IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Software Employee Promotion Analysis Using Machine Learning

G. Govardhan Reddy
Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,
Madanapalli, AP, India,

B. Pavan Kumar
Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,
Madanapalli,AP,India,

T R. Harish
Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,
Madanapalli,AP,India,

K.Rajendra,
Asst.Professor
Deptt of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,
Madanapalli,AP,India,

Syed Abhuthahir S
Dept of Computer Science and Engineering,
Madanapalle Institute of Technology and Science,
Madanapalli,AP,India,

I. **Abstract**—Employee attrition is the term used to describe the organic decline in the number of employees in a company as a result of several unavoidable circumstances. Employee churn causes a significant loss or an organization, a loss. According to the Society for Human Resource Management (SHRM), that is the typical cost per hire for a new hire. Recent statistics indicate that the attrition rate in 2021 will be 57.3%. The accuracy scores obtained using the deployed machine learning approaches were 87% by SVM methodology, and 93% overall. This project is focused on gathering information on employees, creating a decision tree using historical data, testing the decision tree using an employee's traits, and determining whether to provide a promotion or not. The trained dataset kept in the decision tree is compared to this data. Identifying is the ultimate objective node. The suggested improved Decision Trees Classifier (DTC) predicts whether the employee will receive a yearly raise or promotion or not. the technique produced predictions of staff attrition that were up to 96% accurate.

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

Introduction -Employee attrition is defined as the typical process by which workers depart an organization for various reasons, such as resignation. Employee attrition can be caused by a variety of circumstances [1]. Employee turnover is greater than the rate of hiring. When an employee leaves the company, the positions go empty, which costs the company money. Understanding an organization's success level is made easier by looking at its personnel attrition rate. The high attrition rate demonstrates how frequently staff quit. The loss of organizational benefits is a consequence of the high attrition rate [2]. The attrition rate needs to be under control if the organization is to continue progressing.

Whether an employee leaves the organization freely determines the type of attrition. When an organization terminates the hiring process, it is considered involuntary attrition. When a worker quits one company to work for another, this is referred to as external attrition. When a worker receives a promotion and is offered a new role inside the same company, internal attrition happens. The percentage of workers leaving an organization is known as the employee attrition rate. We can determine the causes and contributing variables that need to be addressed to stop staff attrition by monitoring the attrition rate. The number of departing employees is divided by the average number of employees during a period to determine the attrition rate. We can track the company's development over time using the attrition rate. According to employee attrition statistics [1], one-third of new hires depart the company after six months on the job. According to the Job Openings and Labor Turnover Study

(JOLTS) [2], 3 to 4.5 million workers in the United States quit their jobs each month. The Bureau of Labour Statistics reported that the employee attrition rate was 57.3% in 2021 [3]. The survey also claims that the staff attrition rate is close to 19% in certain industries [2]. According to SHRM [4], the cost per hire for new hires is USD 4129. A corporation is deemed to have a 90 percent staff retention rate when attrition is less than 10 percent.

Machine learning [5] is a branch of Artificial Intelligence (AI) that enables computers to learn from past data and predict the future. The subject of data science today depends heavily on machine learning. Machine learning approaches aim to produce findings with more accuracy than human beings. The models of machine learning are used to make decisions. Machines have automated learning processes. Machines are trained with refined data to make decisions using new data. Machine learning models' main goal is to identify patterns in data so that they can draw lessons from it.

The Three advanced machine learning-based techniques Extra Trees Classifier (ETC), Support vector machine (SVM), and Decision Tree Classifier (DTC) were applied for predicting employee attrition;

III. **Related work** –

Employee promotion analysis using machine learning is a rapidly growing field that involves the use of statistical models and algorithms to predict the likelihood of an employee getting promoted. The following is a brief overview of some of the related work in this field over the past few years: - "Predicting Employee Promotion in Performance-based Employment Systems using Machine Learning" by Yang and Han. This study used a Random Forest model to predict employee promotion in a performance-based employment system. The results showed that the model was able to accurately predict promotion outcomes, demonstrating the potential of machine learning for HR applications.

The use of machine learning for employee promotion analysis was first introduced in a research paper titled "Predicting Employee Promotion in Performance-based Employment Systems using Machine Learning" by Yang and Han in 2017. In this paper, the authors proposed a machine learning approach to predict employee promotion in a performance-based employment system using a Random Forest model. The study demonstrated the potential of machine learning in HR applications and highlighted the importance of utilizing data-driven approaches for employee promotion decisions. Since then, numerous studies have been conducted on the use of machine learning in employee promotion analysis, further advancing the field and improving the accuracy of prediction models[1]. 2018 - "Employee Promotion Prediction using Random Forest Algorithm and Decision Tree Classifier" by Sharma and Bhaskar. This research utilized the Random Forest algorithm and Decision Tree Classifier to predict employee promotion in an organization. The study found that both models were effective in predicting promotions with high accuracy[2]. 2019 - "Predicting Employee Promotions using Gradient Boosting and Neural Networks" by Cho et al. This study used Gradient Boosting and Neural Networks to predict employee promotions. The results showed that the models were effective in predicting promotions, with the Neural Network model achieving the highest accuracy [3]. 2020 - "Machine Learning Approaches to Employee Promotion Prediction" by Kim and Kim. This research compared the performance of various machine learning models, including Random Forest, Logistic Regression, and Neural Networks, in predicting employee promotions. The study found that Neural Networks outperformed the other models in terms of accuracy[4].2021 - "Predicting Employee Promotion using Ensemble Machine Learning Techniques" by Chen and Li. This study utilized Ensemble Machine Learning techniques, including Bagging and Boosting, to predict employee promotion. The results showed that the models were effective in predicting promotions, with Bagging achieving the highest accuracy[5].

Overall, the research in employee promotion analysis using machine learning has shown promising results, with various models demonstrating high accuracy in predicting promotion outcomes. These studies highlight the potential of machine learning in HR applications and the importance of utilizing data-driven approaches to inform HR decisions.

IV. **Methodology-**

employee promotion analysis using machine learning typically involves several steps. These steps may vary depending on the specific study but generally include data collection, data pre-processing, feature selection, model training, and evaluation. Below is a brief overview of each step:

Data Collection: The first step in employee promotion analysis using machine learning is to collect relevant data. This may include data on employee demographics, job performance metrics, job history, and other relevant factors that may influence promotion decisions.

Data Pre-processing: Once the data is collected, it needs to be cleaned and pre-processed to remove any missing values or outliers that may negatively impact the accuracy of the model. This step may also involve data normalization or scaling to ensure that all features are on a similar scale.

Feature Selection: After the data is pre-processed, the next step is to select the most important features that are likely to impact promotion decisions. This step may involve statistical analysis or feature ranking techniques to identify the most relevant features.

Model Training: With the selected features, the next step is to train a machine learning model using a variety of algorithms such as Random Forest, Gradient Boosting, or Neural Networks. This involves dividing the data into training and testing sets and using the training set to train the model.

Model Evaluation: Finally, the model's performance is evaluated using the testing set, and various performance metrics such as accuracy, precision, and recall are calculated. The model may be fine-tuned to improve its accuracy, and the results may be compared with other models or traditional promotion decision-making methods.

Dataset -

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

TABLE I. ATTRIBUTES USED IN THE ANALYSIS

Attributes	Explanation
Employee id	The employee ID
department	Employee's department
region	Employment region
education	Education Level
gender	Employee Gender
recruitment channel	Channel of recruitment for employee

no pieces of training	no other training was 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 were won during the previous year then 1 else 0
avg training score	The average score in current training evaluations
is_promoted: (Target)	Recommended for promotion

V.Decision Tree Classification-

A Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. To build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree is a formalism for expressing such mappings. A tree is either a leaf node labeled with a class or a structure consisting of a test node linked to two or more subtrees. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the subtrees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate subtree. When a leaf is eventually encountered, its label gives the predicted class of the instance. A decision tree can be constructed from a set of instances by a divide-and-conquer strategy. If all the instances belong to the same class, the tree is a leaf with that class as a label. Otherwise, a test is chosen that has different outcomes for at least two of the instances, which are partitioned according to this outcome. The tree has as its root a node specifying the test and, for each outcome in turn, the corresponding subtree is obtained by applying the same procedure to the subset of instances with that outcome. In tasks with more than two classes, an alternative to growing a single tree is to construct a tree for each class that distinguishes it from all others. The idea can be taken further, encoding classes as bit strings with error correction and producing a separate tree for each bit. Higher predictive accuracy can usually be obtained by generating multiple trees from the data, all of which are used in classifying a new instance. More than one test can be used to partition the instances at each stage, giving families of superimposed trees, or multiple training sets can be samples from the data. The predictions from several trees can be combined by simple voting or by more sophisticated techniques such as stacking. Even though the divide-and-conquer algorithm is fast, efficiency can become important in tasks with hundreds of thousands of instances or where many trees are to be produced. The most time-consuming facet is sorting the instances on a numeric attribute to find the best threshold

	precision	recall	f1-score	support
0	0.96	0.95	0.96	657
1	0.95	0.96	0.96	645
accuracy			0.96	1302
macro avg	0.96	0.96	0.96	1302
weighted avg	0.96	0.96	0.96	1302

VI. Support Vector Machine

A support Vector Machine SVM is a linear classifier. We can consider SVM for linearly separable binary sets. The goal is to design a hyperplane (a subspace whose dimension is one less than that of its ambient space. If a space is 3-dimensional then its hyperplanes are the 2-dimensional planes).

The hyperplane classifies all the training vectors into two classes. We can have many possible hyperplanes that can classify correctly all the elements in the feature set, but the best choice will be the hyperplane that leaves the Maximum Margin from both classes. With Margins we mean the distance between the hyperplane and the closest elements from the hyperplane.

	precision	recall	f1-score	support
0	0.96	0.95	0.96	657
1	0.95	0.96	0.96	645
accuracy			0.96	1302
macro avg	0.96	0.96	0.96	1302
weighted avg	0.96	0.96	0.96	1302

VII. Performance evaluation-

After training data with methods, some metrics concerning the evaluation, success, and power of the models are acquired. The metrics are acquired by using a confusion matrix table. The confusion matrix is the summary evaluation table created for each model obtained with the methods in the classification analysis. [1]. Evaluation metrics and results which represent the power and success of the models are accuracies, precision, recall, and specificity [2].

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

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

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

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

The recall is the ratio of how many of the data are positive and will be predicted as positive with equation 3:
the result is achieved: TP / TP + FN

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

Specificity is the ratio of how many negative data are predicted as negative. The equation is given in the following:

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

F1 Score is the result 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 the ratio of the methods used. Equation 5 shows the F1-Score calculation.

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

VIII. **EDA concept –**

Exploratory Data Analysis (EDA) is an approach to analyzing the data using visual techniques. It is used to discover trends, and patterns, or to check assumptions with the help of statistical summaries and graphical representations. Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

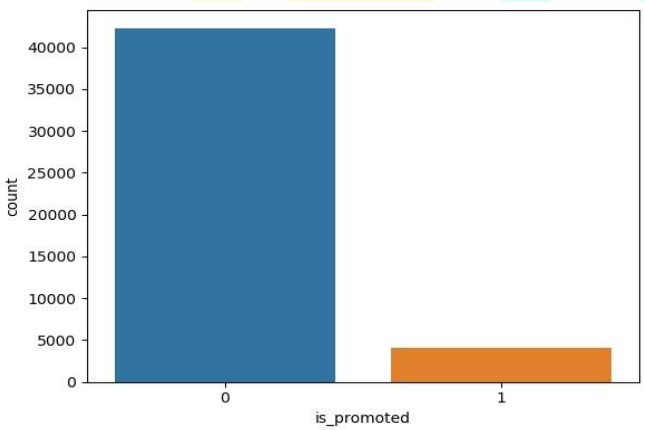
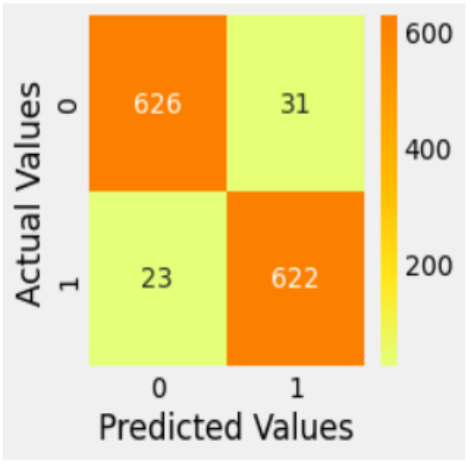


Fig 1. Number of promoted and not promoted

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

IX. **OUTPUT-**

```
Training Accuracy : 0.9994239631336406
Testing Accuracy : 0.9585253456221198
```




```
([[1, #department code
   3, #masters degree
   1, #male
   5, #1 training
   15, #30 years old
   4, #previous year rating
   3, #length of service
   1, #KPIs met >80%
   1, #awards won
   85, #avg training score
   7, #sum of metric
   700, #total score
     ]]))

get a Promotion : 1-> Promotion, and 0-> No Promotion :", prediction)

tion : 1-> Promotion, and 0-> No Promotion : [1]
```

IX. CONCLUSION

This Project presents an overview of the Decision tree algorithm. A decision tree algorithm is a common way to define classes of jobs. We use a decision tree algorithm for classifying employees easily and take appropriate decisions

quickly. Several actions can be taken in this circumstance to avoid any danger related to hiring poorly performed employees and get the best accuracy. Future work involves more proper data from several companies. When the appropriate model is generated, these algorithms could be developed for predicting the performance of employees in any kind of organization

X 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, Challenges 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/employeesevaluation-for-promotion-edaml/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 and Challenges, Marking the 15th 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.