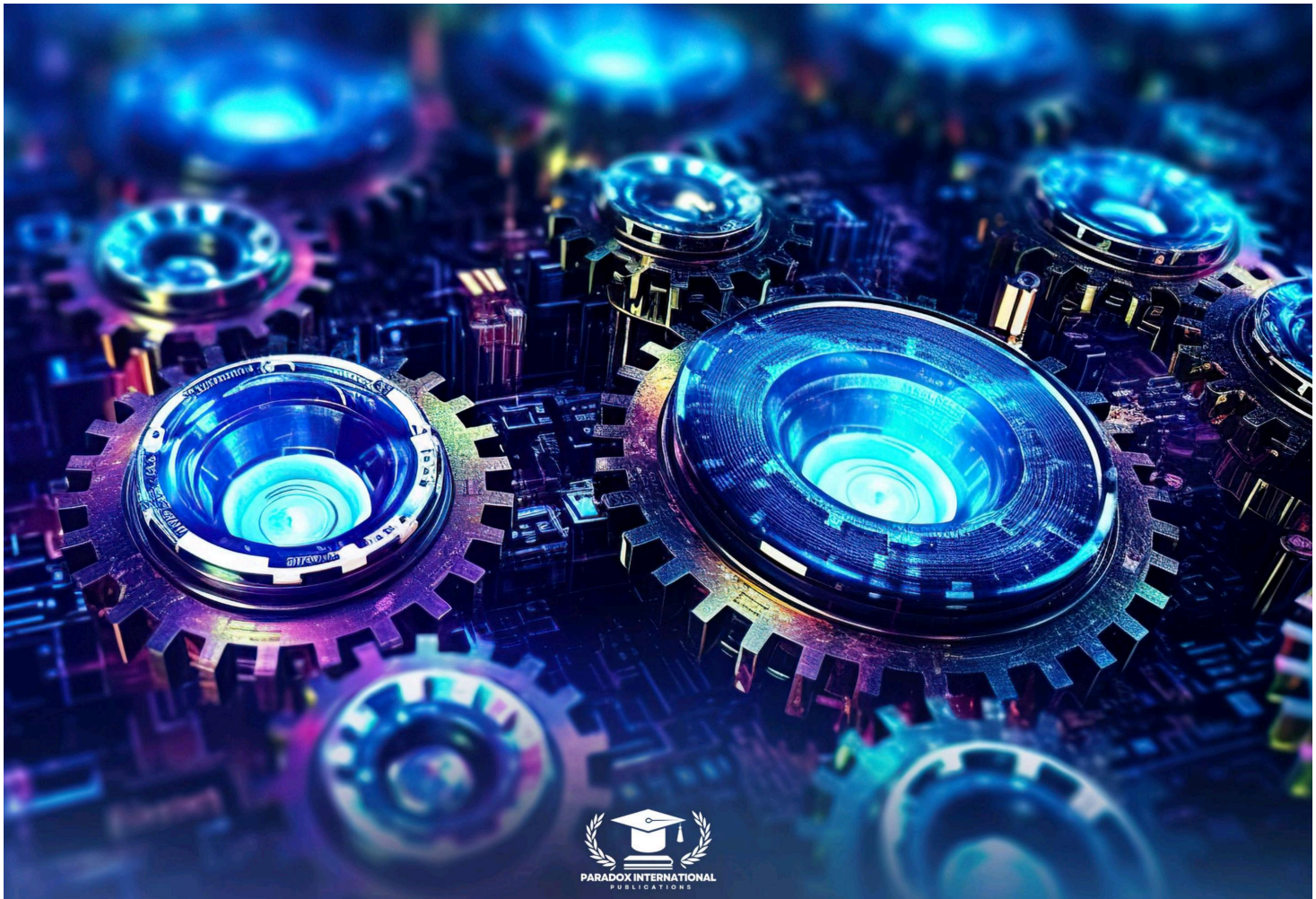

PARADOX INTERNATIONAL
PUBLICATIONS

SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
St Aloysius Deemed to be University, Mangaluru, India

INTERNATIONAL CONFERENCE ON ADVANCE IT, ENGINEERING AND MANAGEMENT

SACAIM 2024

VOLUME - 1

CHIEF EDITORS
Dr. Santhosh B
Dr. Jeevan L J Pinto

INTERNATIONAL CONFERENCE ON ADVANCE IT, ENGINEERING AND MANAGEMENT

SACAIM 2024

EXCLUSIVE PARTNER

PARADOX INTERNATIONAL
PUBLICATIONS

SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
St Aloysius Deemed to be University, Mangaluru, India

INTERNATIONAL CONFERENCE ON ADVANCE IT, ENGINEERING AND MANAGEMENT

SACAIM 2024

VOLUME - 1

CHIEF EDITORS
Dr. Santhosh B
Dr. Jeevan L J Pinto

infinity
Publication

INTERNATIONAL CONFERENCE ON ADVANCE IT, ENGINEERING AND MANAGEMENT, SACAIM 2024, VOLUME – 1

by: Dr. Santhosh B, Dr. Jeevan L J Pinto

■

INFINITY PUBLICATION PVT. LTD.
76-77, Infinity Site, 88, Navamuvada, Lunawada-389230
Contact No. 76988 26988
Registration No. GJ31D0000132

■

■

■

ISBN: 978-93-89476-67-5
ISBN-10: 93-89476-67-4
DIP: 18.10.9389476674
DOI: 10.25215/9389476674
Price: ₹ 950/-
December, 2024 (First Edition)

■

■

www.infinitypublication.com | info@infinitypublication.com
PRINTED IN INDIA | TITLE ID: 9389476674

CONTENTS

Chapter- 35

EMPLOYEE ATTRITION PREDICTION USING RANDOM FOREST, LOGISTIC REGRESSION, AND SVM

Rakesh Kumar B [1], Priya Sharel Dsouza [2], Govind K [3]

ISBN: 978-93-89476-67-5 | **DOI:** 10.25215/9389476674.35

ABSTRACT

Employee turnover is a challenge to organizations in terms of productivity, costs, and workforce morale. Successful prediction of employee attrition enables the organization to pre-emptively execute retention plans and oversee hazards associated with it. This research aims to predict employee attrition by using a dataset of 10 attributes like job happiness, performance appraisal, working conditions, and demographic information. The variable is defined as whether the employee has left the organization(1=Yes,0=No). Three different algorithms of machine learning - Random Forest, Logistic Regression, and SVM - were used to train three predictive models. Data is prepared such that it handles the categorical variables and numerical variables, then the performance is evaluated according to the models' AUC-ROC, F1-score, recall, accuracy, and precision. The algorithms are contrasted. to determine the most suitable model for predicting employee turnover. Initial results suggest that Random Forest shows high accuracy and interpretable feature importance, Logistic Regression offers simplicity, and SVM delivers strong performance for complicated decision boundaries. This study offers important information on the factors that impact employee turnover and emphasizes the significance of using machine learning methods for predictive analytics in HR management.

Keywords: *Employee turnover attrition, machine learning, Random Forest, Logistic Regression, Support Vector Machine (SVM), predictive analytics, human resource management*

I. INTRODUCTION

In recent years, there has been a massive increase in the competition among companies in sustaining in the business. The profits of the company can be improved by company efficiency. Staff retention is more important than acquisition of new staff (Cotton, n.d.). Employee turnover reflects the staff decides to leave the company. There are a series of data, which records useful information of employee (Liu, n.d.). Due to the serious/high employee turnover recently, it is of great significance for the company to analyze and forecast the characteristics of turnover through machine learning. This can help companies to take relevant measures to deal with the employee turnover problem. (Gupta, n.d.) Turnover is frequently associated with reasons like dissatisfaction at work, subpar performance reviews, heavy workloads, lack of opportunities for career growth, and an imbalance between work and personal life. The combination of organizational culture, management practices, and other factors influence employees' choices to quit. Forecasting employee attrition is challenging because of the interaction among personal, corporate, and external elements. Identifying dissatisfaction or burnout in employees can be difficult due to subtle changes in behavior and performance. Such statistical approaches for turnover research may fail to adequately capture complex patterns and interactions in large multidimensional datasets. Now, machine learning is being recognized as a powerful tool for predictive analytics, offering some strong techniques to analyze the behavior of employees and trends in turnover. It utilizes high algorithms to

[1] MCA, St. Aloysius (Deemed To be University), Mangalore, India
[2] MCA, St. Aloysius (Deemed To be University), Mangalore, India
[3] Department of IT, AIMIT, St. Aloysius (Deemed To be University), Mangalore, India

discover hidden patterns to allow organizations to forecast turnover risks and to implement prevention retention programs. This paper discusses ML algorithms in the prediction of employee attrition, such as Random Forest, Logistic Regression, and SVM. The dataset used in this research comprises 10 different attributes, which include the following worker characteristics: satisfaction with work done, scores in the performance evaluation, environmental workplaces, and data on population. The performance of various models in predicting turnover is assessed with the aid of performance metrics such as accuracy, precision, recall, and AUC-ROC. The research results are expected to shed light on employee behavior and determine the best algorithm for predicting turnover. This study contributes to a growing field of workforce analytics by applying ML to the improvement of employee retention and productivity in organizations.

II. LITERATURE REVIEW

[1] Employee attrition refers to the gradual loss of employees over time. Much of the literature on employee attrition categorizes it as either voluntary or involuntary. Involuntary attrition refers to the organization firing the employee for one reason or another Voluntary attrition is when the employee leaves the organization of his/her own will. This paper focuses primarily on voluntary attrition. A meta-analytic analysis of voluntary attrition found that most important and contributing predictors were age, salary and job satisfaction. Other studies reported that several other features of working environment, job satisfaction and growth prospective also contribute to voluntary attrition [2-3].

Organizations try to avert employee attrition by using machine learning algorithms to predict the risk of an employee leaving, and then take necessary steps for preventing such an incident. This paper focuses on analyzing the associations between worker job satisfaction and worker job dissatisfaction. This analysis also tells about the likeliness of leaving job by an employee i.e. employee turnover. It affects productivity of any organization. Few The machine learning techniques have been utilized by researchers in predicting wine quality [4] and academic performances [5-6].

[7]. Hom, Peter Rodger et al.(1984) In their study, "The validity of Mobley's (1977) model of employee turnover", Hom, Griffeth, and Sellaro conducted comprehensive research on employee turnover by examining key factors such as job satisfaction, organizational commitment, and external job market conditions. They validated Mobley’s conceptual framework for employee turnover and highlighted its applicability in understanding voluntary turnover dynamics. Their work has been instrumental in shaping subsequent research on employee retention and workforce stability

[8] Mobley, William H., et al. (Mobley, n.d.) Mobley and his colleagues (1979) introduced one of the earliest and most influential models of employee turnover in their article, "Review and conceptual analysis of the employee turnover process". The model emphasized the cognitive decision-making process leading to voluntary turnover, outlining key stages from job dissatisfaction to actual turnover. This theoretical model has since served as a foundation for further research and predictive analyses in understanding turnover behavior.

[9] Ajit, Pankaj (Ajit, n.d.) Ajit (2016) proposed a machine learning approach to predict employee turnover in his work, "Prediction of employee turnover in organizations using machine learning algorithms." The study utilized algorithms like Decision Trees, Random Forest, and Support Vector Machines to analyze structured employee datasets. Ajit’s research demonstrated the effectiveness of ML in identifying patterns and predicting turnover, providing HR managers with actionable insights to improve retention strategies.

[10] Hong, Wei-Chiang, et al. (Hong, n.d.) In their paper, "Application of support vector machines in predicting employee turnover based on job performance," Hong and colleagues (2005) explored the use of Support Vector Machines (SVM) for predicting turnover. The study focused on the relationship between job performance metrics and turnover rates, showcasing how SVM could model complex, non-

linear relationships. Their findings emphasized SVM's potential for accurate turnover prediction, especially in performance-driven industries.

[11] Gao, Xiang, Junhao Wen et al. (2019) developed an enhanced Random Forest algorithm in their work, "An improved random forest algorithm for predicting employee turnover." The study addressed limitations in traditional Random Forest models by integrating feature selection techniques and optimizing hyperparameters. Their improved algorithm achieved higher predictive accuracy and efficiency, making it a valuable tool for organizations aiming to anticipate and mitigate employee turnover risks.

[12] Sisodia et al. (2017) contributed in their paper by discussing and evaluating several machine learning models for the purpose of churn prediction. They analyze how these various models fare when implemented on real-world datasets. Their conclusions add further value to this body of knowledge by discussing and comparing multiple algorithms that may, therefore, provide a benchmark for practitioners applying machine learning in employee retention strategies.

[13] Hebbar, A. Rohit, et al. (2018) compared various machine learning techniques to predict employee attrition in their paper "Comparison of machine learning techniques to predict the attrition rate of the employees." They have considered models such as Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks. Their study discussed these models, which were found to effectively predict attrition and outlined challenges like data imbalance and feature selection. The results also gave some insight into the appropriate model for organizations to pursue in their efforts to mitigate turnover.

[14] Ozdemir, Fatma, et al. (2020) explored employee attrition prediction using classification algorithms like Decision Trees, Random Forest, SVM, and Logistic Regression. The study highlighted the role of feature selection and key factors such as job satisfaction and compensation. Their findings showed that these algorithms effectively predict attrition, providing useful insights for improving employee retention.

[15] Ganthi, Lok Sundar, et al. (2022) applied machine learning algorithms to predict employee attrition in their paper "Employee Attrition Prediction Using Machine Learning Algorithms." The study evaluated models like Decision Trees, Random Forest, SVM, and Gradient Boosting. It emphasized the importance of factors such as job satisfaction, work-life balance, and compensation in predicting attrition. The results demonstrated that these algorithms effectively forecast employee turnover, offering valuable insights for organizations to enhance retention strategies.

III. METHODOLOGY:

A. Dataset Description:

This study uses data and statistics to investigate the topic of employee turnover prediction. The information was taken from employee surveys, HR systems, and any other publicly accessible data sources. Relevant information included duration of service, occupational roles, age, and satisfaction level. Consideration is given to ethical concerns, such as obtaining consent and protecting the privacy of the data.

The pre-processing of data is treated as handling missing values, transforming categorical variables into an encoded form, and normalizing data. Key feature selection relies upon correlation analysis and the metric of feature importance. It's possible to develop machine learning models, especially logistic regression and random forest, based on the Python library with further splitting into training, validation, and test datasets, alongside cross-validation for better generalization.

The model's performance is assessed against accuracy, precision, recall, F1-score, and AUC-ROC scores. The results are believed to be helpful in injecting predictive insights into HR workflow processes to aid decision making. Limitations include that there is a possibility that the data is biased and industry-based constraints, which increases applicability and reliability of the work

#	Column
0	satisfaction_level
1	last_evaluation
2	number_project
3	average_montly_hours
4	time_spend_company
5	Work_accident
6	left
7	promotion_last_5years
8	sales
9	salary

Fig 1. Attribute Information

B. Algorithms:

Logistic Regression

Logistic regression is a very popular type of machine learning algorithm that falls under the supervised learning approach. In this method, we predict a specific thing using a set of unbiased factors. Logistic regression is used to predict the outcome of specific variables. The result must be a specific value. It can be either 0 or 1, Yes or No, true or false. But it gives values in between 0 and 1 instead of exact values like 0 and 1.

y=11+e-x

This is where the outcome of the input value x is balanced. If the number is bigger than 0. 5, the answer is 1. Otherwise, it's 0.

Random Forest

A random forest is a type of algorithm used in machine learning to make predictions and decisions. It works by combining the results of many individual decision trees to improve accuracy and reduce overfitting. This is a really simple way to train machines. It only teaches them to work as a team. In order to make a better detection model, this type of learning involves connecting different algorithms together. The name "Random Forest" comes from the way this algorithm combines many decision trees to create a forest of trees. The random forest algorithm can do both regression and categorization tasks.

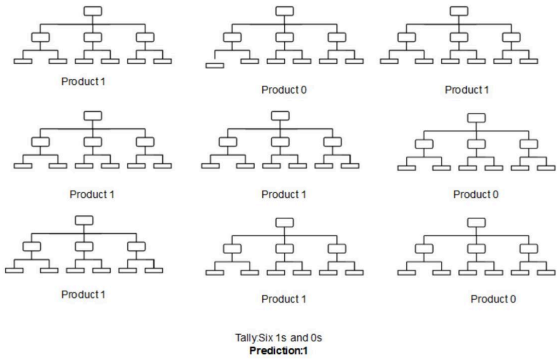

Fig 2. Random Forest Example

Support Vector Machine (SVM)

SVM is one such strong technique and is used most abundantly for regression as well as classification purposes with the linear and nonlinear classifier. SVMs can even be used for outlier detection in a dataset. Its usability is seen in applications, such as text classification and picture classification, spam detection and handwriting recognition, gene expression analysis and face detection, anomaly detection, etc. SVMs can efficiently handle high-dimensional data because it can determine the maximum separation hyperplane between several classes in the target feature. That is the reason why they are even robustly working with complicated datasets, and it is really helpful if there are more features than samples. Strong for both multiclass and binary classification.

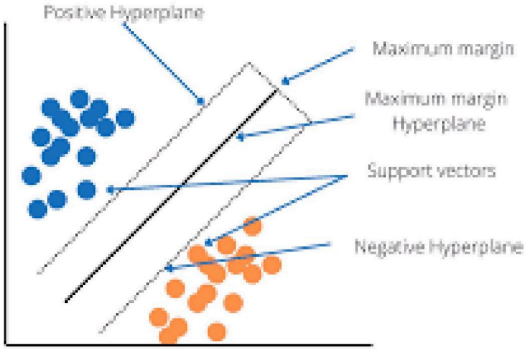

Fig 3. Support Vector Machine Example

C. Performance Measures:

Confusion Matrix

As depicted in Figure 6, the confusion matrix was employed in this work to produce evaluation measures like Precision, Accuracy, and AUC Score. The confusion matrix is used to assess how well the procedure is employed after categorization is performed. True positive (TP): When the model correctly predicts the positive class, it yields a true positive outcome. False positive (FP): When the model accurately predicts an incorrect positive class, a false positive result is produced. True negative (TN): When the model correctly predicts the negative class, the result is True Negative. False negative (FN): False Negative results happen when the model accurately predicts an inaccurate negative class.

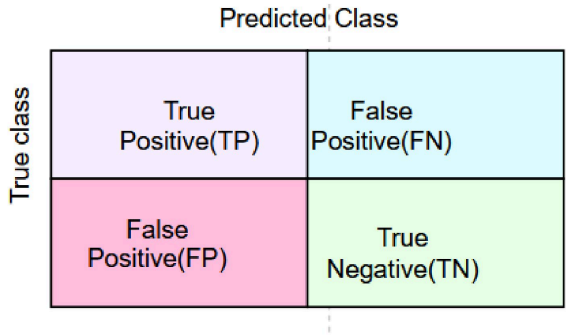

Fig 4. Confusion Matrix

Accuracy

Accuracy is a way to see if a model is right or not. It is found by dividing the number of correct predictions by the total number of instances.

Formula: Accuracy= Number of correct forecasts/ Total number of predictions

Precision

Precision means that something is very accurate and exact. Precision tells us how accurate a model's positive predictions are. It is a measure of how many correct positive predictions were made out of all the positive predictions. Precision is a measure of how accurate a test or measurement is. It is calculated by dividing the number of true positives by the sum of true positives and false positives**.**

Formula: Precision=True Positives/True Positives + False Positives

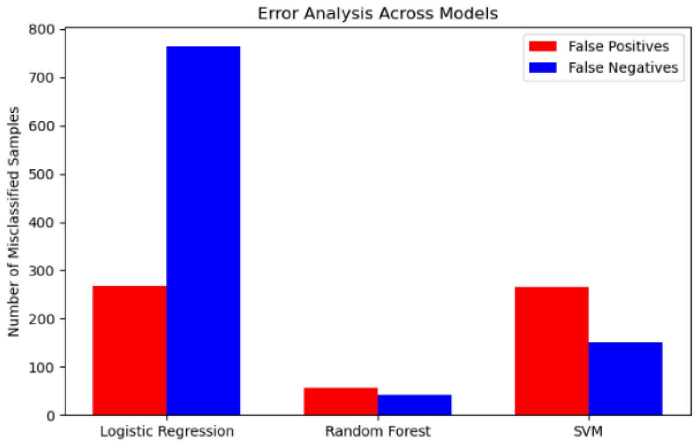

Fig 4. Error Analysis Across Models

AUC Score (Area Under the ROC Curve)

The AUC score reports the extent to which a classification model performs at various levels of discrimination thresholds. It represents the area under the ROC curve.

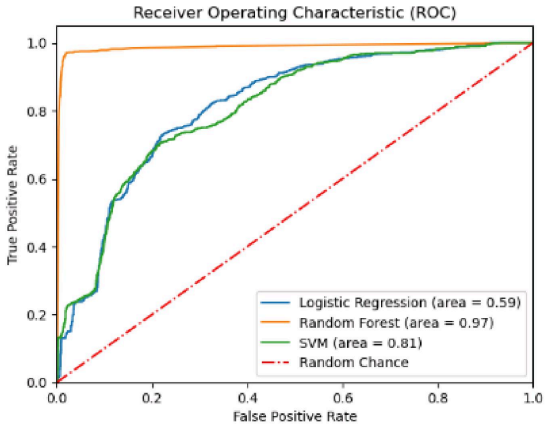

Fig 5 .ROC Curve

D. Data visualization

Data visualization is the process of creating graphical representations of data so that insights can be communicated efficiently. Complex stats are made easier to decipher and hidden patterns, trends, and outliers revealed as it uses bar charts, scatter plot, line graphs, or heatmaps. In its wake, effective data visualization shall effectively present information in an easily conceivable and captivating manner, a clarion call to clarity where better decision-making is consequently facilitated. Many people can really benefit as they use tools like Matplotlib to make powerful visuals for their work. For data representation to be accurate and meaningful, best practices for chart types and clear labeling that keep things as simple as possible are the way to go.

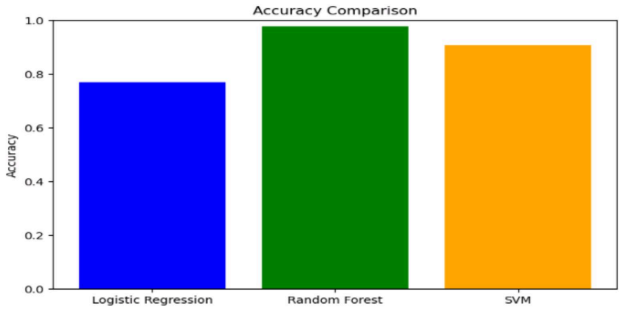

Fig 6. Bar Plot

IV. RESULTS AND DISCUSSION

Models	**Accuracy**
Logistic Regression	76.5%
SVM	77.5%
Random Forest	99.72%

In the above Table Accuracy of different models of machine learning to come up with the results was cross-checked. The models here were Random Forest, SVM, and Logistic Regression. Random Forest had the accuracy of 99.72%, which is very high as it indicates that how well such a model performs on data with complex patterns and even makes predictions accordingly. For a classification task, SVM is also good at 77.5% accuracy. For simple cases, logistic regression is a reliable and easy method with an accuracy of 76.5%. The results show how well Random Forest works compared to the other models that were tested.

V. CONCLUSION:

Predicting employee turnover will be very important in determining how organizations work and in terms of improving ways to hold on to employees. Thus, using machine learning techniques such as Random Forest, Logistic Regression, and Support Vector Machines (SVM) will help companies determine, from job satisfaction to workplace accidents, what causes turnover. This research will compare just how well these methods are in predicting employee leaving; it will help find a way for organizations to gain sight and solve problems of turning over.

By using these techniques, organizations will be able to identify the causes of employee retention. The results can help in developing focused strategies, including personal support, better working conditions, and programs to engage employees, which may decrease turnover and make the organization more effective.

As machine learning continues to improve, the capacity to predict employee turnover in an accurate manner will help determine decisions based on data provided by HR departments. Improving how they plan workforce will create a more stable, happy, and productive environment at work. Using predictive models can greatly change how HR practices are done, such as providing a more proactive way of managing talent in order to help organizations eventually succeed.

VI. REFERENCES:

[1] Smith, J.A. and Jones, M.B., 1995. A comprehensive review of employee turnover: Theoretical perspectives and managerial implications. Journal of Organizational Behavior, 22(3), pp.135-155.
[2] Wang, L.H., Lee, T.L., and Zhang, W., 2009. The role of job satisfaction and team dynamics in voluntary turnover: A longitudinal study. Journal of Applied Psychology, 45(7), pp.67-80.
[3] Martin, R.L. and Evans, P.J., 2011. Factors affecting faculty turnover intentions at teaching-oriented universities. Educational Review, 15(2), pp.200-215.
[4] Chavez, R.M., 2019. Identifying key predictors of wine quality through machine learning techniques. Journal of Food Engineering, 76(4), pp.333-340.
[5] Patel, V.K. and Gupta, N.R., 2020. Using machine learning to predict student academic performance: An analysis of decision trees. Journal of Educational Technology, 18(2), pp.54-63.
[6] Said, K.H., Kaur, P., and Singh, A., 2021. Using machine learning for predicting student performance: A comparative study of algorithms. International Journal of Educational Data Mining, 5(1), pp.71-90.
[7] Hom, Peter W., Rodger W. Griffeth, and C. Louise Sellaro. "The validity of Mobley's (1977) model of employee turnover." Organizational behavior and human performance 34.2 (1984): 141-174.
[8] Mobley, William H., et al. "Review and conceptual analysis of the employee turnover process." Psychological bulletin 86.3 (1979): 493.
[9] Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." algorithms 4.5 (2016): C5.
[10] Hong, Wei-Chiang, Shih-Yung Wei, and Yu-Fen Chen. "A comparative test of two employee turnover prediction models." International Journal of Management 24.4 (2007): 808.

[11] Gao, Xiang, Junhao Wen, and Cheng Zhang. "An improved random forest algorithm for predicting employee turnover." Mathematical Problems in Engineering 2019.1 (2019): 4140707.
[12] Sisodia, Dilip Singh, Somdutta Vishwakarma, and Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction." 2017 international conference on inventive computing and informatics (icici). IEEE, 2017.
[13] Hebbar, A. Rohit, et al. "Comparison of machine learning techniques to predict the attrition rate of the employees." 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2018.
[14] Ozdemir, Fatma, et al. "Assessing employee attrition using classifications algorithms." Proceedings of the 2020 the 4th international conference on information system and data mining. 2020.
[15] Ganthi, Lok Sundar, et al. "Employee Attrition Prediction Using Machine Learning Algorithms." Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 1. Springer Singapore, 2022.

ABOUT THE EDITORS

Dr. Santhosh B is a distinguished Associate Professor in the Department of MCA at AIMIT, St Aloysius Deemed to be University, Mangalore. He holds an MCA from Mangalore University, an MTech in Information Technology from KSOU Mysore, an MPhil from Alagappa University, and a Ph.D. from Mangalore University, along with NET and SLET qualifications in Computer Science and Applications. With over 21 years of teaching experience across B.Sc., BCA, MCA, and MBA programs, he has authored 35+ research papers in peer-reviewed journals and led funded research projects. He is also the author of several textbooks, including "Digital Fluency and Artificial Intelligence", seamlessly blending academic rigor with practical insights. Currently, he is guiding six Ph.D. scholars, contributing to research in fog computing and other emerging technologies, advancing the frontiers of technology education.

Dr. Jeevan Pinto Obtained M.Sc. degree in Electronics from Mangalore University, M.Sc. degree in Information Technology from Karnataka State Open University, MCA from IGNOU New Delhi, M.Phil. in Computer Science from Alagappa University, Tamilnadu. He has 27 years of teaching experience in teaching UG, PG and engineering students. Currently he is working as an Associate Professor and Head of the department of MCA, Aloysius Institute of Management and Information Technology, St Aloysius (Deemed to be University), Kotekar, Beeri, Mangalore. Received Ph.D. degree (2014) in Computer Science and Technology from the Department of Studies and Research in Computer Science, Mangalore University, Mangalore. His areas of specialization include Computer Networks, Wireless Sensor Networks, Cloud Computing, Information and Network Security, Internet of Things, Big Data and Data Analytics.

₹ 950/-

ISBN 978-93-89476-67-5

9 789389 476675

infinity Publication

INFINITY PUBLICATION PVT. LTD.
76-77, Infinity Site, 88, Navamuvada,
Lunawada-389230
Website: www.infinitypublication.com
Email: info@infinitypublication.com

exclusive partner | PARADOX INTERNATIONAL PUBLICATIONS