Analysis and Prediction of Employee Promotions Using Machine Learning

Fatma Ayed Alqahtani
College of Computer Science
King Khalid University
Abha, Saudi Arabia
442813645@kku.edu.sa

Abdulaziz Almaleh
College of Computer Science
King Khalid University
Abha, Saudi Arabia
ajoyrulah@kku.edu.sa

2022 5th International Conference on Data Science and Information Technology (DSIT) | 978-1-6654-9868-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSIT55514.2022.9943959

***Abstract*—Predicting employee performance is essential for organizations. The success or failure of a company often depends upon the competence of its employees, so CEOs and managers who want their organizations to succeed face the difficult task of determining which employees should be promoted. The current promotion process used in most organizations should be considered misleading because it depends on supervisors' judgments. The major aim of this paper is to use classification algorithms to develop predictive models for predicting whether an employee is qualified for promotion or not and identifying the most important attributes affecting employee promotion. The dataset used in this paper is from Kaggle 2020. It contains information on multinational companies arranged in 54,808 rows and 13 columns. This dataset covers nine broad verticals across organizations. Several predictive modeling techniques, including K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Ensemble models (Adaboosting and Gradient Boosting models) were used to predict employee promotion. Based on measurements of accuracy, F1-score, and AUC, Gradient Boosting outperforms the other classification algorithms. The results also show that the most significant factor contributing to predicting employee promotion is the previous year's rating. Department had no effect on employee promotions.**

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

I. Introduction

The human resources (HR) is tasked with evaluating employee performances in terms of their contributions to the company [1]. It gauges the commitment of every company employee, which is paramount to the company's success. It also maintains the employees aligned with their corresponding tasks within the organization. In an organization with a number of employees, completing manual performance evaluations can be challenging. In such cases, an employee might not be promoted based on merit, potentially damaging morale and the company's operations. Thus, there is a need for a more transparent employee evaluation and promotion process based on merit.

A. Employee Promotion

Promotion is critical in any organization, as it has a significant effect on employees' careers and performances and the company's productivity [2]. The HR department guide employees achieve their career goal within organizations. Consequently, an organization builds an experienced workforce by retaining efficient employees, who, in turn, become competent leaders in the future. Promotions raise workers' spirits, boosting their loyalty and productivity. Additionally, promotions increase their overall engagement index.

B. Promotion Effects

Promotion in the workplace is the upward movement of an employee from one job group to another. It comes with additional benefits, such as increased salary, and status, and responsibilities [3]. Consequently, it elevates an employee's power, status, and authority; hence, it acts as a key motivating factor for most employees. It is good organizational practice to fill vacancies in higher job positions through promotions. Doing so motivates others to work hard because their efforts are not going unnoticed.

C. The human resource manager (HRM)

Human resource managers (HRMs) are mainly tasked with handing promotions. They determine which employees have performed in an exemplary fashion and are ready for promotion [4]. In most cases, the HRM relies on supervisors' recommendations from various departments while making critical decisions. However, manual recommendations can be misleading. The supervisor might be biased or provide an erroneous report that adversely affects an employee's chances of promotion. Therefore, the HRM faces a challenge in determining which employees should be promoted. Further, employees may raise questions about the process. Given that an employee's promotion implies new responsibilities, higher pay, and a new leadership role, it is important that they be thoroughly evaluated in terms of various attributes. When processing an employee promotion, the HRM should consider factors such as experience, skills, appraisals, performance, and leadership qualities. Some employee promotions could be time-based, while others are based on different factors. However, it is often difficult to determine how well an individual meets the set criteria for promotion without bias creeping in.

Artificial intelligence can fill this gap by identifying the employees for promotion without human intervention, thus ruling out any possible bias [5]. So the aim of this paper was to predict outstanding employees who are qualified for the promotion using machine learning, based on analyzing features in the dataset. Also, enabling employers and HR managers to improve their promotion processes and contribute to the higher quality of HR decision-making with using machine learning classification models.
objectives of the paper are first to set appropriate evaluation parameters as criteria to assess the performance of a given employee, second to investigate the factors affecting the promotion of employees and third to build a practical and effective prediction model and finally to reduce the time and effort the management would require to identify the suitable candidate.

The paper is organized as follows. Section II describes the related works that exist in the literature reviews. Section III describes the proposed solution that has been discussed in the research done. Section IV, the evaluation measures that employed in this paper. In Section V the results of the algorithms utilized in the paper are presented, along with comparisons between them. Finally, presents the conclusions and future work.

II. Background and Related Works

Advancement in information technology (IT) have revolutionized the operations of various organizations [6]. Departments within these organizations rely on IT for the smooth running of their daily activities. The HR department is one of the core departments in every organization. It maximizes employee productivity and protects the company from any issues arising from poor performances [6]. It is mainly concerned with employees' well-being and links employees with all other departments. Some of the primary responsibilities of an HRM include hiring and firing, processing benefits and compensation, and ensuring that employees are up to date with laws affecting the company's operations. Additionally, they process employee promotions based on set parameters and employees' performances.
Artificial intelligence (AI) is an instrumental tool in HR management. Machine learning, a branch of AI, helps automate the building of analytical models [7]. It is based on the idea that an algorithm can learn from data, identify patterns in them, and make decisions with minimal human intervention. The input values for an HR model can include the level of training an employee has, and their years of service, among other achievements. Consequently, an employees' performance is a key factor in determining whether they qualify for a promotion. The variables that affect an employee's performance in the workspace include motivation, training, and leadership [7].
Machine learning has also been used to determine employee attrition. An organization incurs a considerable loss when experienced employees leave for greener pastures. When long-term employees leave, a company loses customer relationships, among other intangible items. Employees leave due to poor management, low compensation, and a poor working environment [8].
Machine learning has been used to control attrition by identifying patterns in employee behavior that indicate that they are about to resign or move to a different department [9].
Employee performance evaluations measure the contributions and commitments of the members of a given company.
The outcome of the evaluation exercise is key in making key decisions that would shape the company's performance. The AI has proved instrumental in performance evaluations by providing management with key information for making critical decisions concerning such items as raises, promotion, and even layoffs. The employees find such AI exercises helpful because they make their effort known and reward exemplary performance [10]. Employees then strive to join the best-ranked employees because they believe they will receive benefits, such as raises, promotions, and other rewards. Manual evaluations of employees' performances often encounter challenges, such as bias, or misleading information. Hence, unqualified employees may be rewarded at the expense of hardworking ones. AI ensures transparency in the performance evaluation process while selecting only the best candidates for promotion and the poorly performing ones for retrenchment. A positive performance evaluation is one in which the employee completes assigned tasks successfully as required by the management or in observance of predefined standards [10]. Historical performance reviews determines whether a given employee has followed a positive trend. An AI model uses an employee dataset to evaluate their performances. Note that data cleaning is necessary to delete inactive employee from the system and ensure that the data used does not mislead the model.
Employee performance is determined by fulfilling or successfully completing a given task according to the predefined acceptable standards. The performance that the supervisor assesses is directly connected with the compensation or promotion that individual employees get. The success of an individual company leads to the company's collective success. High-performing employees are considered a valuable asset in a company. Employees who consistently exceed expectations are awarded accordingly.

Long et al. [11]. Prediction of Employee Promotion Based on Personal Basic Features and Position Features. In this work, the authors acquire data from a Chinese state-owned enterprise to construct a number of features and applied machine learning methods to predict employee promotion. The study aimed to verify the effectiveness of using basic personal and position information to predict employee promotion. They used common classification models, including k-nearest neighbor, logistic regression, support vector classifier, decision tree, random forest and Adaboost. The random forest model made relatively better predictions because its Area under the ROC (receiver operating characteristic) Curve value was the highest at 0.96. The researchers did not consider additional features, such as the number of trainings and awards.

Liu et al. [12]. A Data-Driven Analysis of Employee Promotion. They conducted the study with data from a Chinese state-owned enterprise to estimate employee prospects and identify staff potential and attempted to validate the effects of organizational position on promotion. They used classification models, including logistic regression, random forest and Adaboost for the estimation. The researchers conclude that the random forest model had the best performance due to an AUC value of 0.856.

Tang et al. [13], follow the utilization classification and network-based methods to enhance the promotion decisions. This study used data contained in a company's human resources information system to analyze drivers of potential promotion among a sample of its workers. The aim was to leverage both supervised learning as well as a graph network analysis to identify the highest-performing employees The researchers used classification models, including logistic regression, random forest and Adaboost, for the supervised learning. The outcomes of their work state that the logistic regression performed relatively better than the rest with an accuracy value of 75.61%. In terms of the network-based algorithms, the algorithm with ∞ set to 5 turned in the best performance. One of the main limitations of their work is the use of a small dataset covering just one year, with the absence of of the leadership features.

Yuan et al. [14] however use the regression model tp help in promotion and resignation in employee networks. The study examined the postings on a social network platform for all 104 employees of Strong Union up until the end of 2013. The aim was to collect a dataset consisting of employees' work-related actions and online social connections to study the correlations between structural features and employees. The researchers used the logistic regression classification model and found that employee who received more attention on the work-related network were more likely to be promoted, while employees who received little attention were highly likely to resign. Although their model revealed interesting facts, the model would be more reliable if compared with other models.

In [15], authors use different supervised classifiers to predict the performance of employee. Their work aims to determine the characteristics of excellent and good employees. Machine learning was used to predict employee performance within a company. The researchers used the cross-industry standard process for data mining, then the logistic regression, decision tree and naive Bayes classification methods were used to develop prediction models. The results showed that logistic regression had higher accuracy than the other two classifiers To further improve the model, feature importance should be considered.

III. Proposed Solution

A. Data Understanding

The dataset is from Kaggle 2020 and describes large multination companies (MNCs) [16]. It covers nine broad verticals across the organizations with 54,808 rows and 13 columns. The fields are department, region, education, gender, age, recruitment channel, number of trainings, previous year's rating, length of service, awards won (yes or no), and average training score. Promotion (yes or no) is the target feature. Table I list of the features that considered in the model's development.

TABLE I
EMPLOYEE PROMOTION DATASET ATTRIBUTES

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

1) Exploratory Data Analysis

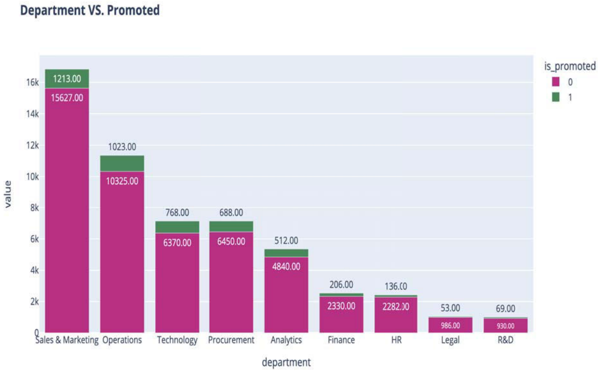


Fig. 1. Popular Departments

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

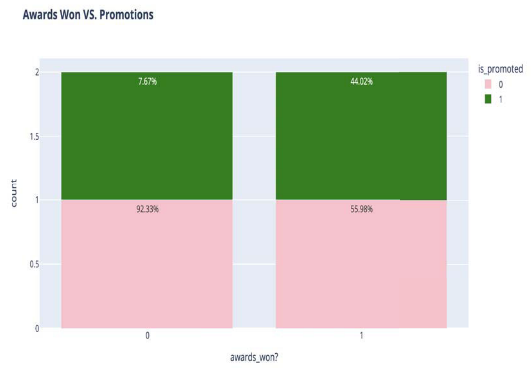


Fig. 2. Awards Won Vs. Promotion

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

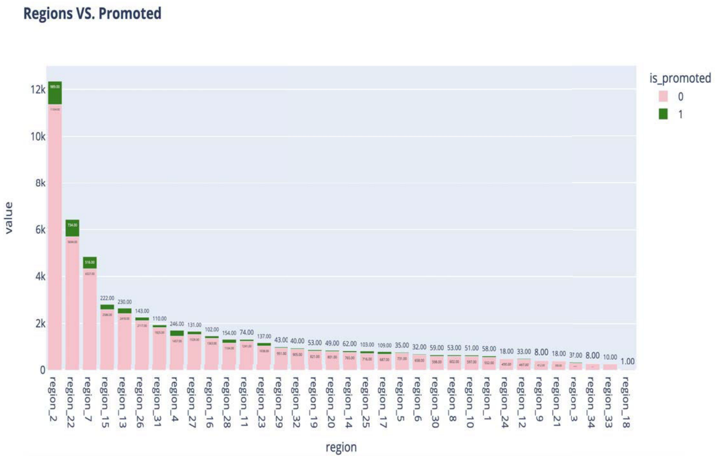


Fig. 3. Regions Vs. Promotion

A pattern in figure 3, there employees were more from the region-2 and most of the employees are promoted from this region,region=7, region-22, region-2 have high promotions as there were more employees from this region.

2) Correlation Analysis

The correlations between independent variables are depicted in Figure 4. Correlations were checked to make sure that there was no multicollinearity (multicollinearity was considered present whenever a correlation coefficient (r) was close to 0.80, which did not happen). Obviously, the Awards won feature was the most important element, with a contribution level of around 20% of the total. There is a correlation between age and length of service of around 66%.

B. Data Preparation

1) Data Pre-processing

Data preprocessing is a crucial stage in machine learning because the quality of the data and the relevant information that can be gleaned from it have a direct impact on models' capacities to learn;
consequently, preprocessing data before feeding it into a model is critical.

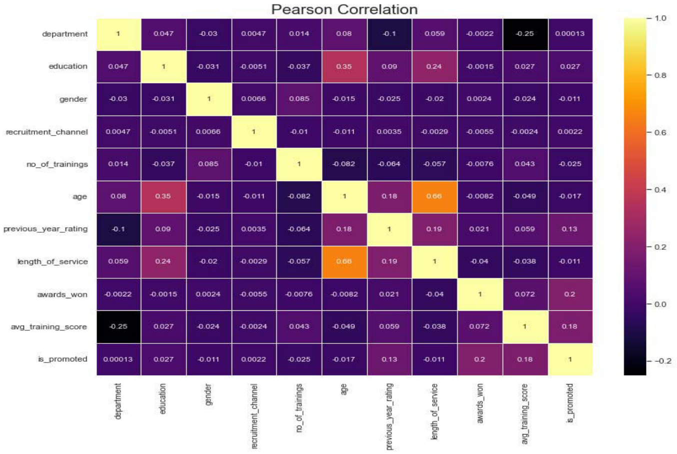


Fig. 4. Pearson Correlation

2) Data Cleaning

The first problem was an imbalanced distribution within the target feature class is_promoted.

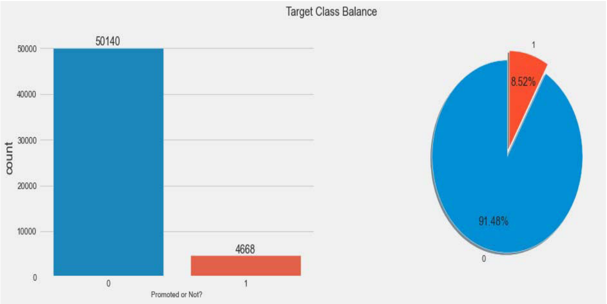


Fig. 5. Target Class Balance

This problem was treated with the synthetic minority oversampling technique (SMOTE). SMOTE is a powerful solution for data imbalances that was proposed by Nitesh Chawla (2002) [17]. Next, irrelevant variables, namely, employee ID and region were removed that had hardly any impact on analyzing and predicting the target variable. Finally, there were no duplicate rows in the dataset.

3) Missing Value Treatment

The dataset had two columns containing missing values, namely, education, and previous year rating.

Education Feature

Using the Mode to replace missing values was not the ideal solution because it increased the imbalanced data problem (most of inputs were at the bachelor's level).

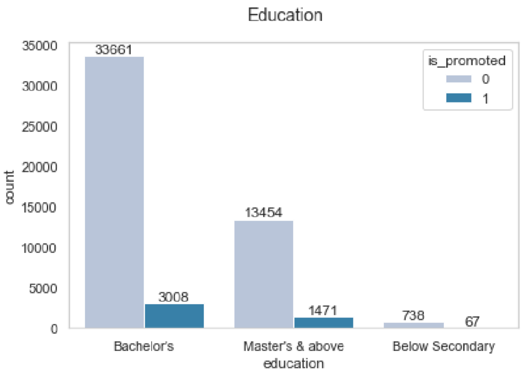


Fig. 6. Education Balance

The selected solution was to use K-Nearest Neighbors (KNN) to predict the missing values.

Previous Year Rating Feature
was choosing zero to populate the Column ”previous year rating” – Why is the data in Column ”previous year rating” missing to begin with? Because such employees were Freshers (i.e. length of service is 1 year), data was not entered.

There would have been no data for these personnel in the data source itself. Because Freshers with less than or equal to 1 year of experience may not have previous year rating at all, were filling missing values with ”0.”

4) *Outliers Treatment*

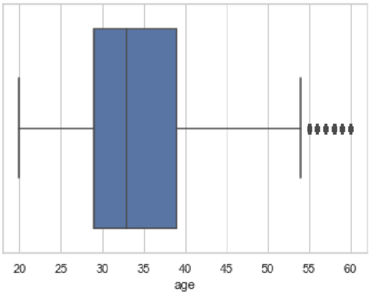


Fig. 7. Age Outlier

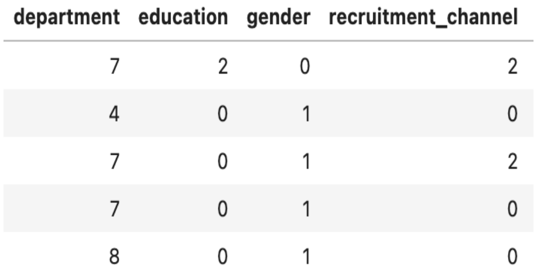
Figure 5 shows outliers for the age feature between the ages of 55 and 60. However, the retirement age in some countries is 60 years old and not any younger. Because age is normally distributed, the age distribution was converted to one based on three standard deviations.

5) *Hypothesis Testing*

It is important to examine the dependence of the target variable on the independent variables [18]. For each pair of dependent and independent variables, if there is no significant relation between the two, then that independent variable should not be used in the model building process. To determine the dependence of the target variable on the independent variables, chi-Square tests were performed for the hypotheses: H0: There is no association between the two variables, and H1: There is an association between the variables. The null hypothesis is rejected when (a) the p-value is less than the level of significance or (b) the calculated test statistic exceeds the critical value. Based on the chi-square test for categorical features, ANOVA for continuous features, all the categorical and continuous variables had an impact.

6) *Categorical Encoding*

The primary reason for converting categorical columns (department, education, gender, recruitment channel) to numerical columns is to allow a machine learning algorithm to understand them. The conversion of categories to numbers is called categorical encoding. This paper used label encoding, a popular encoding technique for handling categorical variables [19]. In this technique, each label is assigned a unique integer based on alphabetical ordering.



department	education	gender	recruitment_channel
7	2	0	2
4	0	1	0
7	0	1	2
7	0	1	0
8	0	1	0

Fig. 8. Label Encoding

7) *Data Scaling*

Data were normalized in this paper using StandardScaler. StandardScaler standardizes a feature’s values by subtracting the mean from each value and dividing the result by the standard deviation.

C. *Modeling*

1) *Data Splitting*

For each model, the dataset was divided into two groups: training and testing. The training set was used to create the model, while the testing set was used to evaluate it. A stratifying parameter was used in each model to partition the data in such a way that the proportion of values in the sample created matches the proportion of values provided to the stratifying parameter. During this phase, 80% of the data was assigned to the training set, while the remaining 20% was assigned to the testing set.

2) *Model Selection*

This paper discovered many models in order to find the best model for the dataset. Also were attempted a variety of techniques were to improve model performance. The KNN, Logistic Regression, Decision Tree, Random Forest, Support

Vector Machine, and Ensemble (Adaboosting, and Gradient Boosting) models were used, briefly described in the following paragraphs:

K-nearest neighbors is a non-parametric machine learning algorithm for classifying and predicting data [20]. The working technique estimates the classes of the independent variables' vectors in which their nearest neighbors are concentrated based on the data.

Logistic regression is a supervised learning technique for predicting the probability of an event occurring by fitting data to a logistic function [21]. The dependent variable in logistic regression is a binary variable with data coded as a 1 or 0 (1 means yes, success, etc., and 0 is no, failure, etc.)

A decision tree is a decision-making tool that uses a tree-like representation of options and possible outcomes, such as chance events, resource costs, and utility [22]. Categorical datasets, which are highly prevalent, function well with this technique. The tree-based splitting facilitates distinguishing between promoted and non-promoted employees.

The Support Vector Machine (SVM) is a supervised learning technique for data classification [23]. A line or a hyperplane separates the data points, and the separation between the two sides divides the entire dataset into two or more categories.

Random forest [24] is an ensemble learning technique based on trees. Instead of using a single decision tree to classify the data, it uses a succession of decision trees to train the model by selecting subsets of data at random.

The AdaBoost classifier [25] combines two poor classifier algorithms to create a powerful classifier. The AdaBoost classifier combines many low-performing classifiers to create a strong classifier with high accuracy.

Gradient boosting is a machine learning technique frequently used in projects with structured or tabular data [26]. Gradient boosting sorts data in sequential order, with new forecasters learning from previous models' mistakes.

Hyperparameters were tuned in the KNN and gradient boosting models using a grid search technique. Grid search defines the search space as a grid of hyperparameter values, with each point in the grid being examined. This method is particularly useful for double-checking previously successful combinations.

3) Features Importance Analysis

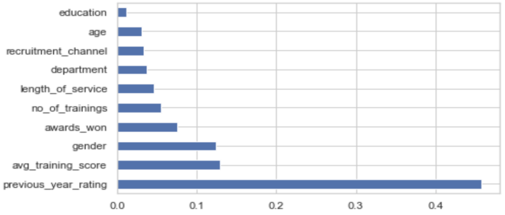


Fig. 9. Gradient Boosting Importance

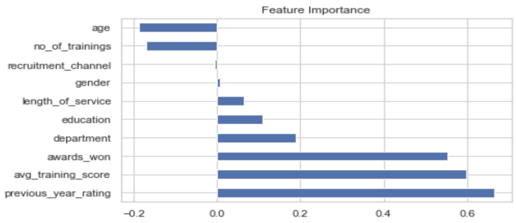


Fig. 10. Logistic Regression Importance

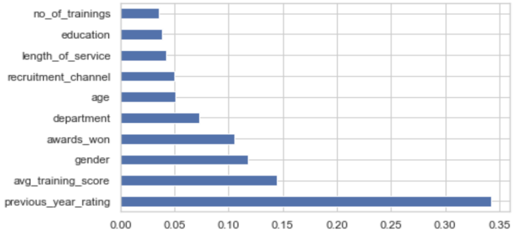


Fig. 11. Decision Tree Importance

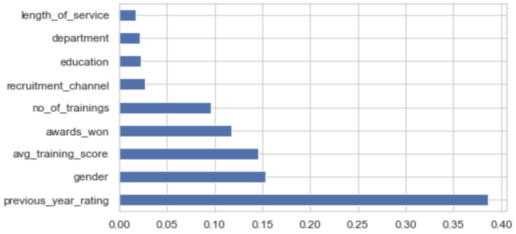


Fig. 12. Random Forest Importance

These plots show Previous year rating, average training score, Gender and Awards won are the most important features that contribute to the predict the promotion. As a result, organizations can emphasize these important factors when deciding whether or not to promote an employee.

IV. Evaluation

Model evaluation is the process of analyzing a machine learning model's performance, strengths, and shortcomings and comparing them to determine the best performer using various evaluation criteria. The evaluation measures employed were the accuracy, precision, recall, F1-score, and ROC (receiver operating characteristic) curve.

True positives and false negatives were used to calculate the precision and recall metrics for the cells in the confusion matrix. The confusion matrix is a table with four different combinations of predicted and actual numbers. The F1-score, defined as the harmonic mean of precision and recall, is a popular statistic that combines precision and memory. The cell values are defined as follows:

True Positive: You correctly predicted that an employee would be promoted.

True Negative: You anticipated an employee would not be promoted appropriately.

False Positive: You incorrectly predicted an employee would be promoted.

False Negative: You incorrectly predicted that an employee would not be promoted.

To calculate accuracy using the following equation:

$$\frac{TP+TN}{TP+FP+FN+TN}$$

To calculate F1-score using the following equation:

$$2\times\frac{Precision\times Recall}{Precision+Recall}$$

To calculate True Positive Rate (TPR) using the following equation:

$$\frac{TP}{TP+FN}$$

To calculate False Positive Rate (FPR) using the following equation:

$$\frac{FP}{FP+TN}$$

V. Experimental Results

The feature selection approaches were used just before feeding the model. Feature selection is crucial to the overall process, which starts with data collecting and ends with modeling. For the feature selection, the scikit-learn software has several built-in options. This mutual information classification method was adopted for the study. This approach calculates the mutual information value of each independent variable in relation to the dependent variable and selects the variables with the greatest information gain. It essentially assesses the link between attributes and the desired outcome. A higher score indicates more dependent variables. The results for mutual information found an average training score of 0.030792, which was higher than the previous year's rating of 0.015075.

1) Voting Classifier

Next, the voting classifier was used to apply the decision tree, random forest, logistic regression, and SVM models. The voting classifier is a machine learning estimator that trains many base models or estimators and then predicts based on the results of each base estimator. For each estimator output, the aggregating criteria can be combined with voting decisions. Hard voting was utilized in this research, and the predicted output class was defined as the class with the majority of votes, i.e., the class with the highest probability of being predicted by each classifier. The accuracy of the voting classifier was 0.902663.

2) Parameter Adjustment

The fundamental concept of a hyperparameter is that pre-set various parameter combinations can identify the best parameter combination with the best score. Table II shows the hyperparameter settings and model scores.

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	

3) Model Testing

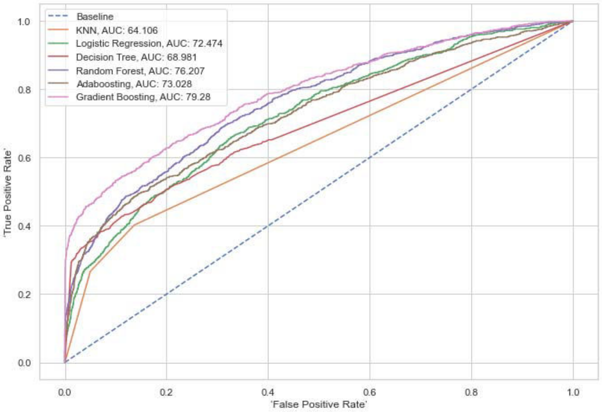
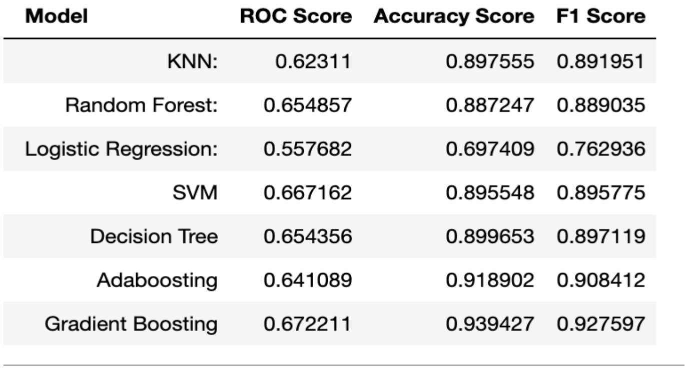


Fig. 13. ROC Curve

This section shows the outcomes of the described methods on both the training dataset and new samplesEmployees who are promoted are assigned to the positive class, while those who are not promoted are assigned to the negative class. The receiver operating characteristic (ROC) curve was used to determine the area under the curve and measure the performance of each model (AUC). A comparison of all models run on the dataset. The different models were compared with the AUC metric, the best model was the Gradient Boosting value of AUC = 79.28



Model	ROC Score	Accuracy Score	F1 Score
KNN:	0.62311	0.897555	0.891951
Random Forest:	0.654857	0.887247	0.889035
Logistic Regression:	0.557682	0.697409	0.762936
SVM	0.667162	0.895548	0.895775
Decision Tree	0.654356	0.899653	0.897119
Adaboosting	0.641089	0.918902	0.908412
Gradient Boosting	0.672211	0.939427	0.927597

Fig. 14. Models Performance Comparison

The essential metrics required for an overall evaluation (accuracy, f1 score, ROC curve, AUC) were calculated, and the best classifier to predict whether an employee will be promoted or not was identified. The algorithm that produced the best results for the available dataset was the Gradient boosting technique: it revealed the best F1 score (0.927), a metric that assesses a classifier's ability to detect all positive instances, and achieved an accuracy 0.939 was the highest from other models. Then, the Ada Boosting classifier technique has high predictions with 0.91 accuracy. After that, the Decision Tree classifier has accurate predictions with 0.899 accuracy.

Conclusions and future work

The goal of this paper was to develop a supervised machine learning classification model for determining promotions. HR data from MNCs were used to predict which employees qualified for promotion. Knowing which employees are eligible for promotion is very important for every company. Further, the larger the company, the more time and effort it takes to specify the employees qualified for promotion manually. Therefore, the creation of a model that identifies candidates for possible promotion is very useful. The following prediction models were developed: KNN, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Ensemble (Adaboosting, and Gradient Boosting) models. The results show that the Gradient Boosting outperformed the other classification algorithms. The results indicate no bias and that the features recruitment channel and department did not play significant roles in promotion. The most important factor amongst the features was the previous year's ratings. The use of machine learning as a predictive decision-making tool is a completely viable solution for the presented problem. The fairly limited data were able to train the algorithms to perform with good accuracy. More data would lead to more optimized solutions. Thus, machine learning in the field of HR analytics can reduce the amount of time that goes into decision making, thereby increasing efficiency.

We will continue to investigate more factors that have high correlation with the promotion problem in future study. Also, attempting to anticipate promotion speed or determining whether a promoted employee is qualified for a higher-level post or whether a candidate shows desirable leadership attributes, and attempting to get this model distributed to practically all Saudi Arabia Companies and looking to find a way to improve predictive performance through the development of new features.

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 position 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 employees' 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 intelligence 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).

[19] "sklearn.preprocessing.labelencoder — scikit-learn 1.0.2 documentation," https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html, (Accessed on 05/10/2022).

[20] O. Kramer, "K-nearest neighbors," in *Dimensionality reduction with unsupervised nearest neighbors*. Springer, 2013, pp. 13–23.

[21] R. Jayadi, R. Jayadi, and H. Firmantyo, "Employee performance prediction using naive bayes," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3031–3035, 2019.

[22] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," in *2011 IEEE control and system graduate research colloquium*. IEEE, 2011, pp. 37–42.

[23] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," *Journal of theoretical and applied information technology*, vol. 12, no. 1, pp. 1–7, 2010.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] G. I. Webb, "Multiboosting: A technique for combining boosting and wagging," *Machine learning*, vol. 40, no. 2, pp. 159–196, 2000.

[26] S. M. Piryonesi and T. E. El-Diraby, "Data analytics in asset management: Cost-effective prediction of the pavement condition index," *Journal of Infrastructure Systems*, vol. 26, no. 1, p. 04019036, 2020.