Analytical Survey on Prediction of Employee Attrition Non parametric tuning Algorithms

G.Pratibha [1], Dr. Nagaratna P Hegde [2]

[1] Research Scholar, Dept of CSE, JNTUH,Hyderabad, India.

[2] Professor, Dept of CSE, Vasavi College of Engineering,Hyderabad, India.

Abstract:

Employee attrition is one of the most serious issues facing companies today. When long-term employees leave the company, it impacts the company's relationship with the customer, which in turn affects the company's revenue if the person who replaces the previous employee is unable to maintain a good rapport with the client. These studies evaluate the employee attrition rate through relevant factors such as Job Role, overtime, and job level, which all have a significant impact on attrition. The study includes a survey of various classification techniques, such as logistic regression, ridge classification, decision trees, and random forests, to forecast the likelihood of attrition of every new employee. A systematic and comprehensive evaluation approach is used to assess the performance of each of these supervised machine learning methods. This survey will assist human resource managers in identifying individuals who are likely to leave the firm and forecasting the reasons for their choice, allowing HR managers to design a retention strategy or seek a replacement..

Keywords: *Employee Attrition, Machine Learning, Random Forest, Naive Bayes, Deep learning, Association technique*

I. INTRODUCTION

Data has become a strategic asset for most organizations in a variety of industries, particularly those involved in business processes. Adoption of new technology improves many organizations [1], and data collection, management, and analysis provide significant benefits in terms of efficiency and competitive advantage. Analyzing vast amounts of data can lead to better decision-making processes, attainment of pre-established company goals, and increased business competitiveness [2, 3]. There are various areas within organizations where the use of artificial intelligence impacts a company's decision-making processes [4,5]. Human resources (HR) have received more attention in recent years, as employee quality and skills are a growth element and a true competitive advantage for businesses [6]. Indeed, after being more widely used in sales and marketing, artificial intelligence is now being used to drive company decisions about their personnel, to base HR management decisions on objective data analysis rather than subjective considerations [7–9].

Predicting employee attrition allows management to act more quickly by improving internal policies and initiatives. Where talented employees who are at risk of leaving might be offered a few recommendations, such as a pay raise or suitable training, to lessen their likelihood of leaving. Machine learning models can assist firms in anticipating staff attrition [10, 11]. Analysts can construct and prepare a machine learning model that predicts workers leaving the firm using previous data stored in human resources (HR) departments. It built such models to investigate the relationship between the characteristics of both active and terminated personnel. Furthermore,

contented, highly motivated, and loyal personnel form the foundation of a firm and have an impact on an organization's productivity.

In the literature, some authors suggested retaining only happy and motivated employees as they tend to be more creative, productive and perform better, which in the end generates and sustains improved firm performance [**12, 13**]. There are many ways of transforming the employee data into a single table. We can create a single table representing every employee present in the organisation on 1 January 2019, with columns for values such as the time they have spent in the organisation, and a final column set to TRUE or FALSE (a Boolean value), indicating whether they left the organisation by 31 January or not. This can be the training data.

Employee ID	Days at pay grade	Days in organisation	Age	Distance	Left?
2218	425	887	42	46	TRUE
2364	254	254	43	58	FALSE
5498	369	369	36	21	FALSE
1548	87	201	29	5	FALSE

Figure 1: An example employee table

The purpose of this study serves to predict the employee who is willing to leave the company and also the employees that could be dismissed with having the least repercussions in the company. It aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes. The objective of this survey is to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes by training the model based on previous attrition data available and predicting it in future for better company HR management. This study uses documented attributes affecting employee attrition to predict and does not consider undocumented factors that may lead to attrition.

II.   RELATED WORKS

The study of Prediction of Employee Attrition with work-place related variables uses classification models for work-place variables rather than using demographic or behavioral variables [**14**]. The results suggest a model for Artificial Neural Network dominates with the maximum accuracy. The most influential variable turned out to be the attrition of the managers which somehow further triggers employee attrition. This approach suggests that work-place related policies are easy to formulate in an organization but demographic and behavioral aspects were yet to be studied in detail for future work.

Problem Statement:

Human Resources are critical resources of any organization. Organizations spend huge amount of time and money to hire and nurture their employees. It is a huge loss for companies if employees leave, especially the key resources. Reasons for attrition can be plenty and range from dissatisfaction due to low salaries, less or no career growth opportunities, inferior employee supervision, eagerness to get into companies with global presence, lack of recognition, lack of freedom of expression in the organization and underutilization of talents and skills of the individuals. Thus in a situation when more and more employees are quitting the organization, the attrition rate is on a rise. So if HR can predict weather employees are at risk for leaving the company, it will allow them to identify the attrition risks and help understand and provide necessary support to retain those employees or do preventive hiring to minimize the impact to the organization.

Review of Employee Attrition Rate Prediction Using Machine Learning techiniques:

In [15], the authors conducted numerical experiments for real and simulated human resources datasets representing organizations of small-, medium- and large-sized employee populations are performed using (1) a decision tree method; (2) a random forest method; (3) a gradient boosting trees method; (4) an extreme gradient boosting method; (5) a logistic regression method; (6) support vector machines; (7) neural networks; (8) linear discriminant analysis; (9) a Naïve Bayes method; and (10) a K-nearest neighbor method. Through a robust and comprehensive evaluation process, the performance of each of these supervised machine learning methods for predicting employee turnover is analyzed and established using statistical methods. Additionally, reliable guidelines are provided on the selection, use and interpretation of these methods for the analysis of human resources datasets of varying size and complexity.

Human resource problems, unlike physical systems, cannot be defined by a scientific-analytical formula. As a result, machine learning approaches are the most effective instruments for achieving this goal. In [16], the authors proposed a feature selection strategy based on a Machine Learning Classifier is proposed to improve classification accuracy, precision, and True Positive Rate while lowering error rates such as False Positive Rate and Miss Rate. Different feature selection techniques, such as Information Gain, Gain Ratio, Chi-Square, Correlation-based, and Fisher Exact test, are analysed with six Machine Learning classifiers, such as Artificial Neural Network, Support Vector Machine, Gradient Boosting Tree, Bagging, Random Forest, and Decision Tree, for the proposed approach. By combining Chi-Square feature selection with a Gradient Boosting Tree classifier improves employee attrition classification accuracy while lowering error rates.

Any organization or company is strongly aware of the significance of employees in gaining and upholding competitive advantage. While putting concentration on earning maximized profit, employee attrition rates should be considered as an interfering factor. In [17], the authors proposed emphasizes on predicting attrition probabilities beforehand by implementing an automated tool. The proposed system implements feed-forward neural network along with 10-fold cross validation procedure under a single platform for predicting employee attrition. This proposed method is evaluated as well as compared with six classifiers such as Support Vector Machine, k-Nearest Neighbor, naïve bayes, Decision Tree, Adaboost, and Random Forest classifiers. Experimental analysis concludes that proposed method outperforms well over aforementioned classifiers in terms of performance measure metrics.

In [18], the authors presented a three-stage (pre-processing, processing, post-processing) framework for attrition prediction. An IBM HR dataset is chosen as the case study. Since there are several features in the dataset, the "max-out" feature selection method is proposed for dimension reduction in the pre-processing stage. This method is implemented for the IBM HR dataset. The coefficient of each feature in the logistic regression model shows the importance of the feature in attrition prediction. The results show improvement in the F1-score performance measure due to the "max-out" feature selection method. Finally, the validity of parameters is checked by training the model for multiple bootstrap datasets. Then, the average and standard deviation of parameters are analyzed to check the confidence value of the model's parameters and their stability. The small standard deviation of parameters indicates that the model is stable and is more likely to generalize well.

Review of Employee Attrition Rate Prediction Using Deep Learning techiniques:

In [19], the authors proposed the deep learning technique along with some preprocessing steps to improve the prediction of employee attrition. Several factors lead to employee attrition. Such factors are analyzed to reveal their intercorrelation and to demonstrate the dominant ones. Our work was tested using the imbalanced dataset of IBM analytics, which contains 35 features for 1470 employees. To get realistic results, derived a balanced version from the original one. Finally, cross-validation is implemented to evaluate our work precisely. Extensive experiments have been conducted to show the practical value of our work. The prediction accuracy using the original dataset is about 91%, whereas it is about 94% using a synthetic dataset.

Review of Employee Attrition Rate Prediction Using Optimization techiniques

In [20], the authors proposed machine learning techniques to predict the probability of employee attrition. The study improves the accuracy of employee attrition prediction by developing an enhanced model using a Deep belief network or DBN. A deep belief network is a form of deep neural network made up of several layers of variables that are latent or hidden. This work uses the Restricted Boltzmann machine, which creates a stack of the network to analyze the pattern of the attrition dataset. The parameters involved in Deep Belief Network are fine-tuned by adapting a novel behavioral inspiration algorithm instead of random assignment of the values. The algorithm used here is the meta heuristic Grey Wolf algorithm which is an optimization algorithm that imitates the hunting behavior of Grey Wolves. Thus, it will increase the performance of the proposed model.

In [21], the authors proposed a new technique inspired by grey wolves called Grey Wolf Optimizer hereby shortened to GWO. This algorithm is inspired by the hunting methodology of grey wolves and their way of leadership in nature. For the purpose of simulation, alpha, beta, omega, and delta type of grey wolves are taken. Also the three steps, that is, searching, encircling, and attacking the prey, utilized for the purpose of hunting by wolves are implemented in the algorithm. After that, bench marking was done against 29 test functions. In [22] with other research scholars presented a new way to provide a solution for economic load–dispatch problem (convex). This was done by using a Grey-wolf inspired meta heuristic known as Grey-wolf optimization. In [23], the authors studied about Particle swarm optimization or PSO in short is an optimization and computational search method which draws its inspiration from biology. The social behaviors of fish schooling and birds flocking formed its basis. The algorithm draws inspiration from the behaviour of those animals who do not have any leader in their swarm or group, such as fish schooling or bird flocking. The reason behind that lies in the fact that such flocks with no leaders find food at random and then will follow an animal of the group that is nearest to the source of food, i.e. a solution (potential).

Review of Employee Attrition Rate Prediction on Big data

In [24], the authors proposed a people analytics approach to predict employee attrition that shifts from a big data to a deep data context by focusing on data quality instead of its quantity. In fact, this deep data-driven approach is based on a mixed method to construct a relevant employee attrition model in order to identify key employee features influencing his/her attrition. In this method, we started thinking 'big' by collecting most of the common features from the literature (an exploratory research) then tried thinking 'deep' by filtering and selecting the most important features using survey and feature selection algorithms (a quantitative method). Secondly, this attrition prediction approach is based on machine, deep and ensemble learning models and is experimented on a large-sized and a medium-sized simulated human resources datasets and then a real small-sized dataset from a total of 450 responses. This approach achieves higher accuracy (0.96, 0.98 and 0.99 respectively) for the three datasets when compared previous solutions. Finally, while rewards and payments are generally considered as the most important keys to retention, our findings indicate that 'business travel', which is less common in the literature, is the leading motivator for employees and must be considered within HR policies to retention.

Table 1. below briefly identifies and documents the literature review findings. Subsequent sections include the studies of different papers indicating the inadequacy of certain models and their recommended models which demonstrate successful results.

Table 1: Summary of previous work.

Reference no.	Techniques used	Recommended model	Dataset used
Ref [25]	Naive Bayes and Random Forest	Naïve Bayes (Accuracy: up to 85%)	Sample dataset of sales agents
Ref [26]	ID3 Decision tree, CART Decision tree	CART Decision Tree (Accuracy: 90%)	Kaggle sample dataset
Ref [27]	Logistic regression model (logit), probability regression model (probit)	Logistic regressio n model (logit) (Accuracy: 90.5%)	Custom dataset drawn from a motor marketing company in Taiwan
Ref [28]	Naïve Bayes, Support Vector Machines, Logistic Regression, Linear discriminant analysis, Random Forests, KNN, XGBoost	XGBoost (Accuracy: 88%)	Dataset of a certain level of employees in a Particular leadership team of a global retailer
Ref [29]	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Treesand Random Forests	SVM (Accuracy: 80%)	Particular Sample dataset from HR department of three companiesin India
Ref [30]	KNN, Naive Bayes, MLP Classifier, Logistic Regression	KNN (Accuracy: 94.32%)	Kaggle sample dataset

III. SYSTEM MODEL

Initially the data is downloaded from kaggle is preprocessed first so that we can extract important features like Monthly Income, Last Promotion Year, Salary Hike and etc. that are quite natural for employee attrition. Dependent variables or Predicted variable are the one that helps to get the factors that mostly dependent on employee related variables. For example the employee ID or employee count has nothing to do with the attrition rate. Exploratory Data Analysis is an initial process of analysis, in which we can summarize characteristics of data to can predict who, and when an employee will terminate the service. The system builds a prediction model by using random forest technique. It is one of the ensembles learning technique which consists of several decision trees rather than a single decision tree for classification. The techniques perform dependent variable analysis and word formation vector to evaluate the employee churn. Hence, by improving employee assurance and providing a desirable working environment, we can certainly reduce this problem significantly. When we apply a wide range of data mining techniques from as simple as Naive Bayes, linear regression and nearest neighbours to more complex techniques as SVM, Random Forests and other ensemble methods.

Usecase:

When an employee quits, an organization stands to lose not only a valuable skill-set but the team structure, rhythm, and overall productivity gets negatively affected. Plus, finding a suitable replacement could be a time-consuming process, resulting in other employees to be burdened with extra work and responsibilities for which they may not possess the requisite skill-set. According to an estimate by the Center for American Progress, replacement costs for a highly-skilled employee might be well over 200% of their salary. More the attrition, the higher will be the cost incurred. Thus, companies are leaning on predictive analytics using machine learning to curb employee churn as shown in Figure 2.
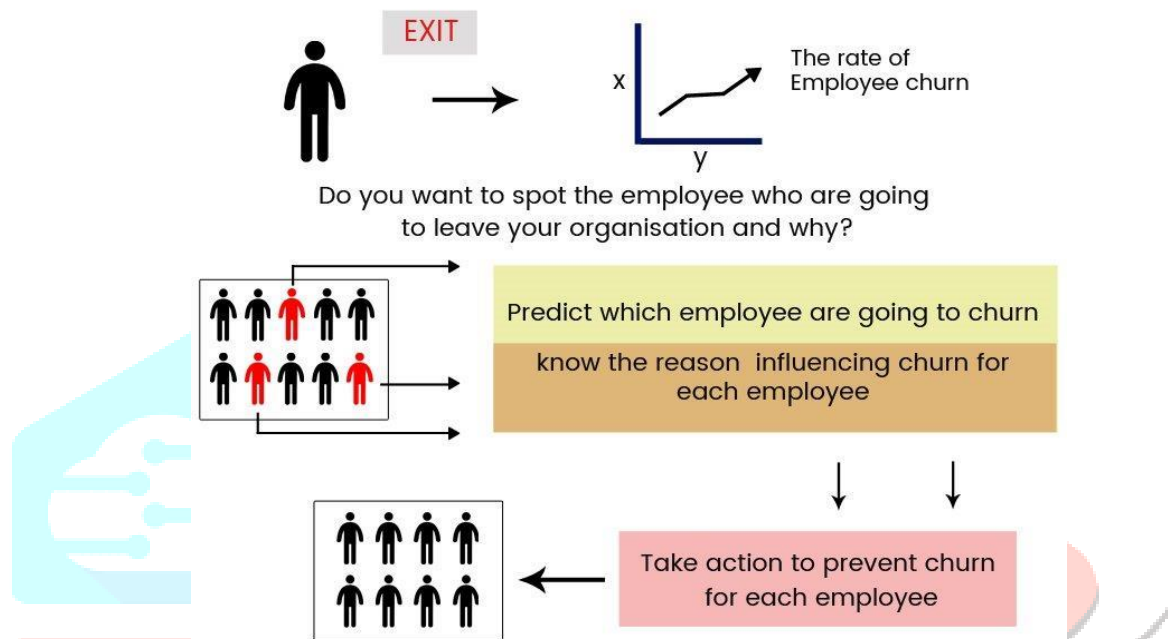
Figure 2: Employee Attrition Prediction

Algorithm steps:

1) Identify the employee dataset that comprises of current also, past workers records

2) Clean the dataset, handle the missing data and determine new features whenever required

3) Select the features among the worker data that are appropriate for the prediction of attrition

4) Try classification algorithms and report the ones most appropriate by looking at the precision, accuracy, recall, and F-measure results on the test data

5) Apply feature selection method, and select the features that are more convenient in order to predict employee attrition

6) Build classification model

7) Further the forecast of employee attrition on utilizing the model

8) Decision on the methodologies of retention

IV. RESEARCH GAPS

This section contains the analysis of the research gaps that was found, based on theory,

- **Impact of partitioning in the used dataset:**

One reason for this could be explained by the input dataset partitioning itself. If the input dataset contains 97% of people who did not quit their job and 3% who did quit their job, then could a model possibly predict that 100% did not quit their job and achieve an accuracy of 100%. This is a well known problem within the field of machine learning and is called overfitting. And this happens when the model learns to predict the training data too well. The risk of overfitting could be lowered in several ways in which some of them are with using cross-validation, train with more data or regularization.

- **Input variable impact**

Another reason could be the chosen input variables and how much they actually affected the predictions. The chosen input variables were clarity, value, efficiency, workload, community and enthusiasm, and their change over two different pulse survey occasions. These were chosen because they were deemed to affect the factor job satisfaction which plays a big role in why people choose to stay or leave their organizations according to the theory.

V. CONCLUSION

Attrition being a major element contributing to the growth of an organization, researchers are still studying ways and methods to identify any possible attrition. Majority of the researches in this survey suggest that the future dominating algorithm in this sector is most likely to be Tree based algorithm. Recent studies have demonstrated that accuracy is not the only factor while evaluating models, but also other performance measures such as recall, precision, F1 score and ROC curve must be considered. In this paper we have studied about different techniques and methods used by the various researchers for employee prediction strategy. Hence, increasing the overall perspective, this survey will contribute to design a robust and reliable model for employee attrition prediction and their respective retention strategies.

REFERENCES

[1]. Cockburn, I.; Henderson, R.; Stern, S. The Impact of Artificial Intelligence on Innovation. In The Economics of Artificial Intelligence: An Agenda; University of Chicago Press: Chicago, IL, USA, 2019; pp. 115–146.

[2]. Jarrahi, M. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Bus. Horiz. 2018, 61, 577–586.

[3] Yanqing, D.; Edwards, J.; Dwivedi, Y. Artificial intelligence for decision making in the era of Big Data. Int. J. Inf. Manag. 2019, 48, 63–71.

[4] Paschek, D.; Luminosu, C.; Dra, A. Automated business process management-in times of digital transformation using machine learning or artificial intelligence. In MATEC Web of Conferences; EDP Sciences: Les Ulis, France, 2017; Volume 121.

[5] Varian, H. Artificial Intelligence, Economics, and Industrial Organization; National Bureau of Economic Research: Cambridge, MA, USA, 2018.

[6] Vardarlier, P.; Zafer, C. Use of Artificial Intelligence as Business Strategy in Recruitment Process and Social Perspective. In Digital Business Strategies in Blockchain Ecosystems; Springer: Berlin/Heidelberg, Germany, 2019; pp. 355–373.

[7] Gupta, P.; Fernandes, S.; Manish, J. Automation in Recruitment: A New Frontier. J. Inf. Technol. Teach. Cases 2018, 8, 118–125.

[8] Geetha, R.; Bhanu Sree Reddy, D. Recruitment through artificial intelligence: A conceptual study. Int. J. Mech. Eng. Technol. 2018, 9, 63–70.

[9] Syam, N.; Sharma, A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Ind. Mark. Manag. 2018, 69, 135–146.

[10] Vahabzadeh, Ali & Yusuff, Rosnah. (2012). Greening your reverse logistics: Decision-making model can help organizations recapture value. Industrial Engineer. 44.

[11] Grothoff, Julian & Kleinert, Tobias. (2021). Mapping of Standardized State Machines to Utilize Machine Learning Models in Process Control Environments. 10.1007/978-3-030-69367-1_4.

[12] Martin, L. How to retain motivated employees in their jobs? Econ. Ind. Democr. 2018, 34, 25–41.

[13] Zelenski, J.M.; Murphy, S.A.; Jenkins, D.A. The happy-productive worker thesis revisited. J. Happiness Stud. 2008, 9, 521–537.

[14]. Vishnuprasad Nagadevara, "Prediction of employee attrition using workplace related variables", Review of Business Research, 2012, Research Gate.

[15] Zhao, Yue & Hryniewicki, Maciej & Cheng, Francesca & Fu, Boyang & Zhu, Xiaoyu. (2018). Employee Turnover Prediction with Machine Learning: A Reliable Approach. 10.1007/978-3-030-01057-7.

[16] Subhashini, M & R., Dr. Gopinath. (2020). Employee attrition prediction in industry using machine learning techniques. International journal of advanced research in engineering & technology. 11. 3329-3341. 10.34218/IJARET.11.12.2020.313.

[17] Dutta, Shawni & Bandyopadhyay, Samir. (2020). Employee attrition prediction using neural network cross validation method. International Journal of Commerce and Management. 6. 80-85.

[18] Najafi, Saeed & Shams Gharneh, Naser & Arjomandi Nezhad, Ali & Zolfani, Sarfaraz. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. 10.3390/MATH9111226.

[19] Al-Darraji, Salah & Honi, Dhafer & Fallucchi, Francesca & Ibrahim, Ayad & Giuliano, Romeo & Abdulmalik, Husam. (2021). Employee Attrition Prediction Using Deep Neural Networks. Computers. 10. 141. 10.3390/computers10110141.

[20] Usha, P.M. & Balaji, N.V. (2021). A Grey Wolf Optimization Improved Deep Belief Network for Employee Attrition Prediction. International Journal of Engineering Trends and Technology. 69. 72-81. 10.14445/22315381/IJETT-V69I10P210.

[21] Mirjalili, Seyedali & Mirjalili, Seyed & Lewis, Andrew. (2014). Grey Wolf Optimizer. Advances in Engineering Software. 69. 46–61. 10.1016/j.advengsoft.2013.12.007.

[22] Sharma S, Mehta S, Chopra N et al (2015) Int J Eng Res Appl 5(4, Part-6):128–132. ISSN 2248-9622.

[23] Mehta, Sonam & Shukla, Deepak. (2015). Optimization of C5.0 classifier using Bayesian theory. 1-6. 10.1109/IC4.2015.7375668.

[24] Ben Yahia, Nesrine & Jihen, Hlel & Colomo-Palacios, Ricardo. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3074559.

[25] Mauricio A. Valle & Gonzalo A. Ruz (2015) Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms, Applied Artificial Intelligence, 29:9, 923-942, DOI: 10.1080/08839513.2015.1082282.

[26] Gao, Ying. "using decision tree to analyze the turnover of employees." (2017).

[27] W. C. Hong, S. Y. Wei, and Y. F. Chen, "A comparative test of two employee turnover prediction models", International Journal of Management, 24(4), 808, 2007.

[28] Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. algorithms, 4(5), C5.

[29] Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999-2006.

[30] Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A., & Kulkarni, D. (2018). Employee Attrition Prediction. arXiv preprint arXiv:1806.10480.