International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

# COPY RIGHT



IJIEMR Transactions, online available on 10th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

**10.48047/IJIEMR/V12/ISSUE 04/106**

Title **PREDICTION OF EMPLOYEE ATTRITION USING MACHINE LEARNING**

Volume 12, ISSUE 04, Pages: 844-850

Paper Authors

**Dr. B. Sai Jyothi, Ch. Nityanandakari, B. Divya, D. Mary Kumari, G. Bhavitha**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# Prediction of Employee Attrition using Machine Learning

**Dr. B. Sai Jyothi**, Professor, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

**Ch. Nityanandakari**, **B. Divya**, **D. Mary Kumari**, **G. Bhavitha**
UG Students, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
chnityanandakari@gmail.com, bandarudivyachowdary@gmail.com,
marydasari1228@gmail.com, gujjulabhavitha@gmail.com

**Abstract**

An organization's success heavily relies on its employees, who are considered its foundation. The departure of skilled, capable, and knowledgeable personnel for better prospects can have serious implications for businesses. By uncovering the sources of employee dissatisfaction, organizations can take necessary measures to reduce turnover rates. The status of employee attrition i.e whether the employee departs the organisation, can be determined by using an employee attrition prediction system. This system can be beneficial for HRM teams, who can use it to plan for future employee hiring or to devise strategies to reduce the attrition rate. The prediction is made using machine learning techniques, which take into account various factors such as the work environment, career satisfaction, employee behavior, working hours, and most importantly, income or incentives.

**Keywords:** Employee attrition, Logistic Regression, Decision tree, KNN, SVM, Random Forest

## 1. Introduction

Due to intense industry competition today, timely delivery of any service or product is the main objective of every organization. If a skilled employee unexpectedly leaves the company, the task at hand may not be completed within the specified deadline. Retaining skilled employees is a significant challenge for organizations today, as losing valuable personnel results in considerable losses, while recruiting new employees necessitates both time and effort. Thus, by creating a workplace of satisfaction, organizations can improve employee satisfaction and reduce retention. There are a lot of factors that drive people out of the company. Some of the reasons could be better compensation, interpersonal relationships with colleagues and superiors, dissatisfaction with the current position, lack of professional advancement, extended working hours, excessive workload, and so on. To tackle this issue this paper proposes a system that considers employee feedback to predict attrition and takes necessary actions, such as hiring new employees or providing opportunities for career growth. Therefore, the prediction model for determining employee attrition is of utmost importance.

This paper opted to create a system that predicts employee attrition using the IBM HR Analytics Employee Attrition dataset available on Kaggle. For the purpose of predicting employee attrition various supervised machine learning algorithms are utilized.

## 2. Literature Survey

Employee attrition is a term used to describe the loss of employees, which can be categorized as either voluntary or involuntary. Voluntary attrition occurs when an employee leaves a company of their own accord, while involuntary attrition occurs when an employer terminates an employee for various reasons. This study focuses on voluntary attrition and identifies job satisfaction, age, and compensation as the main predictors[1].

To forecast employee performance, Qasem A, A. Radaideh, and Eman A Nagi developed a classification model using data mining techniques [2]. They utilized the CRISP-DM approach and the decision tree classification technique to construct multiple classification rules [3]. Through experiments and real data from various companies, they validated the model's effectiveness, which can be used to predict the performance of new candidates.

In their study, Amir Mohammad EsmaieelISikaroudi, RouzbehGhous, and Ali EsmaieeliSikaroudi et al. utilized real data from a manufacturing facility to perform knowledge discovery methods [4]. They analyzed a range of worker characteristics such as age, technical aptitude, and work history, and assessed the relevance of these attributes using the Pearson Chi-Square test.

John M. Kirimi, Christopher Moturi, et al. developed a prediction model for employee performance forecasting to aid human resource professionals in improving their performance evaluation process [5].

In another study, Rohit Punnoose and Pankaj Ajit et al. examined the use of the Extreme Gradient Boosting (XGBoost) method, which demonstrated more reliability than other commonly used supervised classifiers in predicting employee turnover [6].

Implementing machine learning algorithms to predict the likelihood of employee turnover and taking preventative measures can help organizations avoid employee attrition.

## 3. Problem Identification

I. Which employees are most likely to leave the company?
II. Which elements contribute to employee attrition?

Managing employee attrition is crucial for achieving low and healthy turnover rates and maintaining organizational performance (Shaw, 2010). To add value to the company, HR departments have turned to data-driven decisions and machine learning (Tomassen, 2016). The perception of high employee attrition rates as a problem for firms places added pressure on HR departments to maintain attrition rates at a manageable level (Park and Shaw, 2013).

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

## 4. Methodology:

Developing a system for predicting employee attrition involves the following stages.

### 4.1 Dataset Collection

In this paper, IBM HR Analytics Employee Attrition and Performance dataset is utilized, which is accessible on the Kaggle website [7]. IBM data scientists designed this fabricated dataset for the purpose of data analysis. It comprises 35 distinct attributes with a total of 1470 entries. This dataset provides the following attributes.

Age, Attrition, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Num Companies Worked, Over18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Curr Manager.

### 4.2 Data Analysis and Visualization

To gain insight into the attributes, bar charts can be used to visualize the data and to determine how the attributes differ concerning attrition. Additionally, correlations can be examined among the attributes to analyse the strength of the relationship between two attributes.

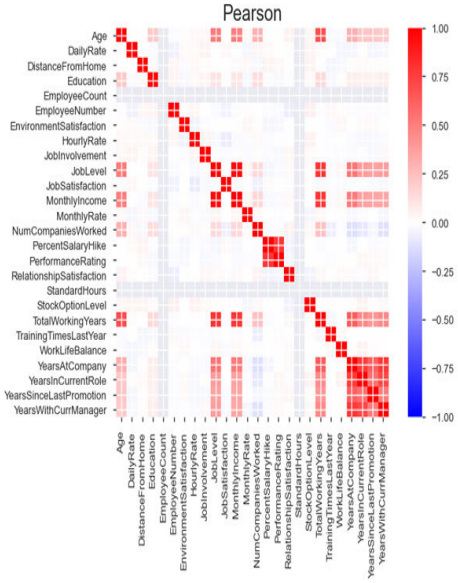


**Fig. 1. Correlation between attributes**

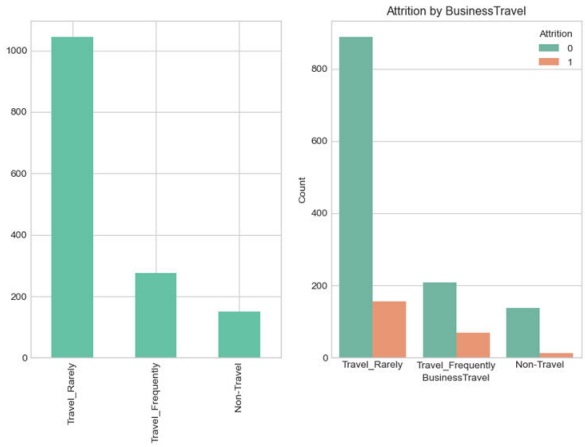


**Fig. 2. Attrition vs Business Travel**

In Figure 2, a bar graph is presented to show a comparison of the attrition rates among different business travel categories. The attrition rate for employees who travel rarely is 14.96%, while it is 24.91% for those who travel frequently, and only 8.00% for non-travelers. The higher attrition rate for employees who travel frequently can be attributed to the fact that there are 277

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

employees in that category, and 69 of them are leaving the organization. On the other hand, for employees who travel rarely, there are 1043 employees, and only 156 of them are leaving, resulting in a lower attrition rate. For non-travelers, there are a total of 150 employees, out of which only 12 are leaving the organization.

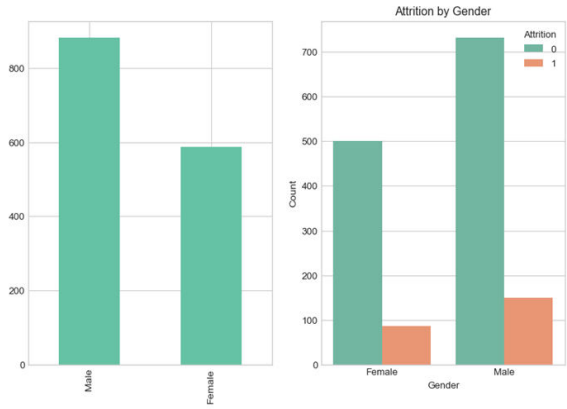


**Fig. 3. Attrition vs Gender**

The bar graph in Figure 3 depicts the attrition rates of male and female employees based on gender. Out of a total of 882 male employees, 150 are leaving, which leads to an attrition rate of 17.01%. Similarly, out of a total of 588 female employees, 87 are leaving, resulting in an attrition rate of 14.80%.

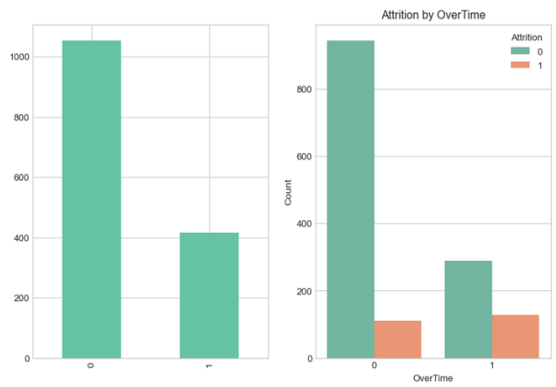


**Fig. 4. Attrition vs OverTime**

The bar graph shown in Figure 4 illustrates the relationship between overtime and attrition rates. Out of a total of 416 employees who work overtime, 127 employees are leaving the organization, resulting in a relatively high attrition rate of 30.53%. On the other hand, out of 1054 employees who do not work overtime, only 110 employees are leaving, leading to a lower attrition rate of 10.44%.

### 4.3 Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for use in machine learning models. One aspect of pre-processing is handling missing values, but in our dataset, it is observed that there are no such values. However, some attributes such as EmployeeCount, Over18, and StandardHours have constant values across all 1470 entries and are unnecessary for our prediction. Therefore, these attributes are removed from the dataset as they could potentially lead to false outcomes. The categorical variables are converted into numerical values using one-hot encoding.

### 4.4 Machine Learning Model

Machine learning employs a variety of strategies to categorise new or previously unclassified data into the appropriate class. The machine learning algorithms utilised by the system are listed below:

#### 4.4.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm that is commonly used to predict categorical

dependent variables [8]. The algorithm involves calculating the weighted sum of the input features in the dataset such as age, salary, department, BusinessTravel, gender etc and bias term, followed by passing the result through a sigmoid function to obtain the predicted probability of the positive class. The weights are then adjusted iteratively through the process of gradient descent, with the aim of minimizing the logistic loss function. The optimization process continues until convergence, which is achieved when the change in the loss function between iterations is below a certain threshold. Once the optimization is complete, the model can be used to make predictions of attrition status on the new data of an employee.

#### 4.4.2 K-Nearest Neighbour

K-Nearest Neighbour is a simple supervised machine learning algorithm that operates on the basis of similarity between new data and available data. The algorithm classifies new data into the category that most closely matches the available categories [9]. This algorithm takes the number of neighbours and the distance metric such as Euclidean or Manhattan as the input parameters and the model is trained on the training data. During the training process the model stores the input data and the corresponding class labels in memory. While predicting the data the model initially finds the optimal number of neighbours for the new data of an employee and assigns the attrition value of the new data to yes or no.

#### 4.4.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning method that can be utilized for both Classification and Regression problems [10]. The SVC function is used to create a support vector classifier model with the specified kernel function and other hyperparameters. The kernel function is specified as a parameter to the SVC class and is used to transform the data into a higher dimensional space, where it can be more easily separated. The most commonly used kernel functions are linear, polynomial, and radial basis function (RBF). The SVM algorithm seeks to find the hyperplane that maximizes the margin between the classes. The margin is the distance between the hyperplane and the closest data points from each class. The SVC function outputs a trained SVM model that can be used to make predictions of the attrition status on the new data of an employee.

#### 4.4.4 Decision Tree

Decision trees are a type of tree structure that sorts instances by evaluating the values of their features. Its primary objective is to construct a model that can forecast the value of a target variable by utilizing a set of input variables such as age, salary, department, gender of an employee. The process of creating a tree involves splitting the source employee dataset into subgroups based on an

attribute value, which is commonly referred to as training [11]. This process of dividing and subdividing subsets is called recursive partitioning. The recursive partitioning continues until the split no longer improves predictions of the attrition status thereby indicating the end of the recursion.

#### 4.4.5 Random Forest

Random Forest is a supervised machine learning algorithm that can be used to address classification and regression problems in ML. The technique employs ensemble learning, which involves combining several classifiers to solve complex problems and improve model performance. Unlike single decision trees, Random Forest makes predictions based on multiple decision trees, and the final outcome of attrition status of an employee is determined through majority voting [12]. In random forest each decision tree is created by using a random subset of the employee attrition dataset. Each tree is trained separately for making prediction of the attrition status. The accuracy of the model increases with the number of trees, which helps to prevent the problem of overfitting.

## 5. Implementation

The available dataset has been bifurcated into two segments, where one part serves as the training data and the other as the test data. The training data constitutes 70% of the total data, while the remaining 30% is reserved for testing purposes. During the training phase, the primary objective is to achieve the closest possible output to the expected response. On the other hand, the test data helps evaluate the ability of a computer to predict future outcomes and validates the performance of a machine learning model. Subsequently, several machine learning algorithms are employed to train the model, and the accuracy scores obtained are as follows:

| S. No | Algorithm | Accuracy Score | Accuracy CV 10-fold |
|---|---|---|---|
| 1 | Logistic Regression | 90.48 | 88.92 |
| 2 | K Nearest Neighbour | 89.21 | 82.60 |
| 3 | Support Vector Machine | 88.92 | 86.10 |
| 4 | Decision Tree | 100.0 | 78.52 |
| 5 | Random Forest | 100.0 | 85.52 |

**Table 1. Accuracy of algorithms**

## 6. Results

To assess the precision of a specific algorithm, the anticipated values are compared with the test values in the created model. The given table demonstrates the accuracy scores and the cross-validation 10-fold accuracy scores of each algorithm, which enables us to swiftly identify the appropriate algorithm for our model. Based on the table, it can be deduced that the Random Forest and Logistic Regression algorithms exhibit the highest accuracy on the HR-Employee-Attrition dataset.

## 7. Conclusion

The focus of this study is on the importance of anticipating voluntary employee turnover. The study proposes several classification strategies based on

supervised learning to predict this issue. With the help of this research, organizations can identify the factors that lead to employee attrition and take appropriate measures to reduce it.

## 8. Limitations and Future Scope

This study focuses on predicting employee attrition using a dataset containing only 35 attributes. It suggests that using a dataset with a larger number of attributes could result in improved outcomes. In the future, more advanced machine learning techniques such as XGBoost and Graphical Neural Networks can be employed for more precise prediction [13].

## References

[1] Dilip Singh Sisodia, Somdutta Vishwakarma,Abinash Pujahari, "Evaluation of machine learning models for employee churn prediction, "International Conference on Inventive Computing and Informatics (ICICI 2017).

[2] Mrs. Jaya Sharma, "Employee Attrition and Retention in a Cut-Throat Competitive Environment in India: A Holistic Approach," Paripex - Indian Journal of Research (PIJR), Volume 4, Issue 2, Feb 2015.

[3] Dr. B. Latha Lavanya, "A Study on Employee Attrition: Inevitable yet Manageable," International Journal of Business and Management Invention, Volume 6, Issue 9, September. 2017, pp. 38-50.

[4] Heng Zhang , Lexi Xu , Xinzhou Cheng , Kun Chao , Xueqing Zhao, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning," The 18th International Symposium on Communications and Information Technologies (ISCIT 2018).

[5] Diwakar Singh, "A Literature Review on EmployeeRetention with Focus on Recent Trends,"International Journal of Scientific Research in Science and Technology (IJSRST 2019), Volume 6, Issue 1, pp. 425-431.

[6] L. E. Peterson (2009), "K-nearest neighbor," Support Vector Machines (SVM). [Online]. Modern College of Engineering.

[7] Kaggle, "HR-Employee-Attrition." [Online]: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[8] Logistic Regression Algorithm [Online]: https://www.javatpoint.com/logistic-regression-in-machine-learning.

[9] K-Nearest Neighbor(KNN) Algorithm for Machine Learning [Online]: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[10] Support Vector Machine Algorithm [Online]: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[11] Decision Tree [Online]: https://www.geeksforgeeks.org/decision-tree/.

[12] Random Forest Algorithm [Online]: https://www.javatpoint.com/machine-learning-random-forest-algorithm.

[13] Guerranti, F.; Dimitri, G.M. A Comparison of Machine Learning Approaches for Predicting Employee Attrition. *Appl. Sci.* **2023**, *13*, 267.