www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Machine Learning for Employee Promotion Analysis and Prediction

[1]Mr. Ch Surya Prakash, [2]Kattula Sunitha,

[1]Associate Professor, Department of MCA, Rajamahendri Institute of Engineering & Technology. Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

[2]Student,Department of MCA, RajamahendriInstitute of Engineering & Technology. Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

Abstract

Organisations rely on performance forecasts to guide decision-making. Executives and managers who care about their companies' performance must deal with the challenging decision of promoting certain personnel since the success or failure of the business is often dependent on the competency of its employees. Because it is based on supervisors' assessments, the existing promotion procedure used by most firms should be seen as deceptive. This paper's primary objective is to identify the most relevant variables influencing employee promotion and to utilize classification algorithms to build prediction models for determining whether an employee is suitable for a promotion. This research makes use of the Kaggle 2020 dataset. In its 54,808 rows and 13 columns, you may find data on global corporations. This dataset includes information from nine different types of businesses. To forecast which employees will be promoted, we employed a number of predictive modeling approaches, such as K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Ensemble models composed of Adaboosting and Gradient Boosting. Gradient Boosting is the best classification technique when compared using accuracy, F1-score, and area under the curve (AUC). Additionally, the statistics demonstrate that an employee's rating from the previous year is the single most important factor in determining whether or not they will be promoted. Promotions within the department were unaffected.

Keywords—employee promotion, machine learning, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Adaboosting, Gradient Boosting

INTRODUCTION

Performance reviews are to be conducted by human resources (HR) in light of employees' contributions to the organization [1]. It is crucial to the company's performance since it measures the dedication of every employee. It also keeps workers focused on what they should be doing for the company. It could be difficult to manually complete performance reviews in a large firm. A lack of merit-based promotions might have a negative impact on morale and business operations in such a situation. An open and merit-based system for evaluating and promoting employees is, thus, necessary. King Khalid University at Abha, Saudi Arabia's Abdulaziz Almaleh College of Computer Science Dear Sir/Madam, A. Job Advancement A promotion's impact on an employee's career, performance, and the company's productivity is substantial, making it an essential component of any organization's success [2]. When working for a company, the HR department is there to help workers reach their professional goals. Keeping productive workers on staff allows a company to amass a pool of experienced workers, some of whom may go on to assume leadership roles with relative ease. Employee morale, loyalty, and output are all positively impacted by promotions. A higher overall engagement index is a side effect of promotions. B. The Impact of Promotions Moving up the corporate ladder from one job category to another is what's known as a promotion. There are perks to it as well, such a higher pay, more prestige, and more responsibility [3]. Because of the increase in power, position, and authority it brings to an employee, it is a major motivator for the majority of workers. Promotions are an effective way for organizations to fill open jobs at higher levels. Doing so shows that people are making a difference and encourages them to keep up the good job. C. The HR manager Promotions are one of the primary responsibilities of human resource managers. They decide whether workers have shown exceptional performance and are prepared for promotion [4]. When making important

choices, the HRM usually follows the advice of supervisors from other departments. Nevertheless, suggestions made by humans could be deceiving. When a supervisor is prejudiced or gives an inaccurate report, it might hurt an employee's advancement prospects. Consequently, it is difficult for the HRM to decide which workers should be promoted. In addition, workers could dispute the procedure. Because a promotion brings more responsibility, more money, and maybe even a leadership position, it's crucial to assess the candidate carefully based on a number of criteria. Human resource management should take an employee's leadership abilities, performance, appraisals, skills, and experience into account when promoting them. While some promotions may be time-based, others may be based on other criteria. Still, it's not always easy to objectively assess how well a candidate satisfies the promotion requirements. This void may be filled by AI, which can automatically select deserving personnel for promotions and eliminate prejudice in the process [5]. Therefore, this article set out to use machine learning to analyze dataset attributes in order to forecast which exceptional workers are eligible for the promotion. Additionally, by using machine learning categorization models, HR managers and employers may enhance the quality of HR decision-making and enhance the promotion processes. The goals of this paper are as follows: first, to establish valid evaluation criteria for judging an employee's performance; second, to study what variables influence promotions; third, to develop a workable prediction model; and lastly, to lessen the workload of HR in finding the right candidate. The following is the outline of the paper. The literature reviews' connected works are detailed in Section II. After reviewing the findings, a potential remedy is outlined in Section III. The paper's assessment tools are detailed in Section IV. The algorithms used in the article and their results are described and compared in Section V. Lastly, it lays forth the findings and plans for further research.

BACKGROUND AND RELATED WORKS

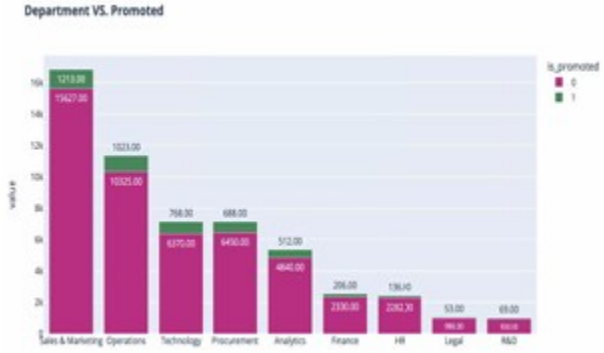
Many businesses' processes have been utterly transformed by the rapid development of information technology (IT) in recent years [6]. Every day, these companies' departments depend on IT to keep things operating smoothly. An essential aspect of every company is the human resources department. It safeguards the business against problems caused by subpar performance while maximizing staff productivity [6]. As a hub connecting all other departments, it primarily cares about the health and happiness of workers. An HR manager's main duties include managing the recruiting and firing process, handling salary and benefits, and keeping workers informed about regulations that impact the company's operations. Promotions are also handled by them according to predetermined criteria and workers' actual performance. Human resource management greatly benefits from the use of artificial intelligence (AI). One subfield of artificial intelligence, machine learning, may automate the process of creating analytical models [7]. An algorithm can learn from data, find patterns in it, and make judgments with little to no human interaction; this is the basic premise. An HR model's input values might comprise a variety of accomplishments, such as an employee's years of service and their degree of training. So, whether or not an individual is eligible for a promotion is heavily dependent on their performance. Workplace performance is influenced by factors such as leadership, training, and employee motivation [7]. The rate of staff turnover has also been calculated using machine learning. When skilled workers start looking for work elsewhere, it can be devastating for a company. Among other intangibles, a firm loses client ties when long-term staff go. Employees depart because of unprofessional management, inadequate pay, and an unpleasant workplace [8]. Using machine learning to detect trends in employee behavior that suggest they are about to quit or switch departments is one way attrition may be controlled [9]. In order to gauge how much of an impact each employee has on a business, performance reviews are conducted. Decisions that significantly impact the company's success rely on the results of the review. By supplying management with crucial information for making important choices about things like increases, promotions, and even layoffs, AI has proven to be an invaluable tool in performance reviews. Workers benefit from AI exercises like this because they showcase their efforts and recognize outstanding performance [10]. When workers see that the top-ranked workers are getting promotions, pay rises, and other perks, they want to be among them. Problems like prejudice or incorrect information are common in manual assessments of workers' performance. Consequently, it's possible to praise incompetent workers while letting dedicated ones go unrewarded. Artificial intelligence makes sure that everyone can

see how their work is being evaluated, allowing for fair promotions and layoffs. If an individual consistently meets or exceeds management's expectations or established benchmarks while working on given duties, the assessment will be favorable [10]. By looking at how an employee has performed in the past, we can see whether there has been an upward tendency. To determine how well a worker is doing, an AI model consults a dataset of their past work. To avoid misleading the model, it is important to remove inactive employees from the system, which requires data cleansing. One way to measure an employee's performance is by looking at how well they do certain tasks in comparison to established benchmarks. Workers' pay and advancement opportunities are strongly related to how their supervisors rate their work. Collective success is the result of each company's success. An organization's most precious resource are its high-performing workers. Recognizing and rewarding employees that continuously go above and above is our top priority. According to Long et al. [11]. Prospective Employee Advancement Prediction Using Job and Individual Characteristics. Based on data collected from a Chinese state-owned company, the authors of this study used machine learning techniques to forecast which employees will be promoted. To what extent may fundamental personal and positional data be used to forecast employee advancement? That was the primary goal of the research. Adaboost, decision trees, logistic regression, k-nearest neighbor, decision trees, and support vector classifier were among the popular classification models used. With an Area under the ROC (receiver operating characteristic) value of 0.96, the random forest model produced somewhat better predictions. The researchers failed to take supplementary characteristics like the quantity of trainings and rewards into account. Report by Liu et al. [12]. An Analysis of Employee Promotion Based on Data. In an effort to verify the impacts of organizational position on promotion, they used data from a Chinese state-owned business to do the study, which aimed to assess employee prospects, identify staff potential, and conduct the research. Logistic regression, random forest, and Adaboost were among the categorization models used for the estimate. With an area under the curve (AUC) of 0.856, the researchers determined that the random forest model performed the best. Improve your promotion choices with the help of Tang et al. [13], who use network-based methodologies and usage categorization. In order to determine what factors may lead to a promotion for a subset of an organization's employees, this research mined the company's HR database. To find the best performers, we used a combination of supervised learning and graph network analysis. Logistic regression, random forest, and Adaboost were among the categorization models used for supervised learning by the researchers. According to their findings, logistic regression outperformed the other methods with an accuracy rate of 75.61 percent. The method that yielded the greatest performance among the network-based algorithms was the one with [] set to 5. The use of a tiny dataset encompassing just one year, without the leadership traits, is one of the primary drawbacks of their study. On the other hand, Yuan et al. [14] apply the regression model to assist with employee network advancement and resignation. All 104 of Strong Union's workers' social media posts from 2013 were analyzed in the research. Gathering data on workers' work-related activities and social media contacts allowed us to examine the relationships between structural aspects and personnel. Using a logistic regression classification model, the researchers discovered that workers who were given more attention on the work-related network had a higher chance of being promoted, while those who were given less attention were more likely to quit. Comparing their model to others might provide more trustworthy results, despite the fact that it did disclose some intriguing facts. The authors of [15] forecast workers' output using a variety of supervised classifiers. The goal of their job is to identify what makes a good employee great. A corporation employed machine learning to forecast how well employees will do on the job. The researchers built prediction models using logistic regression, decision trees, and naive Bayes classification after using the cross-industry standard procedure for data mining. Among the three classifiers tested, logistic regression had the best accuracy. Considering the value of features might help refine the model even more. Section III: The Suggested Fix Part A: Analyzing Data Dataset describing big multinational corporations (MNCs) [16] is from Kaggle 2020. The 54,808 rows and 13 columns span nine main verticals throughout the companies. Subjects covered include: department, area, education, gender, age, recruitment channel, trainings completed, ratings from prior years, duration of service, awards received (yes or no), and average training score. The goal characteristic is whether the promotion is yes or no. The characteristics taken into account during the model's creation are listed in Table I.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

TABLE I
EMPLOYEEPROMOTIONDATASETATTRIBUTES

Features	Data Type	Description
employee_id	int64	Unique ID for employee
department	object	Department of employee
region	object	Region of employm (unordered)
education	object	Education Level
gender	object	Gender of Employee
recruitment_channel	object	Channel of recruitment for ployee
no_of_trainings	int64	no of other trainings compl in previous year on soft sk technical skills etc.
age	int64	Age of Employee
previous_year_rating	float64	Employee Rating for the previ year
length_of_service	int64	Length of service in years
awards_won?	int64	if awards won during previ year then 1 else 0
avg_training_score	int64	Average score in current trai evaluations
is_promoted	int64	(Target) Recommended for motion



In the figure 1 the most of the employees in the company were promoted from sales marketing, operations and Technology.

In the figure 2 there is a high chance of getting promoted if an employee has won an award.

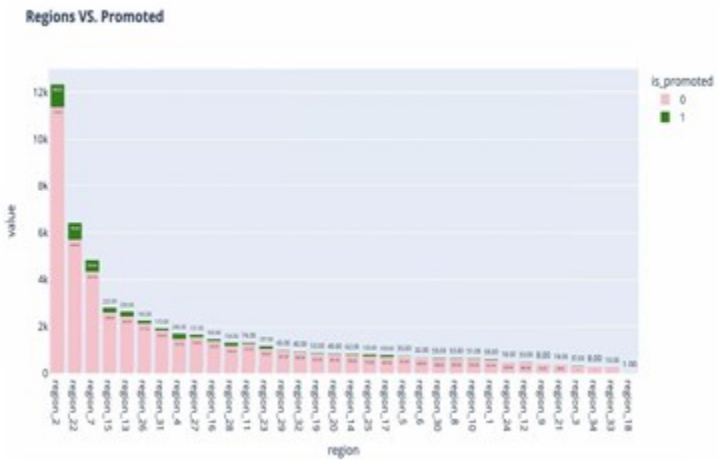


Fig. 3. Regions Vs. Promotion

Figure 3 shows a trend: people from region-2 get the most promotions, and there are a lot of them. This is also true for region=7, region-22, and region-2, all of which have a large number of workers. 2) Analysis of Correlations See how the independent variables are related in Figure 4. We looked for multicollinearity by comparing the correlations; multicollinearity was defined as a correlation coefficient (r) close to 0.80, which was not the case here. With a weight of almost 20%, the Awards earned feature was clearly the most essential component. Age and duration of service are correlated to a degree of around 66%. Section B: Preparing Data 1) Cleaning and Preparing Data Because machine learning relies on high-quality, relevant data, data preparation is an essential step in the process.
Data Expansion In this work, we use StandardScaler to normalize the data. To normalize feature values, StandardScaler first finds the average, then subtracts it from each value. Then, it divides the result by the standard deviation. C. Simulating First, dividing up data There were two sets of data used for each model: training and testing. To build the model, we utilized the training set, and to test it, we used the testing set. Each model made use of a stratifying parameter to divide the data in a manner that ensures the sample's percentage of values corresponds to the proportion of values fed into the parameter. At this stage, the training set received 80% of the data while the testing set received 20%. 2. Choosing the Right Model In its search for the optimal model for the dataset, this article uncovered a plethora of models. To boost the model's efficiency, we tried a number of different approaches. Support, Decision Tree, Logistic Regression, KNN, and Random Forest

Index in Cosmos
MAY 2025, Volume 15, ISSUE 2
UGC Approved Journal

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

The models utilized, which are briefly detailed below, were Ensemble (Adaboosting and Gradient Boosting), and Vector Machine. One non-parametric machine learning approach for data classification and prediction is K-nearest neighbors [20]. Based on the data, the working approach predicts the classes of the independent variable vectors that include the most of their neighbors. By fitting data to a logistic function, logistic regression, a supervised learning approach, may estimate the likelihood of an event happening [21]. In logistic regression, the dependent variable is a binary one, with data coded as either 1 for success or 0 for failure. An effective decision-making tool, a decision tree displays potential outcomes, alternatives, resource costs, and random occurrences in a tree-like structure [22]. Most datasets are categorical, and this method works well with them. Differentiating between promoted and non-promoted personnel is made easier by the tree-based splitting. One method for classifying data that uses supervised learning is the Support Vector Machine (SVM) [23]. Each data point is divided into two or more categories by a line or hyperplane that runs through the middle of the dataset. An ensemble learning approach that relies on trees is random forest [24]. It trains the model using a series of decision trees that randomly choose subsets of data, rather than relying on a single decision tree for data classification. The AdaBoost classifier [25] is an effective classifier that merges two less effective techniques. By merging many underperforming classifiers, the AdaBoost classifier is able to produce a robust and accurate classifier. Projects using structured or tabular data often make use of gradient boosting, a machine learning approach [26]. By sorting data sequentially, gradient boosting allows new forecasts to learn from the errors made by older models. The KNN and gradient boosting models' hyperparameters were fine-tuned by using a grid search approach. The search space is defined by a grid of hyperparameter values in grid search, and each point in the grid is investigated. For combinations that have already been successful, this approach is great for verifying them again. 3) Analyzing the Importance of Features

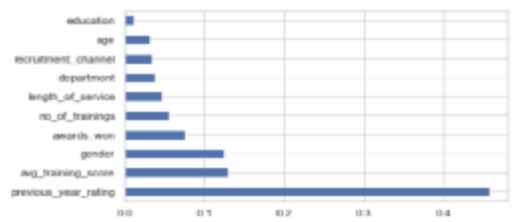


Fig. 9. Gradient Boosting Importance

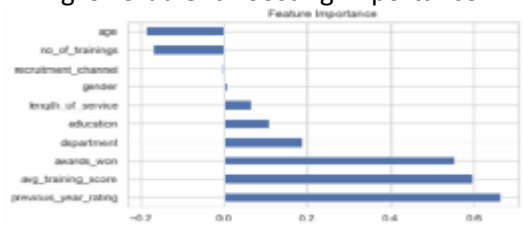


Fig. 10. Logistic Regression Importance

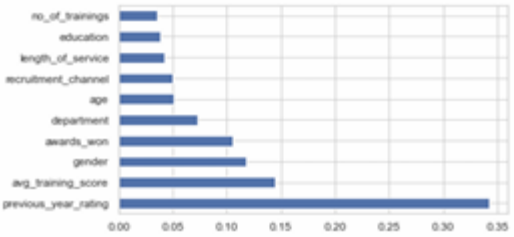


Fig. 11. Decision Tree Importance

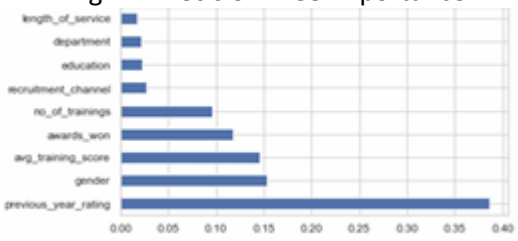


Fig. 12. Random Forest Importance

In these diagrams, It is possible to forecast a person's promotion based on their average training score, gender, previous year's rating, and awards obtained. Consequently, businesses may highlight these crucial aspects when determining a worker's promotion prospects.

EVALUATION Model

assessment involves looking at the strengths and weaknesses of several machine learning models and comparing them to choose the best one based on a set of criteria. The receiver operating characteristic

Index in Cosmos
MAY 2025, Volume 15, ISSUE 2
UGC Approved Journal

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

(ROC) curve, accuracy, precision, recall, and F1-score were the assessment metrics used. The accuracy and recall measures for the confusion matrix cells were determined by comparing the number of true positives and false negatives. Four distinct permutations of the actual and expected values make up the confusion matrix. In order to measure both accuracy and memorization, statisticians often use the F1-score, which is the harmonic mean of recall and precision. Here are the definitions of the cell values: A confirmation that your prediction of an employee's promotion was accurate.

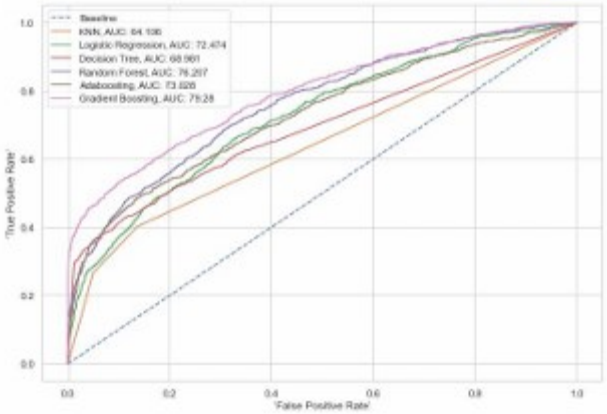
PROCEDURES FROM EXPERIMENTS

Prior to feeding the model, the feature selection procedures were used. From data collection to modeling, the whole process relies on feature selection. The scikit-learn package comes with a number of choices for feature selection. For the investigation, they used this mutual information categorization approach. This method chooses the independent variables that provide the most information gain by calculating their mutual information value with respect to the dependent variable. It basically checks how certain qualities relate to the end result. More dependent variables are indicated by a higher score. The average training score for mutual information was 0.030792, up from 0.015075 the year before, according to the data. classifier for voting The decision tree, logistic regression, random forest, and support vector machine models were then applied using the voting classifier. A machine learning estimator, the voting classifier first trains a large number of base models or estimators before making predictions using the average of all of them. The aggregating criteria and voting choices may be applied to the output of each estimator. In this study, we used a hard voting system where the projected output class was the one that had the most votes, meaning it was the class that each classifier was most likely to predict. A score of 0.902663 was recorded by the voting classifier.

One of the main ideas behind hyperparameters is that by setting a number of different combinations of parameters, the optimal combination may be found. The hyperparameter is shown in Table II.

TABLE II HYPER-PARAMETERS VALUES

Model	Hyper-parameter	Value	Score
K-Nearest Neighbors	n neighbors	2	0.891
	weights	uniform	
Gradient Boosting	learning rate	1	0.9395
	loss	deviance	
	n estimators	100	
Support Vector Classifier	C	1000	0.895
	kernel	RBF	
Random Forest	max features	log2	0.885
	n estimators	100	
	max depth	8	
Decision Tree	criterion	entropy	0.897
	min samples	8	
	splitter	random	



Roc curve

Results for both the training dataset and fresh samples are shown in this section, along with the methodologies that were presented. Those workers who get promotions are placed in the positive class, while those who do not are placed in the negative class. The area under the curve (AUC) and performance of each model were determined using the receiver operating characteristic (ROC) curve. An analysis of each model that was applied to the dataset. After comparing each model using the AUC measure, the Gradient Boosting model emerged as the clear winner with an AUC score of 79.2.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Model	ROC Score	Accuracy Score	F1 S
KNN:	0.62311	0.897555	0.89
Random Forest:	0.654857	0.887247	0.88
Logistic Regression:	0.557682	0.697409	0.76
SVM	0.667162	0.895548	0.89
Decision Tree	0.654356	0.899653	0.89
Adaboosting	0.641089	0.918902	0.90
Gradient Boosting	0.672211	0.939427	0.92

Fig. 14. Models Performance Comparison

Accuracy, f1 score, ROC curve, and area under the curve (AUC) were computed as part of the overall assessment, and the top classifier for predicting a worker's promotion status was found. The Gradient boosting approach outperformed all other algorithms on the given dataset, revealing the greatest F1 score (0.927)—a measure for a classifier's capacity to recognize all positive instances—and achieving the maximum accuracy (0.939) among the models tested. Afterwards, the Ada Boosting classifier method produces accurate predictions with a high level of 0.91. The Decision Tree classifier then produces 0.899 correct predictions.

CONCLUSIONS AND FUTURE WORK

A supervised machine learning classification model for promotion determination was the aim of this work. To determine who was eligible for a promotion, HR data from multinational corporations was analyzed. It is critical for any firm to know which workers have the potential to be promoted. Additionally, carefully specifying which workers are suitable for promotions requires more time and effort when the organization is bigger. Consequently, it is very beneficial to develop a model that may identify individuals who may be considered for a promotion. We created a number of prediction models, including KNN, LR, Decision Tree, SF, Random Forest, and Ensemble (Adaboosting and Gradient Boosting). Among the tested classification methods, Gradient Boosting proved to be the most effective. Promotion was unaffected by the featured recruiting channel or department, and the findings show no evidence of bias. The ratings from the previous year were the most essential consideration among the characteristics. One reasonable approach to the issue at hand is to use machine learning for the purpose of generating predicted decisions. The algorithms were trained to function well with very little data. Greater data would result in solutions that are more optimal. Therefore, HR analytics powered by machine learning can speed up decision-making and cut down on wasted time. In subsequent research, we will go more into other elements that are highly correlated with the promotion issue. In addition, we are trying to figure out how to improve our predictive performance by adding new features, getting this model distributed to almost all Saudi Arabian companies, and predicting how quickly an employee will be promoted or whether they are qualified for a higher-level position or have desirable leadership qualities.

REFERENCES

[1]. G. R. Ferris, M. R. Buckley, and G. M. Allen, “Promotion systems in organizations.” Human Resource Planning, vol. 15, no. 3, 1992.

[2]. P. Khatri, S. Gupta, K. Gulati, and S. Chauhan, “Talent management in hr,” Journal of management and strategy, vol. 1, no. 1, p. 39, 2010.

[3]. J. Schwarzwald, M. Koslowsky, and B. Shalit, “A field study of employees attitudes and behaviors after promotion decisions.” Journal of applied psychology, vol. 77, no. 4, p. 511, 1992.

[4]. N. Suleman, Mahyudi, Ansir, and M. Masri, “Factors influencing po sition promotion of civil servants in north buton district government,” vol. Volume 21, pp. PP 19–33, 04 2019.

[5]. P. Cunningham, M. Cord, and S. J. Delany, “Supervised learning, in machine learning techniques for multimedia,” 2008.

[6]. G. Randhawa, Human resource management. Atlantic Publishers & Dist, 2007.

[7]. P. Hamet and J. Tremblay, “Artificial intelligence in medicine,” Metabolism, vol. 69, pp. S36–S40, 2017.

[8]. S. S. Alduayj and K. Rajpoot, “Predicting employee attrition using machine learning,” in 2018 international conference on innovations in information technology (iit). IEEE, 2018, pp. 93–98.

[9]. P. K. Jain, M. Jain, and R. Pamula, “Explaining and predicting employ ees attrition: a machine learning approach,” SN Applied Sciences, vol. 2, no. 4, pp. 1–11, 2020.

[10]. B. L., “Artificial intelligence and human resource,” 01 2022.

[11]. Y. Long, J. Liu, M. Fang, T. Wang, and W. Jiang, "Prediction of employee promotion based on personal basic features and post features," in Proceedings of the International Conference on Data Processing and Applications, 2018, pp. 5–10.

[12]. J. Liu, T. Wang, J. Li, J. Huang, F. Yao, and R. He, "A data driven analysis of employee promotion: the role of the position of organization," in 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 4056–4062.

[13]. A. Tang, T. Lu, Z. Lynch, O. Schaer, and S. Adams, "Enhancing promotion decisions using classification and network-based methods," in 2020 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2020, pp. 1–6.

[14]. J. Yuan, Q.-M. Zhang, J. Gao, L. Zhang, X.-S. Wan, X.-J. Yu, and T. Zhou, "Promotion and resignation in employee networks," Physica A: Statistical Mechanics and its Applications, vol. 444, pp. 442–447, 2016.

[15]. M. G. T. Li, M. Lazo, A. K. Balan, and J. de Goma, "Employee performance prediction using different supervised classifiers."

[16]. "Hr analytics: Employee promotion data — kaggle," https://www.kaggle.com/datasets/arashnic/hr-ana, (Accessed on 05/06/2022).

[17]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," Journal of artificial intel ligence research, vol. 16, pp. 321–357, 2002.

[18]. "Hypothesis testing- statistics how to," https://www.statisticshowto.com/ probability-and-statistics/hypothesis-testing/, (Accessed on 05/10/2022).