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

DETECTING EMPLOYE ATTRITION IN AN ORGANIZATION USING DECISION TREE, LOGISTIC REGRESSION AND RANDOM FOREST

[1]N.V. Murali Krishna Raja, [2]L.N.V..Radhika, [3]P.E.N. Venkateswara Rao,[4]G. Mahesh,[5]P.L.S. Sandhya, [6]SK.Yaseen

[1]Sr.Asst. Professor, [2]Student, [3] Student ,[4] Student ,[5] Student ,[6]Student
[1]Computer Science and Engineering,
[1]SRI VASAVI ENGINEERING COLLEGE, TADEPALLIGUDEM, INDIA.

Abstract: **Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees. Challenges that an organization has to face due to employee attrition are Expensive in terms of both money and time to train new employees. Loss of experienced employees' impact in productivity, and also impact in profit. Whenever an employee leaves an organization, there is a source of advantages for the business competitor. To be continuously competitive in the business, the organization should minimize the employee attrition. The major aim of the project is to analyze the employee data set of an organization and find the reasons, why the best and most experienced employees leave the company prematurely and also try to predict which valuable employees are probable to leave the organization subsequently.**

INTRODUCTION

Employee attrition, the phenomenon of employees leaving an organization, is a critical concern for businesses across various industries. The cost of hiring and training new employees, coupled with the potential loss of valuable skills and knowledge, makes it imperative for organizations to proactively identify factors leading to attrition. Predictive modeling using machine learning techniques has emerged as a powerful tool in this context, enabling HR professionals and organizational leaders to anticipate and mitigate employee attrition.

This project aims to explore and compare three popular machine learning algorithms—Decision Tree, Random Forest, and Logistic Regression—for the task of predicting employee attrition. Each of these models offers unique strengths and insights into the complex dynamics that contribute to employee turnover. By leveraging historical data on employee demographics, job-related factors, and workplace satisfaction, we can build predictive models that assist organizations in making informed decisions to retain their valuable talent.

By combining the strengths of Decision Trees, Logistic Regression, and Random Forest, this project aims to provide organizations with accurate and actionable insights to mitigate employee attrition and foster a healthier and more sustainable work environment.

The foundation of any predictive modeling project is the dataset. We will collect and preprocess relevant data, which may include employee demographics, job satisfaction scores, performance metrics, and other pertinent variables. The dataset will be divided into training and testing sets to evaluate model performance.

The final models will be implemented in a practical setting, allowing organizations to integrate them into their employee management systems. Real-time monitoring and periodic updates will ensure the continued relevance and effectiveness of the models.

2.PROBLEM STATEMENT

High employee turnover refers to the frequent departure of employees from a company, which poses significant challenges and obstacles for organizations. This issue has far-reaching consequences that impact various aspects of a company's functioning, including productivity, employee morale, and overall operational effectiveness.

Firstly, high employee turnover negatively affects productivity within a company. When employees leave, it disrupts the workflow and creates gaps in knowledge and skills. New employees need time to get up to speed, learn the job, and become fully productive. This transition period can result in decreased output and efficiency, as the company must invest time and resources in training and onboarding new hires. Additionally, the constant turnover can lead to a loss of institutional knowledge, as experienced employees leave and take their expertise with them. This loss of knowledge can hinder productivity and require additional time and effort to regain.

Secondly, high employee turnover has a detrimental impact on employee morale. Frequent departures can create a sense of instability and uncertainty among the remaining employees. They may feel demotivated, as they witness their colleagues leaving and question their own job security. This can lead to decreased job satisfaction, lower engagement, and a decline in overall morale. A negative work environment can further contribute to increased turnover, creating a vicious cycle that is difficult to break.

Moreover, high employee turnover affects the overall operational effectiveness of a company. Constantly recruiting, hiring, and training new employees is a time-consuming and costly process. It requires significant resources, including advertising job openings, conducting interviews, and providing training. These resources could have been allocated to other areas of the business, such as innovation, expansion, or improving existing processes.

3 PROPOSED SYSTEM

This project aims to address the pressing concern of employee turnover in our company by suggesting the creation and utilization of predictive models that leverage three advanced machine learning algorithms: decision trees, random forests Employee turnover is a significant issue that can have detrimental effects on a company's productivity, morale, and overall success. By implementing predictive models, we can proactively identify employees who are at a higher risk of leaving the company, allowing us to take preventive measures to retain them.

The first algorithm, decision trees, will analyze various factors such as job satisfaction, salary, work-life balance, and career growth opportunities to identify patterns and create a predictive model.

The second algorithm, random forests, will build upon the decision tree model by creating an ensemble of multiple decision trees. This ensemble approach will enhance the accuracy and robustness of the predictive model, as it will consider different perspectives and reduce the risk of overfitting.

3.1 EXISTING SYSTEM

As a result, many organizations are turning to more advanced technologies such as predictive analytics and machine learning to better understand and address employee attrition. These tools can analyze large amounts of data in real-time, allowing companies to identify patterns and trends that may be contributing to attrition rates. By leveraging these technologies, organizations can make more informed decisions and take proactive measures to retain their top talent.

Additionally, some companies are implementing employee engagement platforms that use artificial intelligence to track employee sentiment and satisfaction levels. By monitoring factors such as workload, communication, and job satisfaction, these platforms can provide valuable insights into potential attrition risks and help managers address issues before they escalate.

3.2 SYSTEM DESIGN

The system's architecture is designed to be modular, providing flexibility and scalability. This architecture serves as a solid foundation for developing an efficient employee attrition prediction system. By combining machine learning models with user-friendly interfaces for HR professionals, the system becomes more effective. Additional refinement and detailed specifications can be incorporated based on the specific requirements and constraints of the project.

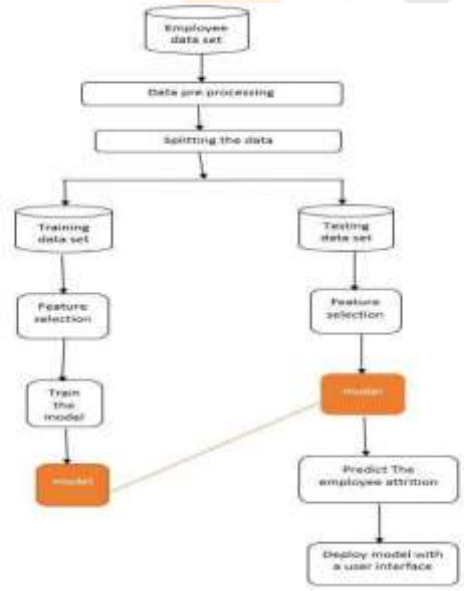

Fig : System design.

4.ALGORITHMS

4.1 Decision Tree:

Decision trees are powerful tools that provide a visual representation of the various choices and potential outcomes available to a business in a given situation. They enable decision-makers to analyze complex scenarios and make informed decisions by systematically evaluating the potential consequences of each choice.

By mapping out the different decision paths and their associated outcomes, decision trees allow decision-makers to identify the optimal course of action that maximizes desired outcomes or minimizes risks. This visual depiction helps in understanding the potential consequences of each decision and enables decision-makers to consider multiple factors simultaneously.

Decision trees empower decision-makers by providing a structured framework for evaluating choices and their potential outcomes. They allow for a systematic analysis of the various factors that may influence a decision, such as market conditions, customer preferences, resource availability, and competitive landscape.

4.2 Random Forest:

Random forest (RF) is a powerful and widely used supervised machine learning algorithm that excels in producing accurate classifications and regressions. It achieves this by harnessing the collective strength of multiple decision trees to train on a given dataset.

In RF, each decision tree is constructed using a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and increase the diversity among the trees. Each tree independently makes predictions based on its own set of rules and features.

When it comes to classification tasks, RF assigns a class label to a specific dataset based on the majority vote of the decision trees. The class label that receives the most votes becomes the final classification decision made by the RF model.

4.3 Logistic Regression:

Logistic regression is a statistical model that is widely used in machine learning and data analysis. It is a type of supervised learning algorithm that is utilized for classification tasks, specifically to forecast the likelihood of a target variable. The target variable in this case is binary, indicating that there are only two potential classes.

The main objective of logistic regression is to find the best fitting line or curve that separates the two classes in the data. It does this by estimating the probability of the target variable belonging to a particular class based on the input features. The output of logistic regression is a probability score between 0 and 1, which can be interpreted as the likelihood of the target variable belonging to one of the classes.

The logistic regression algorithm uses a mathematical function called the logistic function or sigmoid function to model the relationship between the input features and the target variable. This function maps any real-valued number to a value between 0 and 1, which is useful for representing probabilities the target variable belonging to one of the classes.

5.RESEARCH METHODOLOGY

5.1 Requirements Gathering and Analysis:

The data set contains different features such as Satisfaction level Rating, number of projects, average monthly hours, time spent in company, work problems, last evaluation rating, promotion in last five years, last evaluation Rating, department, salary, left or not. analyze the data and remove the missing values from the data.

5.2 Design and Planning:

This means that in supervised learning, the training set consists of data points that have both input variables and corresponding output variables. The input variables are used as the features or attributes that the algorithm analyzes, while the output variable is the target or label that the algorithm tries to predict or classify. The algorithm uses these data points to learn the relationship between the input variables and the output variable. It analyzes the patterns and correlations in the training data to create a model or function that can make predictions or classifications on new, unseen data.

5.3 Accuracy:

When assessing categorization models—including those intended to identify staff attrition—accuracy is a frequently employed statistic. By comparing the number of accurately predicted instances to the total number of occurrences, it calculates the overall correctness of predictions. Other measures that provide further information about the model's performance in these situations include precision, recall, and F1-score.

Accuracy can be measured by:

Accuracy = $\boldsymbol{TP + TN \div (TP + TN + FP + FN)}$

5.4 Deployment

We determine the optimal approach to accurately identify the project's outcome by applying several machine learning algorithms during the data's testing and training procedures.

6. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
Satisfaction_level	0.090	1.000	0.6128	0.24631
Last_evaluation	0.360	1.0000	0.7161	0.171
Number_of_projects.	2.0000	7.0000	3.802	1.2325
Average_monthly_hours.	96.00000	310.000	201.050	49.943
Years_of_company.	2.0000	10.0000	3.498	1.4601

Table 4.1 displayed mean, standard deviation, maximum, minimum values of the employes data set

Accuracy table

ALGORITHMS	ACCURACY
LOGISTIC REGRESSION	**77.81**
DECISSION TREE	**97.1**
RANDOM FOREST	**98.0**

REFERENCES

[1] J. Xu-rui, W. Ming-gong, W. Xiang-xi and W. Ze-kun, "Application of ensemble learning algorithm in aircraft probabilistic conflict detection of free flight," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, 2018, pp. 10-14.

[2] D. Salmond, "Blind estimation of wireless network topology and throughput," 2019 53rd Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2019, pp. 1-6.

[3] F. M. Rueda, S. Lüdtke, M. Schröder, K. Yordanova, T. Kirste and G.A. Fink, "Combining Symbolic Reasoning and Deep Learning for Human Activity Recognition," 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kyoto, Japan, 2019, pp. 22-27.

[4] F. Sasso, A. Colucci and G. Notar Stefano, "Interaction-Based Distributed Learning in Cyber-Physical and Social Networks," in IEEE Transactions on Automatic Control, 2019.

[5] S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[6] D. G. Gardner, L. V. Dyne and J. L. Pierce, "The effects of pay level on organization-based self-esteem and performance: a field study," Journal of Occupational and Organizational Psychology,
vol. 77, no. 3, pp. 307-322, 2004.

[7] E. Moncarz, J. Zhao and C. Kay, "An exploratory study of US lodging properties' organizational practices on employee turnover and retention," International Journal of Contemporary Hospitality Management, vol. 21, no. 4, pp. 437-458, 2009.

[8] Q. A. Al-Radaideh and E. A. Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," international Journal of Advanced Computer Science and Applications, vol. 3, no. 2, p. 144–151, 2012.