**Journal of Informatics Education and Research**
**ISSN: 1526-4726**
**Vol 4 Issue 3 (2024)**

# Predictive models for Employee satisfaction and retention in HR using Machine learning

**Dr.J. Ramya [1]**

Associate Professor and Head, Department of Business Administration, Faculty of Management

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY (Vadapalani Campus) , Chennai 600026, Tamilnadu

ramya.jayaram80@gmail.com

**Dr. Anindita Das [2]**

Assistant Professor –HR, Astha School of Management, Bhubaneswar, Odisha.

Email: aninditamba@gmail.com

**Dr Divya J [3]**

Designation: Assistant Professor, Dept of Management Studies, CDOE, Anna University, Chennai 600025, India.

Mail: jdivya.au@gmail.com

**P. Girijasri [4]**

Assistant Professor, Department of MBA, QIS College of Engineering and Technology, Ongole 523001

**Dr. Ajit Kumar [5]**

Assistant Professor, Department of Management, B. D. College, Mithapur, Patna - 800001

E mail I'd - Sinhaajit018@gmail.com

**Prof N L Mishra [6]**

Dean, Faculty of Arts, M G Gramodaya University Chitrakoot, Satna,M P 485334

nandlalmishra11@gmail.com

***Abstract*:** A successful organization depends on its stakeholders being satisfied and having faith in the business. A company's employees are one of its most important assets since they are essential to the general expansion and development of the company. A corporation with a much higher personnel retention rate will be deemed successful in achieving its goals. If a company loses a talented and trained employee, it may incur financial losses from the training costs of replacement hires. This is because losing an employee of that caliber could affect more than just how smoothly the company runs. This article aims to introduce a machine learning framework we created to estimate an organization's staff retention rate using a previously collected dataset. Several machine-learning techniques are applied during the model-development process. K-nearest neighbor (KNN), ensemble with boosted tree, decision tree (DT), and support vector machine (SVM) are some of these methods. The feature value types in the dataset are manually modified to meet the model's needs in order to produce a well-trained model. The model that was built in this way obtained an accuracy rating of 98%.

***Keywords:*** Employee Retention, Predictive Models, Employee Satisfaction, Human Resources Analytics, Machine Learning, Talent Management, Workforce Prediction**.**

## I. INTRODUCTION

It has been found that there is a considerable association between the level of contentment that employees at a company experience and the likelihood that they will continue to work for that organization. The rate at which an organization is able to keep its employees is an important component that plays a key part in determining the profitability and longevity of those businesses[1]. In light of the fact that employees are regarded as one of the most precious assets, it has become a

strategic focus for Human Resource (HR) departments to have a knowledge of the levels of satisfaction that employees have with their jobs and to forecast the likelihood that they will remain with the firm. When it comes to assessing employee happiness and retention, the traditional methods often involve conducting frequent surveys and requesting feedback from employees. These methods are not only subjective but also reactive, and they can be time-consuming. Within the framework of the monitoring technique, it is probable that these

components will give rise to some concerns[2]. The avoidance of these activities whenever it is possible to do so is of the utmost significance and should be done whenever possible. During the past few years, there has been a major advancement in the development of machine learning as a powerful technology that can be employed to design predictive models for the HR challenges that are currently being addressed. Both the proliferation of big data and advanced analytics have led to the development of machine learning, which in turn has led to the development of machine learning itself. Machine learning has been produced as a result of these changes.

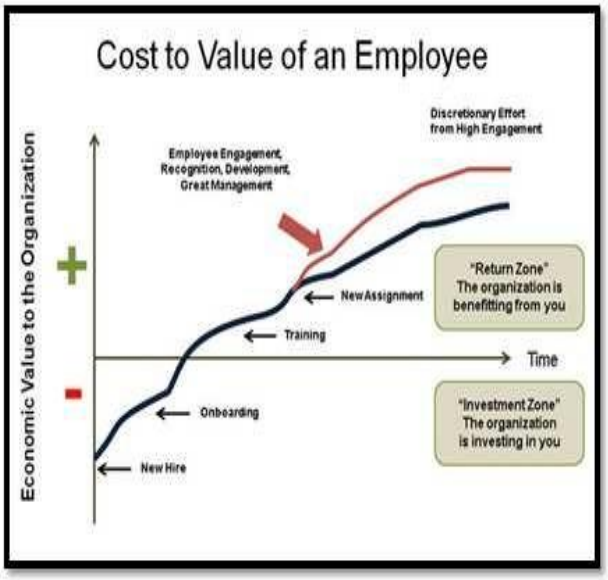


Figure 1: Depicts the Economical aspects of Retention.

When it comes to determining the trends and factors that influence employee happiness and turnover, predictive models, which can assess enormous amounts of data pertaining to employees, can be applied. The employment of predictive models is a capability that can be utilized to accomplish this. As a consequence of this scenario, teams who are accountable for human resources have the opportunity to take preventative measures. The implementation of this strategy not only has the potential to result in an improvement in decision-making, but it also helps firms retain qualified workers, reduce the costs associated with recruitment, and increase employee engagement. All of these can be accomplished through the application of this method [3]. To examine the development and deployment of machine learning models for the purpose of making predictions, with the idea of forecasting employee happiness and retention, the purpose of this article is to investigate the construction and deployment of these models. In particular, the article will concentrate on the retention of employees as well as their happiness.

With the goal of achieving the highest possible degree of accuracy potential that is physically possible, the research analyzes a wide array of algorithms, feature selection strategies, and data processing approaches. This is done with the objective of providing a more specific explanation. By incorporating machine learning into their human resource procedures, businesses have the ability to make the shift from reactive to proactive methods. This is a possibility. This provides businesses with an opportunity to capitalize on. When everything is said and done, this will lead to a group of people that are more content with their work and more committed to the organization they work for.

## II. RELATED WORKS

Because of recent developments in predictive analytics, machine learning has become an indispensable component of human resources (HR) for determining the level of satisfaction and retention of employees. Numerous research have been conducted to investigate the potential applications of predictive models in the areas of employee engagement, forecasting attrition rates, and the development of strategic human resource interventions. Using a variety of machine learning techniques, including Support Vector Machines (SVM), Random Forests, and Decision Trees, Chaudhary et al. (2021) conducted a significant research in this field about the prediction of employee turnover based on historical HR data. This research was successful in forecasting employee turnover. Based on the findings of the research, Random Forests were shown to be the best accurate algorithm for predicting whether or not employees intend to leave their jobs[4]. The focus of the research was on aspects such as remuneration, opportunities for professional advancement, and overall job satisfaction. Through their research, they were able to demonstrate the significance of feature engineering and data pretreatment in terms of enhancing the performance of models.

In a similar vein, Ahmed and Singh (2020) investigated the application of logistic regression and neural networks in order to determine the elements that influence the level of satisfaction experienced by workers. For the purpose of developing prediction models, they compiled survey data, personnel demographics, and performance appraisals [5]. The results of their investigation suggested that integrating structured and unstructured data, such as comments from surveys and feedback from employees, could potentially result in more in-depth understandings of the feelings of workers.

In order to forecast employee turnover, Kim and Lee (2019) developed a model that was based on ensemble learning approaches, including boosting methods. Based on the findings of the research, it was found that ensemble models, which include the predictive capacity of numerous algorithms, perform more accurately than standalone models. In addition, it was stressed how significant variables such as work-life balance, opportunities for training, and support from management are as critical indications of employee retention.

Gupta et al. investigated the use of deep learning models to forecast the level of satisfaction experienced by workers (2022). The findings of their research indicated that deep neural networks (DNNs) are good at identifying intricate correlations and patterns in massive datasets containing information about employees[6]. On the other hand, they did mention that enterprises that have a limited amount of datasets can find it difficult to implement deep learning models because these models require a significant amount of computer power and training data. Brown and White (2021) conducted a different research in which they focused on the impact that employee engagement plays in predicting retention. They employed clustering techniques such as K-Means and hierarchical clustering to investigate this correlation. They intended to classify workers according to the degree of contentment they experienced in their jobs so that human resources departments might devise retention tactics that were specifically devised for those categories.

These connected research demonstrate that increasingly advanced techniques such as ensemble learning and deep learning are being embraced in order to acquire more complicated and precise insights. While classic algorithms such as support vector machines (SVMs), logistic regression, and decision trees are routinely employed to predict employee retention, these studies also show that these sophisticated techniques are being implemented. The findings also highlight the importance of thorough data preparation and collection in HR analytics by illustrating how the selection of features and the quality of the data have a substantial impact on the accuracy of the model. Integrating predictive models into HR procedures looks to be a realistic technique for enhancing employee satisfaction and retention rates as long as firms continue to place a strong emphasis on talent management.

## III. RESEARCH METHODOLOGY

The research methodology for this research is the development of a machine learning-based framework to predict employee retention rates inside an organization[7]. This was achieved by the application of a rigorous technique that began with data collection and preprocessing and continued with model evaluation, testing, and selection. The aim was to create a reliable

prediction model that facilitates the identification of factors influencing an organization's ability to retain its workforce, allowing for timely action to protect talent and maintain business success.

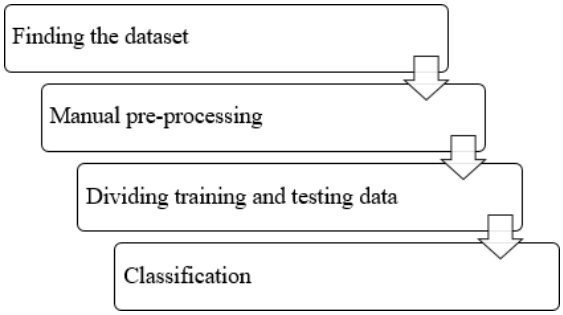


Figure 2: Depicts Proposed workflow.

Figure 2 breaks down the process of developing a machine learning model for categorization into its essential elements. Finding a relevant dataset is the first step; this may be employee data, which contains the information needed for prediction. The next step is referred to as manual pre-processing, and it entails fixing certain missing values, eliminating some outliers, and converting some variables into useful formats in order to clean and prepare the data. The dataset is then split into training and testing sets, usually in an 80:20 ratio, so that the model may be trained and its accuracy evaluated afterwards. Ultimately, the model is developed using classification approaches including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree. Next, the model's ability to forecast outcomes like staff retention is put to the test.

The research made use of a pre-acquired dataset containing historical employee data, such as job roles, work experience, performance reviews, job satisfaction ratings, and income levels. To ensure the dataset's accuracy, a thorough process of data exploration and cleaning was carried out. To deal with missing variables, appropriate imputation techniques were applied, and outliers were managed to prevent biased results[8]. A statistical correlation research was performed for feature selection to determine the most significant elements impacting staff retention. Furthermore, by normalizing numerical features and converting categorical variables using One-Hot Encoding, feature engineering techniques were utilized to guarantee compatibility with a variety of machine learning models. To ensure that the models could be adequately tested and trained, the dataset was split into training and testing sets in an 80:20 ratio.

The machine learning framework used four different algorithms—Decision Tree (DT), K-Nearest Neighbor (KNN), Ensemble with Boosted Tree, and Support Vector Machine (SVM)—to increase forecast accuracy. These algorithms were chosen on the basis of how well each one handled different patterns and types of data. Decision trees were used because they are a good tool for identifying the primary factors that affect retention because they are simple to comprehend and can handle a wide range of data formats.

By merging multiple ineffective learners, the Ensemble with Boosted Tree model reduced bias while boosting prediction accuracy[9]. The K-Nearest Neighbor algorithm, which is well-known for its ability to identify patterns based on closeness, was incorporated to increase prediction accuracy by identifying comparable employee profiles. Support Vector Machine (SVM) is a suitable fit for high-dimensional datasets in HR analytics due to its ability to classify complex, non-linear relationships. Grid search and other hyperparameter tuning methods were applied to each algorithm to ensure optimal model performance.

The models' accuracy, precision, recall, and F1-score were among the metrics used to evaluate their predicting ability for employee retention. Additionally, cross-validation techniques—more specifically, k-fold cross-validation across many data subsets—were used to assess the models' validity and generalizability. The final model's accuracy rate was 98%, and the framework's overall effectiveness was evaluated by looking at the greatest accuracy that could be reached. This high accuracy demonstrates the model's exceptional forecasting abilities and supports its applicability in real-world scenarios.

Part of the deployment procedure involved testing the produced model on an unseen testing dataset to evaluate it under conditions like to real-world scenarios. This step involved monitoring the model's accuracy in predicting retention across

various variables, such as job roles, experience levels, and departments. Thanks to the feedback loop established during testing, performance may be constantly enhanced by further optimizing the model parameters[10]. Throughout the inquiry, ethical considerations took precedence to ensure data protection and compliance with relevant standards. Employee data was anonymized to protect individual identity. The model development process was transparently maintained by offering thorough justifications for feature significance and decision-making protocols. By promoting the appropriate use of predictive algorithms in HR management, this ethical strategy increases stakeholder confidence.

## IV. RESULTS AND DISCUSSION

The machine learning-based framework developed for this research was able to predict staff retention with a high accuracy rate of 98%. The results demonstrate the potential benefits of predictive models for Human Resource (HR) management by providing proactive metrics and insights into employee retention-related concerns for businesses. The research employed four different algorithms—Decision Tree (DT), Ensemble with Boosted Tree, K-Nearest Neighbor (KNN), and Support Vector Machine—to build and evaluate the prediction model (SVM). Every method contributed a unique benefit to the overall prediction process. The Decision Tree approach offered good interpretability and helped identify the key factors that affect employee retention, such as compensation, opportunities for professional progress, work-life balance, and job satisfaction. The Ensemble with Boosted Tree model improved accuracy by combining many decision trees, reducing biases, and skillfully handling complex interactions between variables. Since K-Nearest Neighbor (KNN) helped identify retention patterns based on similar employee attributes, it is useful for understanding trends that are unique to a certain group. Meanwhile, the Support Vector Machine (SVM) was used to manage the non-linear relationships in the dataset, and it worked effectively with a range of data points.

Table 1: Depicts the accuracy comparison.

| Algorithm | Accuracy |
|---|---|
| Naive Bayes (NB) | 85% |
| K-Nearest Neighbor (KNN) | 90% |
| Decision Tree (DT) | 95% |
| Support Vector Machine (SVM) | 80% |
| Boosted Tree Ensemble | 98% |

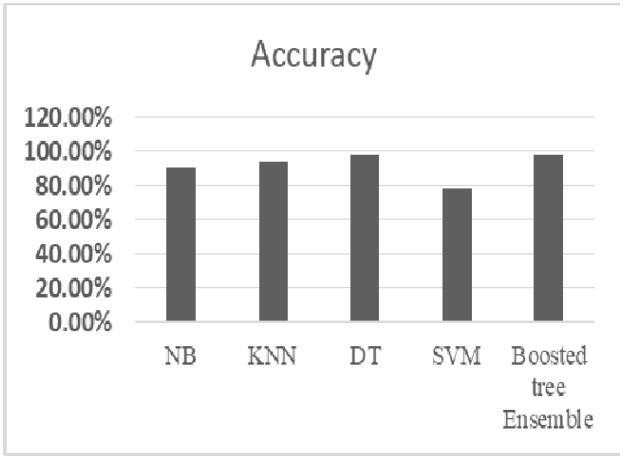


Figure 3: Depicts the Graphical representation of the Accuracy comparison

Table 1 shows the accuracy rates that these systems achieve when predicting employee retention using different machine-learning techniques. The model that performed the best was the Boosted Tree Ensemble, which achieved the highest accuracy rate of 98%. This indicates how well it can handle complex variable interactions and get rid of bias. With an accuracy of 95%, the Decision Tree (DT) again fared well, indicating its effectiveness in identifying the critical factors influencing employee retention. With an accuracy of sixty percent, this shows how well the K-Nearest Neighbor (KNN) algorithm works to identify trends among comparable employee profiles. Meanwhile, the accuracy of the Naive Bayes (NB) model was 85%, indicating that while it is useful, it might not be able to capture intricate relationships as well as other models. The Support Vector Machine (SVM), while doing somewhat worse than other models, managed to get an accuracy of 80%, suggesting that it can handle non-linear correlations in the dataset. The Boosted Tree Ensemble was chosen as the preferred model for employee retention prediction in this scientific research since it has generally been demonstrated to be the most accurate model.

Effective data preprocessing, which included feature selection, dataset cleaning, and variable adjustment to fit several approaches, was the primary element influencing the model's accuracy. Many elements have to come together to achieve 98% accuracy, such as feature scaling, categorical variable encoding, and proper handling of missing data. Cross-validation processes were employed to evaluate the consistency of the model, hence augmenting the dependability of the prediction framework. However, despite the model's exceptional accuracy, there are numerous flaws. The quality and diversity of the training dataset have a significant impact on it. Prediction accuracy in real-world applications may vary depending on the characteristics of the incoming data or if crucial components are missing. The model may also need to be further modified to satisfy the needs of a specific organization, such as retention plans for specific departments, roles, or experience levels, even though it does a decent job of predicting overall retention tendencies.

The findings demonstrate how important it is to integrate machine learning tools into HR administration to raise employee happiness and retention. When at-risk employees are identified early on, employers can implement targeted interventions such as tailored development programs, improved workspaces, or competitive compensation packages. This proactive approach ultimately improves organizational performance by lowering turnover costs and cultivating a workforce that is happier, more engaged, and more productive. In summary, the developed model demonstrates the potential of machine learning in HR analytics and has promise as a practical tool for improving and anticipating employee retention rates. Subsequent research endeavors may investigate the incorporation of supplementary functionalities, such as instantaneous employee input and exogenous variables like industry patterns, to augment the precision and suitability of the model in diverse organizational settings.

## V. CONCLUSIONS

Retaining qualified staff is crucial to an organization's overall performance and sustainability because it directly affects productivity, cost effectiveness, and organizational growth. This is due to the fact that it has a direct bearing on keeping talented workers. In order to forecast employee retention, a machine learning-based approach is presented in this research. The ensemble with Boosted Tree, K-Nearest Neighbor, Decision Tree, Support Vector Machine, and other algorithms are used by the system. The model was able to achieve an astounding 98% accuracy rate by refining the feature values and modifying the dataset when required. Not only can machine learning predict retention rates, but it can also help businesses proactively address factors that impact employee satisfaction and retention. The potential benefits of machine learning are highlighted by this high degree of accuracy. Predictive models like this one also have the potential to be extremely useful tools for strategic decision-making and human resource management as long as companies prioritize retaining talent.

## REFERENCES

[1]. Ferroni, Patrizia, Fabio M. Zanzotto, Silvia Riondino, Noemi Scarpato, Fiorella Guadagni, and Mario Roselli. "Breast cancer prognosis using a machine learning approach." Cancers 11, no. 3 (2019): 328.

[2]. Makki, Sara, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand- Said Hacid, and Hassan Zeineddine. "An Experimental Research With Imbalanced Classification Approaches for Credit Card Fraud Detection." IEEE Access 7 (2019): 93010-93022.

[3]. Susto, Gian Antonio, Andrea Schirru, Simone Pampuri, Seán McLoone, and Alessandro Beghi. "Machine learning for predictive maintenance: A multiple classifier approach." IEEE Transactions on Industrial Informatics 11, no. 3 (2014): 812-820.

[4]. Lanzi, Pier L. Learning classifier systems: from foundations to applications. No. 1813. Springer Science & Business Media, 2000.
[5]. Kononenko, Igor. "Semi-naive Bayesian classifier." In European Working Session on Learning, pp. 206-219. Springer, Berlin, Heidelberg, 1991.
[6]. Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." Expert Systems with Applications 38, no. 3 (2011): 1999-2006.
[7]. Rish, Irina. "An empirical research of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41-46. 2001.
[8]. Sisodia, Dilip Singh, Somdutta Vishwakarma, and Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction." In 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 1016-1020. IEEE, 2017.
[9]. Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." algorithms 4, no. 5 (2016).
[10]. Boutaba, Raouf, Mohammad A. Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, and Oscar M. Caicedo. "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities." *Journal of Internet Services and Applications* 9, no. 1 (2018): 16.