.

[7] Sarma, A. Lakhtaria, Dr. K.(2013). Data Mining Based Predictions for Employees Skill Enhancement Using ProSkill-Improvement Program and Performance Using Classifier Scheme Algorithm. International Journal of Advanc

[8] Wang,Q. Li,B. Hu,J.(2009). Human Resource Selection Based on Performance Classification Using Weighted Support Vector Machine. Journal of Advanced Computational Intelligence and Intelligent Informatics. Volume 13(4).pp 407-415.

[9] Wie-Chiang H., Ruey-Ming C. (June, 2007). A Comparative Test of Two Employee Turnover Prediction Models. International Journal of Management. Vol.24 No.2. pp. 216–

[10] Yee,C. Chen,Y.(2009). Performance appraisal system using multifactorial evaluation model. World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering, Vol. 3, No: 5.pp 231-25

[11] Yuan,J. Zhang,Q.M. Gao,J. Zhang,L. Wan,X.S. Zhou,T.(2016). Promotion and resignation in employee networks. Physica A 444. pp 442–447. doi:10.1016/j.physa.2015.10.039