Employee Promotion Prediction by using Machine Learning Algorithms for Imbalanced Dataset

Kevser Şahinbaş
Department of Management Information System,
Istanbul Medipol University, Istanbul, Turkey
Email: ksahinbas@medipol.edu.tr

2022 2nd International Conference on Computing and Machine Intelligence (ICMI) | 978-1-6654-7483-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMI55296.2022.9873744

***Abstract*—Promotion processes are one of the most important processes in terms of human resources. A promotion process organized fairly within the organization is a managerial tool that motivates employees and contributes to business continuity. Promotion is an important extrinsic motivation for many employees. It ensures the employee's engagement and commitment to the organization and contributes to the continuity of his current performance. It is also an important rewarding and performance control mechanism for the organization. Many factors such as seniority, performance level, competencies, age, awards, training score, organizational commitment of the personnel who will be promoted are taken into consideration. In this study, a prediction methodology will be studied based on the criteria evaluated for the employees in the promotion processes by Machine Learning algorithms such as Support Vector Machine, Artificial Neural Network, and Random Forest. Random Forest achieved the highest performance with 98% accuracy, 96% precision, 1.0% recall and 98% f1-score values with ROS approach. This study could be used by HR and manager to predict the probability of promotion so that managers can find the right parameters for someone to get promoted.**

Keywords—employee promotion, prediction, HR dataset, imbalanced dataset, data management, ANN, RF, SVM

I. INTRODUCTION

One of the most delicate topics in any employee's life is promotion. Promotion is the assignment of an employee to a higher-level position in terms of responsibility, authority and pay. When the promotion procedure is applied correctly, the company's success grows as well as the employees' motivation and devotion to the organization. Seniority and qualification are the most important factors in employee advancement. The company's career management success is contingent on establishing a clear and objective promotion policy and applying it fairly. Under what conditions, by whom and how promotions will be made, what qualifications are required for promotion to each position should be determined in advance and presented to all personnel.

Greater responsibility, increased duties, increased privileges, and increased authority are all markers of job advancement. According to studies [1-2] promotions have a favorable and significant impact on employee work performance. Haryano et al. [3] stated that promotion has a beneficial impact on job performance.

Promotion is the transition of any personnel in the enterprise from their position to another task that will increase their authority and responsibilities even wider and increase their status. The fact that the employees come to a better place than they are in the institution be an increase in motivation for the workers. Because it is of great importance for the employee to know that his/her labor will be rewarded and that he/she will have the opportunity to rise, both in terms of increasing his/her commitment to the job and working more efficiently. When this is the case, both production, efficiency and quality will increase. This will enable the company to continue its production with great profitability. As a result, both the employee and the employer will be happy. If an employee is promoted and given more benefits, they become more satisfied and committed to their work. Satisfied employees work harder and more readily, and employee turnover is minimized [4]. According to Dean and Joseph [5], work promotion is defined as a growth in the workforce or employees already employed in better jobs, as measured by more responsibilities, facilities, achievements, higher qualification demands, higher status, and increased earnings or salaries. Tessema and Soeters [6] have the same research that performance practices have sufficient relationship with employee promotion. According to Knowles et al. [7], the goal of job promotion is to provide high-performing individuals with more recognition, position, and acknowledgment. Obtain personal fulfillment, pride, a greater social status, and a higher salary. Workplace motivation, discipline, and productivity are all on the rise. Ensure employee stability, evaluate employee promotions using assessment indicators, evaluate employees on a timely basis, and be transparent. Because of new occupations, job promotion chances have a variety of effects on businesses. Providing employees with opportunities to improve their creativity and innovation for the benefit of the firm. Other employees are motivated by expanding employee knowledge and job experience. Due to a mutation in the role, a new organizational structure is being implemented. According to the research findings of Shahzad et al. [8], there is a significant relationship between employee performance and promotion. He stated that there should be certain principles regarding promotions in company policies since promotions have a positive relationship with employee performance and organizational productivity.

Promote the post to another so that it does not become empty. Employees who are promoted to the appropriate position appreciate and benefit from their work environment, enhancing their productivity. To make job promotion prospects more accessible to applicants.

This paper proposes a decision support system designed for a Human Resource (HR) departments about eligibility of employees' promotion. The study's contribution is the using of imbalanced dataset techniques to cope with imbalanced problem. Another contribution of the paper is to focuse on parameter tuning. Employees who may be promoted as a result of this study will be identified, and HR will be able to use this information to improve key performance indicator

(KPI) KPIs in promoted jobs. RF outperformed the other algorithms with 98% accuracy, 96% precision, 1.0 recall and 98% f1-score rate obtained among SVM and ANN in this study.

The paper is organized as described below. Section II describes the related work. Section III presents the methods and materials. Imbalanced dataset, ANN, SVM and RF algorithms are explained in detail. Section IV indicates the proposed system overview. Section V gives the experiment results and compare the performance metrics of the algorithms. Lastly, Section VI concludes the paper.

II. RELATED WORK

According to McIntyre [9], staying with a company for 10 years and getting promoted 3-4 times in that 10 years is ideal. "At a time like this, those numbers look good on a resume too," McIntyre says. According to him, it seems reasonable for an employee with a career life of 30 years to be promoted an average of 10-15 times. The study proposed by Hameed et al. [10] used Variable adaptive momentum (BPVAM) backpropagation and principal component analysis (PCA) analyzes together, which can increase the classification accuracy, for the identification of epilepsy cases from EEG signals, and they also performed a comparative analysis with some automated techniques. Mutlu et al. [11] proposed a Convolutional Neural Network (CNN)-based model for the diagnosis of liver disease using the BUPA and ILPD datasets. CNN's performance was compared to machine learning approaches such as Naive Bayes (NB), Support Vector Machine (SVM), K-neaest Neighbors (KNN), and Logistic Regression (LR). Their work indicated that CNN is effective in classifying liver disease, achieving 75.55% and 72.00% accuracy in the BUPA and ILPD datasets, respectively. Rasheed et al. [12] proposed a model that predict Parkinson's disease in the early stage. After using the variable adaptive moment-based backpropagation algorithm of ANN, known as BPVAM, to classify the same dataset, they combined BPVAM and PCA to use the size reduction technique. They showed with their studies that BPVAM-PCA is more efficient than BPVAM. Ufuk [13] proposed a model of the regression analysis that found a strong correlation between emotional commitment, continuation commitment, and normative commitment and how people perceive promotion techniques. The proposed algorithm was used to determine the personnel to be promoted in a business and priority values were determined for the candidate personnel. Linguistic variables were used to evaluate candidate personnel based on factors, and clarification of fuzzy weights was done with a clarification process developed on the basis of α-shear and optimism index [14]. In Pakistan's banking sector, this study examined the relationship between employee advancement, performance appraisal, and work satisfaction with employee performance. In Smart PLs, 280 bank personnel were sampled using the SEM analysis method. The findings show that private bank executives should give more importance to recognition and incentive policies, as employees expect to be rewarded for their achievements [15]. Employee promotion in both contexts goes with either a superior grade or an increase in pay within the group. It is a continuous procedure based on professional qualification and length of service and is accepted to be one step ahead in the job within a promotion [16]. The author stated that human resources activities such as remuneration, promotion and performance evaluation have a significant and positive connection with the job performance of the employees.

III. METHODS AND MATERIALS

A. Dataset

In this study, a data set of Kaggle's publicly accessible employee values was used [17]. In Table 1, attributes in the analysis is indicated.

TABLE I. ATTRIBUTES USED IN ANALYSIS

Attributes	**Explonation**
employeeid	The employee ID
department	Employee's department
region	Employement region
education	Education Level
gender	Employee Gender
recruitmentchannel	Channel of recruitment for employee
no of trainings	no of other trainings completed in the previous year on soft skills, technical skills, etc.
age	Age of Employee
previous year rating	Employee Rating for the previous year
length of service	Length of service in years
awards_ won	if awards won during the previous year then 1 else 0
avg training score	Average score in current training evaluations
is_promoted: (Target)	Recommended for promotion

B. Imbalanced Dataset

The problem of learning from unstable data has increased with the continued expansion of data availability. This is related to the performance of learning algorithms. The nature of the problem is about efficient transformation. Standard algorithms have deviated to methods such as wrong cost analysis due to even distribution policies. This is a recurring problem. If learning from imbalanced datasets cannot be achieved, prediction accuracy will not increase. In this respect, both internal and external imbalances create problems [18]. Hybrid approaches to this issue are gaining increasing popularity. The minority class may be more important in data mining than the non-minority class because it contains very important and useful information and can affect the general information [19].

a) SMOTE

SMOTE is an approach where the Minority class is instantiated by creating synthetic instances. Oversampling is done by getting a sample of each minority class and presenting synthetic samples along line segments joining any of the nearest neighbors [20]. In order to reduce the problems caused

by oversampling, it was deemed appropriate in SMOTE to create new minority samples instead of weighting the data points. Interpolation is made between neighboring minority class instances. In this way, it focuses on the concept of feature space [21].

b) ROS

The most basic and oldest approach of dealing with unbalanced categorization difficulties is ROS. The random oversampling (ROS) approach balances the class distribution by randomly doubling minority class label samples to approximate the large class label while training the classifier until the desired class ratio is reached [22]. Assume that S is the unbalanced dataset, that S_{Neg} is the negative class (majority class), and that S_{Poz} is the positive class (minority class). ROS's job is to add K samples from the minority/positive class to S at random, ensuring that the two classes in S have the identical sample counts. The number of positive class samples is equal to S_{Pos} ++K in this case, with different K values corresponding to varying degrees of equilibrium for the training dataset. The ROS is complete when S_{Pos}==S_{Neg}.

C. Support Vector Machine (SVM)

The Support Vector Machine is used for classification - linear and non-linear - and for Regression. It is determined which hyperplane is selected to be optimal. The hyperplane is the decision boundary. It works well with linear classification because the decision boundary can be drawn properly, but in non-linear classification, residual data is generated, and it becomes difficult to classify. Kernel function is used when classifying them. It can be used for handwriting analysis, facial analysis etc. [23]. The SVM's ability to do examples such as Handwriting analysis is because it can learn to assign labels to objects as an example.

D. Artificial Neural Network (ANN)

The way the brain works is modeled at this stage. It provides learning Socratically. ANNs designed to simulate human information processing are trained through experience. Neural networks collect every information that occurs in the form of patterns and shapes in the data and gain a great power in the name of learning – predicting and recognizing. Because information is processed in this process, ANNs are known as PEs and balance these inputs and outputs. Since it is a connector-based system, they are also called connective models. In supervised learning, hidden neurons communicate with other neurons. The most commonly used learning algorithm at this stage is the error back propagation. In unsupervised learning, a neuron responds more strongly to a particular input and suppresses the output of other neurons. In collaborative learning, each neuron works together to reinforce their output [24].

E. Random Forest (RF)

This method is preferred to use in the classification and regression analysis that based on creating more than one decision trees. This method stands up to “ensemble learning”. In this learning method, a vast number of decision tree is created with the selection more than one subset in the dataset is called as “bagging method”. Especially random forest is constituted with created a vast number of decision tree and these are combined in order to obtain decision tree which is giving the best results as a result of the classification process of the data is carried out [25].

RF is used to construct a classification or regression rule with good accuracy, or to investigate the relationship of candidate prediction variables for the prediction problem. In the first case, this area is tried to be protected by considering the classification error rate and the mean error square. These two transferred goals can also become common goals. The RF Classifier uses a series of CARTs to make predictions. Trees are created by plotting a subset of training examples using a bagging approach. To produce forest trees, Ntree and Mtry must be known. Random Forest can work quickly on large datasets and offers efficient computational capabilities

F. Evaluation

After trained data with methods, some metrics with reference to evaluation, success and power of the models are acquired. The metrics are acquired by using confusion matrix table. The confusion matrix is the summary evaluation table created for each model which obtained with the methods in the classification analysis. [26]. Evaluation metrics and results which representing to the power and success of the models are accuracy, precision, recall and specificity [27-28].

Accuracy is the ratio of the number all correct classifications to the number of the all classifications. In the other words, it is the result of how many of the data are correctly classified. The equation is given as follow:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (1)$$

Precision is the ratio of how many of the data classified as positive are actually positive. With the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Recall is the ratio of how many of the data are actually positively will be predicted as positive with the equation 3: result is achieved: TP / TP + FN

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Specificity is the ratio of how many of data which are actually negative are predicted as negative. Equation is given on the following:

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

F1 Score is the result which obtained by combining the calculation of the “recall and precision” values. In addition, this value (f1 score) shows the power and performances of the methods, by giving classification results and ratio of the methods used. Equation 5 shows the F1-Score calculation.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

IV. SYSTEM OVERVIEW

Data for employee promotion is obtained from the Kaggle website. The diagram of the overall architecture is indicated in Figure 1. Firstly, data are balanced with imbalanced data techniques and then passed to the next stage. ANN, SVM and RF are applied to balanced data by SMOTE and ROS. Consequently, the model generates the classes that labeled non-promoted as 0 and promoted as 1.

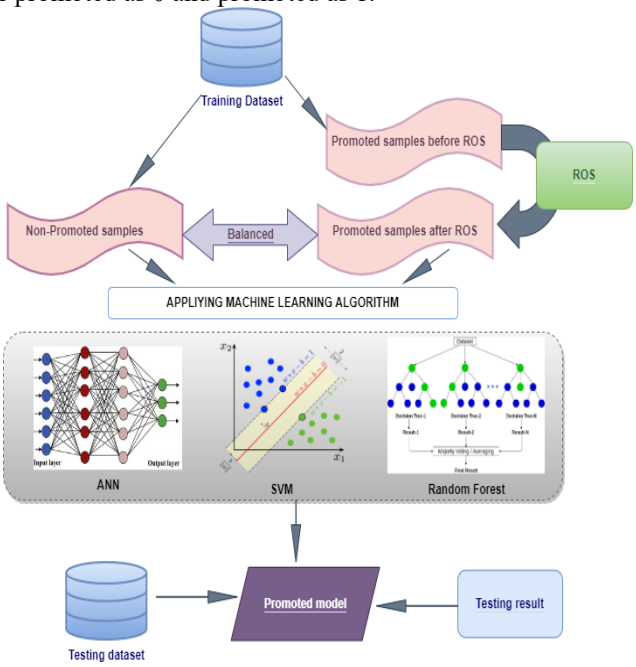

Fig. 1 The proposed system models

V. EXPERIMENTS AND FINDINGS

In this section, data is balanced by SMOTE and RUS techniques and a model is proposed on the prediction of employee promotion by SVM, ANN, and RF algorithms and the findings are shown in detail by Python.

A. Imbalanced Data Methods

In this study, the data set consists of 52399 employees. 47853 non-promoted employees are labeled 0 and 4546 promoted employees are labeled 1.

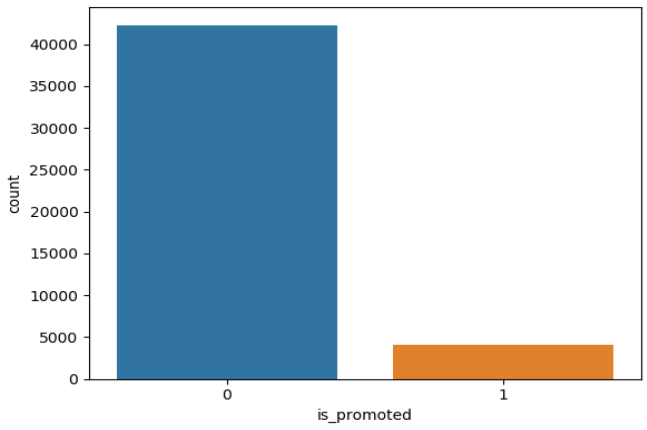

Fig 2. Number of promoted and not promoted

The distribution of the highly imbalanced dataset is indicated in Figure 2. In the study, SMOTE and Random Oversampling techniques are used to reduce the negative effect of the problem of class imbalance on classification.

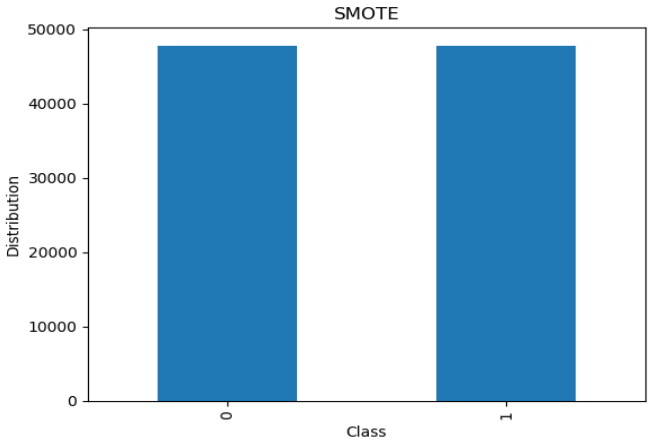

Fig. 3. Dataset after balanced methods

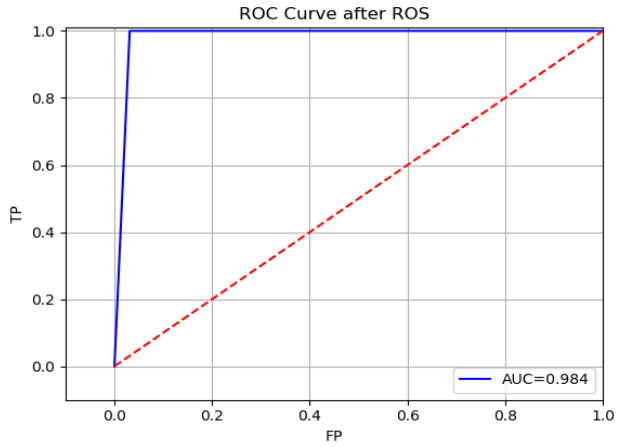

Fig. 4. ROC Curve of RF after ROS

TABLE II. PERFORMANCE METRICS

Algorithms	Performance Metrics			
	Accuracy	*Precision*	*Recall*	*F1-Score*
SVM	0.92	0.867	0.09	0.17
SVM_ROS	0.726	0.756	0.667	0.70
SVM_SMOTE	0.756	0.766	0.736	0.751
ANN	0.9234	0.8266	0.15	0.25
ANN_ROS	0.729	0.788	0.627	0.698
ANN_SMOTE	0.77	0.795	0.728	0.76092
RF	0.9314	0.84	0.247	0.38
RF_ROS	0.9842	0.96	1.0	0.98
RF_SMOTE	0.9157	0.9118	0.92	0.915

This study presents the performance of the SVM, ANN and RF based model in terms of accuracy, precision, recall and F1-Score performance metrics.

The findings from Table II stated that accuracy rate is not enough to select the accurate model for imbalanced dataset. Although accuracy rates are high, recall and f1-score values

are very low without applying any imbalanced techniques. F1-score are evaluated in selection of the model.

The performance of the SVM model was calculated, and the results are shown in Table II. The F1-Score of the SVM, SVM_ROS, SVM_SMOTE are calculated as 0.17, 0.70 and 0.75 respectively.

ANN (0.25), ANN_SMOTE (0.69), ANN_SMOTE (0.76) scores were obtained according to F1-score performance values.

According to the findings obtained as a result of the analysis in Table II, RF model has an obvious advantage for employee promotion prediction and achieve the highest classification performance of predicting promotion with 98% F1-score by using ROS imbalanced technique.

VI. CONCLUSION

Promotions have a favorable, significant and beneficial impact on employee work performance in human resources process. In this study a prediction model for employee promotion is proposed by using RF method.

A decision support system designed for a Human Resource (HR) departments about eligibility of employees' promotion. SMOTE and ROS imbalanced techniques are used. Then, classification algorithms are applied to predict employee promotion such as SVM, ANN and RF. RF outperformed the other algorithms with 98% accuracy, 96% precision, 1.0 recall and 98% f1-score rate obtained among SVM and ANN. This study indicates that F1-score that is the harmonic mean of precision and recall, should be used. The key reason for utilizing F1 Score instead of accuracy is to avoid selecting an inappropriate model in datasets with imbalanced distribution. Furthermore, the F1-score is critical since it is necessary to have a measurement metric that includes not only False Negative or False Positive, but also all mistake costs. This study can be used by HR in the time efficiency of their performance to improve key performance indicator (KPI) KPIs in promoted positions. Besides, it could assist managers in minimizing a person's handicap after receiving a promotion due to a mistake made in the selection of a promotion candidate. For the future work, it is planned to add feature engineering and feature importance to the study by using other data balance techniques.

REFERENCES

[1] O. O. Awosusi, & A. O. Jegede, Motivation and job performances among nurses in the Ekiti State Environment of Nigeria. International Journal of Pharma and BioScience, 2(2), 2011, 583-595.

[2] E. Kiruja, & E. Mukuru, Effect of motivation on employee performance in public middle level Technical Training Institutions in Kenya. IJAME, 2018.

[3] S. Haryono, S. Supardi, & U. Udin, The effect of training and job promotion on work motivation and its implications on job performance: Evidence from Indonesia. Management Science Letters, 10(9), 2020, 2107-2112.

[4] M. R. B. Rubel and D. M. H. Kee, Perceived Fairness of Performance Appraisal, Promotion Opportunity and Nurses Turnover Intention: The Role of Organizational Commitment, Asian Social Science; Vol. 11, No. 9, 2015.

[5] M. R. W. Dean, & M. J. Joseph, Human Resource Management (14th Edition). UK: McGraw-Hill Education, 2005.

[6] M. T, Tessema, J. L. Soeters, Challanges and practices of HRM in developing countries: testing the HRM-performance link in the Eritrean civil service. Int. J. Hum. Res., 17(1): , 2006, 86-105.

[7] M. S., Knowles, E. Holton, & R. Swanson, The adult learner: the definitive classic in adult education and human resource development, 6th ed., Burlington, MA: Elsevier, 2005.

[8] K. Shahzad, S. Bashir and M. I. Ramay, Impact of HR practices on the perceived performance of University teachers in Pakistan. Int. Rev. Bus. 4 (2), 2008.

[9] M. G. McIntyre, Secrets to winning at office politics: How to achieve your goals and increase your influence at work. St. Martin's Griffin, 2005.

[10] A. A. Hameed, A. Jamil, N. Ajlouni, J. Rasheed, A. Özyavaş and Z. Orman, "Classification of Epileptic Seizures using Artificial Neural Network with Adaptive Momentum," *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020, pp. 1-4, doi: 10.1109/ICDABI51230.2020.9325688.

[11] E. N. Mutlu, A. Devim, A. A. Hameed, A. Jamil, "Deep Learning for Liver Disease Prediction". In: Djeddi, C., Siddiqi, I., Jamil, A., Ali Hameed, A., Kucuk, İ. (eds) Pattern Recognition and Artificial Intelligence. MedPRAI 2021. *Communications in Computer and Information Science*, 2022, vol 1543. Springer.

[12] J. Rasheed, A. A. Hameed, N. Ajlouni, A. Jamil, A. Özyavaş and Z. Orman, "Application of Adaptive Back-Propagation Neural Networks for Parkinson's Disease Prediction," *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020, pp. 1-5, doi: 10.1109/ICDABI51230.2020.9325709.

[13] Ü. Ufuk, Bankalarda Terfi Uygulamaları Algısının Örgütsel Bağlılık Üzerindeki Etkisini Belirlemeye Yönelik Bir Araştırma. BDDK Bankacılık ve Finansal Piyasalar Dergisi, 13(2), 2019, 161-184.

[14] M. Dağdeviren, Bulanik analitik hiyerarşi prosesi ile personel seçimi ve bir uygulama. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 22(4), 2007.

[15] A. Kibria, A. S. Kazi, A. A. Banbhan, A. K, Shahani & I. Junejo, Investigating linkages of performance appraisal, employee promotion and job satisfaction with employee performance in banking sector of Pakistan. Journal of Contemporary Issues in Business and Government Vol, 27(2), 2021.

[16] R. B. R. Mohammad and M., H., K. Daisy, Perceived Fairness of Performance Appraisal, Promotion Opportunity and Nurses Turnover Intention: The Role of Organizational Commitment, Asian Social Science; Vol. 11, No. 9, 2015.

[17] https://www.kaggle.com/code/flaviocavalcante/employees-evaluation-for-promotion-eda-ml/data?select=employee_promotion.csv

[18] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

[19] N. V. Chawla, SMOTE: Synthetic Minority Over-sampling Technique, 2002.

[20] A. Ferńandez, SMOTE for Learning from Imbalanced Data: Progress andChallenges, Marking the 15-year Anniversary, 2018.

[21] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, 2016.

[22] F. Rayhan, S. Ahmed, A. Mahbub, R. Jani, S. Shatabda & D.M. Farid,. Cusboost: cluster-based under-sampling with boosting for imbalanced classification. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (pp. 1-5). IEEE, 2017.

[23] M. Somvanshi, A review of machine learning techniques using decision tree and support vector machine. IEEE, 2017.

[24] S. Agatonovic-Kustrin, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Elsevier, 2000.

[25] T. Chen, XGBoost: A Scalable Tree Boosting System. ACM, 2016.

[26] T. M. Oshiro, How Many Trees in a Random Forest?, 2012.

[27] S.Umadevi, & D. Marseline, A Survey on Data Mining Classification Algorithms. International Conference on Signal Processing and Communication, 2017, 64-268.

[28] I. A. Zriqat, , A. M. Altamimi & M. Azzeh, A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods, 2017, 868-879.