# A Machine Learning Model for Predicting Employees Retention: An Initiative towards HR through Machine

2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC) | 978-1-6654-5401-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/PDGC56933.2022.10053249

Bobbinpreet Kaur
ECE department
Chandigarh University
India
bobbinece@gmail.com

Ayush Dogra
Researcher
Ronin Institute
Mont Clair,NJ,US

***Abstract*— The success of an organization depends upon the satisfaction and trust of its stake holders. Employees being one of the most important asset of an organization plays an integral role for uplifment of an organization. An organization with higher retention ration will be a successful one in terms of achieving its goals. As losing a Skilled and trained employee can not only disrupt the functionality of an organization but on the other hand incur financial losses with respect to training costs of new employees. In this article we built a Machine learning based framework of Prediction of retention rate of employees for an organized using a pre-acquired dataset. The machine learning algorithms used for development of model are Decision tree (DT), Ensemble with boosted tree, K-nearest neighbor (KNN), Support Vector machine (SVM). The feature value types in the dataset are manually modified as per the requirement of the model to achieve a well-trained model. The model thus developed provided the accuracy rate of 98%.***

**Keywords—Machine learning, Classification, HR through machine, Employee retention, Prediction**

## I. Introduction

Machine learning is an evolutionary technology that learns by experience and from the learning thus acquired the system can predict the future trends. The machine learning applications are moving at a high phase touching almost all kind of applications [1]. Every industry now a days is trying to inculcate a model based on machine learning in various departments to utilize its efficiency and problem solving capability. One such department of the organization is Human resource department wherein the application of machine learning is not that widespread till date due to a human developed notion that Human resources is all about managing Human resources through trained professionals in the field [2]. The Human resource department basically works towards Recruitment, Compensation and benefits to the employees and defining the work goals of an employee [3]. One of the biggest challenge faced by HR is employee retention. The retention of an employee is an important aspect and a highly thoughtful strategy need to be developed and proposed for increasing the retention rates as depicted in Fig.1. During restructuring of the policies, HR department is responsible for planning, developing and implementing strategies for retention of employees.

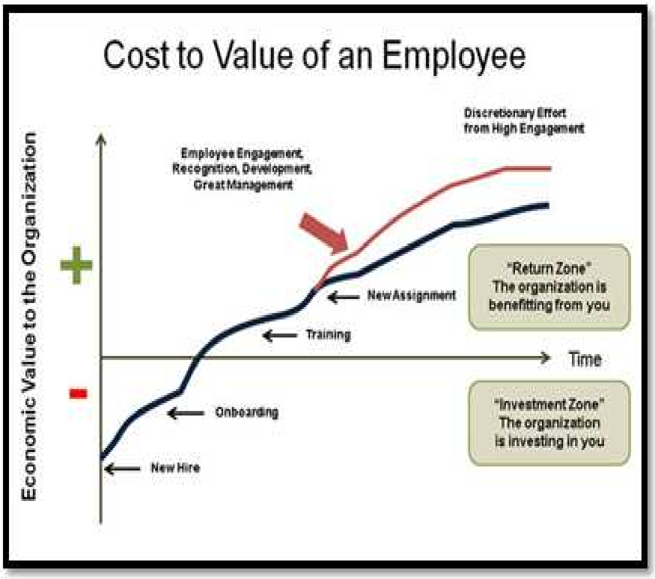

*Fig. 1 Economical aspects of Retention [4]*

Despite the fact that reorganizations resulting from falling profits are unlikely to provide wage hike, such reorganizations could provide other benefits or incentives to their employees, such as additional time off work, flexibility in the work schedule or opportunities on-site. With the increasing industrialization and opportunities for working as freelancer the retention rate is declining in the recent times thereby posing a challenge on Human resource department [5]. Therefore a trained manger is required to uplift this ratio and reduce the losses to the organizations.

Using the past data can help analyze the situations prevailing in the organization for finding out root causes of problem and then making decisions. This process is called Data driven decision making. Many researches have tried to prove the fact that this principle can actually boost up the organizations and uplift the profits. There comes the role of Machine learning. These algorithms are capable of learning from the data and then predict the consequences and make the decisions. Since the human brain cannot handle a large amount of data at same time and as the complexity of data increases the computations are getting tougher, the use of machine learning based analysis come into picture [6]. One of the application in making a HR through machine model is drafting and predicting Employee retention. This paper addresses the employee retention prediction. Fig.2 shows the relationship between Machine learning and data analysis.

Authorized licensed use limited to: Birla Inst of Technology and Science Pilani Dubai. Downloaded on June 02,2023 at 14:02:30 UTC from IEEE Xplore. Restrictions apply.

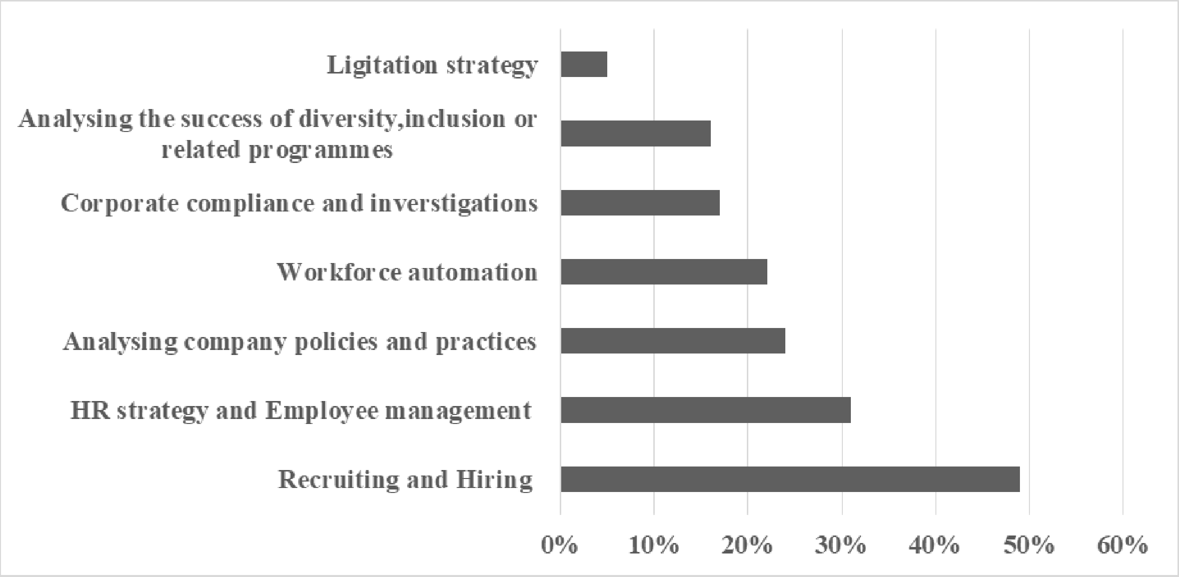

*Fig. 2 Machine learning and data analysis in HR [7]*

This will help HR managers to redraft the polices which can satisfy the employees on one hand and on the other ensures organization growth.
A proper blend of Human judgment strategies and Machine learning algorithms capabilities like analysis and prediction is a key towards successful HR model for an organization. The HR is a store house of the data but hardly the mangers have the capabilities to do in-depth analysis of the data. Therefore in conjunction with the application of this model the mangers can make successful decisions and policies for retention and predict the future perspectives of the organizational model. This paper will implement the state of art classifiers to design best suited model for this prediction problem based on the dataset and its parameters.

## II. METHODOLOGY

### A. Dataset and its arrangements

For making this predictive analysis we have applied machine learning algorithms on the HR data available on Kaggle [8]. The dataset contains records for 14999 employees based on 10 parameters. The dataset need to be pre-prepared manually in order to get the efficient model developed.

The process followed to predict attrition analyses the reasons behind the attrition of employees. The predictive model thus build can predict the retention possibilities by considering the input statistical parameters.

The step by step process is listed as below:

1. Dataset collection and identification: The acquired dataset must contain preset and past observations of diverse departments across the company.
2. Dataset preconditioning: The data preconditioning and cleaning is required in order to reduce the false predictions and training time.
3. Analyze the data subjectively for detecting the most important attributes and trends in the data.
4. Apply various machine learning models and split the data for training and testing scenarios. Choose the best model in terms of accuracy to develop predictive solution.
5. From the comparative analysis done, select the best fit model for the data and release the developed solution to the company.

*Table I Dataset description*

| Attribute | Type | Values |
|---|---|---|
| satisfaction_level | Numeric ( Fractional) | 0.09 to 1 |
| last_evaluation | Numeric ( Fractional) | 0.36 to 1 |
| number_project | Numeric ( whole) | 2 to 7 |
| average_montly_hours | Numeric ( whole) | 96 to 310 |
| time_spend_company | Numeric ( whole) | 2 to 10 |
| Work_accident | Binary | 0 or 1 |
| promotion_last_5years | Binary | 0 or 1 |
| salary | String | low, medium, high |
| Department | String | sales,technical, support, IT, product_mng |
| Left | Binary | 0 or 1 |
| Total number of records- 14999 | | |
| Total retained- 11428 | | |
| Total left- 3571 | | |

Table I gives the detailed description of the dataset used for this predictive analysis. This data is to be subjected to Classification model for prediction. Table II gives details of four instances taken from dataset with ten attributes.

Authorized licensed use limited to: Birla Inst of Technology and Science Pilani Dubai. Downloaded on June 02,2023 at 14:02:30 UTC from IEEE Xplore. Restrictions apply.

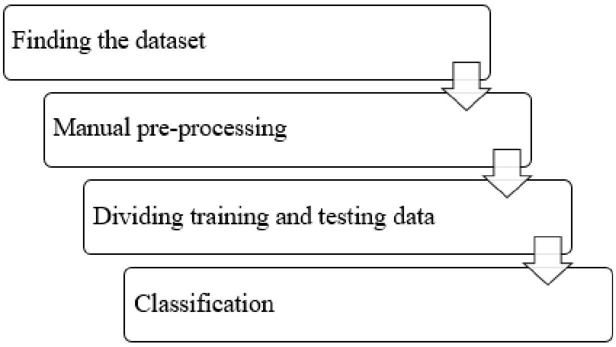

*Fig. 3 Proposed workflow*

Fig. 3 shows the proposed workflow for prediction model development. We have applied certain pre-processing on the data for making the model efficient one.

a) The attribute department is deleted from the dataset because we are making the decision for whole organization irrespective of the department. This will be required for predictive analysis of a particular department.

b) The attribute salary carries string values which is not suitable for classification model. Therefore the string is mapped to numeric values 0, 1, 2 corresponding to low, medium, and high.

c) The ratio of one class record vs. other is high i.e. the retained records are high in number as compared to the left. For better analysis a careful selection of records is to be made.

d) Moving the attribute named left describing the employee retention to the rightmost column. This is done so as to make it a response variable used for prediction with two classes 0 and 1.

*Table II Instances from dataset*

| instance | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_company | Work_accident | left | promotion_last_5years | Department | salary |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | sales | low |
| 1 | 0.8 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | sales | medium |
| 2 | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | sales | medium |
| 3 | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | sales | low |
| 4 | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | sales | low |

*Table III Dataset statistical analysis*

| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_company | Work_accident | left | promotion_last_5years |
|---|---|---|---|---|---|---|---|---|
| count | 14999 | 14999 | 14999 | 14999 | 14999 | 14999 | 14999 | 14999 |
| mean | 0.612834 | 0.716102 | 3.803054 | 201.0503 | 3.498233 | 0.14461 | 0.238083 | 0.021268 |
| std | 0.248631 | 0.171169 | 1.232592 | 49.9431 | 1.460136 | 0.351719 | 0.425924 | 0.144281 |
| min | 0.09 | 0.36 | 2 | 96 | 2 | 0 | 0 | 0 |
| 25% | 0.44 | 0.56 | 3 | 156 | 3 | 0 | 0 | 0 |
| 50% | 0.64 | 0.72 | 4 | 200 | 3 | 0 | 0 | 0 |
| 75% | 0.82 | 0.87 | 5 | 245 | 4 | 0 | 0 | 0 |
| max | 1 | 1 | 7 | 310 | 10 | 1 | 1 | 1 |

Authorized licensed use limited to: Birla Inst of Technology and Science Pilani Dubai. Downloaded on June 02,2023 at 14:02:30 UTC from IEEE Xplore. Restrictions apply.

From the data analysis as in Table III, it can be inferred that the employees not getting promotion and salary hike will tend to leave the company. So the redesigned HR model will look for alternative option to satisfy employees and save the interest of the organization.
Using these Modifications on the dataset, following information is pertained:

a) 8predictors:- satisfaction_level,last_evaluation,number_project,average_montly_hours, time_spend_company,Work_accident, promotion_last_5years, salary.
b) 1 response variable

The target feature in the dataset is represented by the variable Attrition, where "No" denotes a departed employee and "Yes" denotes a retained employee. With the use of this dataset, the machine learning system will be able to learn from actual data as opposed to explicit programming. The predictions produced in the output will be more precise if this training process is carried out repeatedly over time and on relevant data.
The dataset is now prepared for the classification process. There are no missing records otherwise the missing records are to be replaced with an alternative value or have to be deleted from the data.

### *B. Training setup*

The huge data available need to be segregated for training and testing purposes. We have taken a ratio of 80:20 for training and testing processes [9].

Fig. 4 shows the division of data done in order to present the training data to the classifier. As the number of cases belonging to class label 0 i.e. retained employees is very high as compared to the class label 1, this is the case of imbalanced classification. Class label 0 will be treated as majority class and class label 1 will be treated as majority class. A suitable classifier is to be selected to get the accurate results [10].

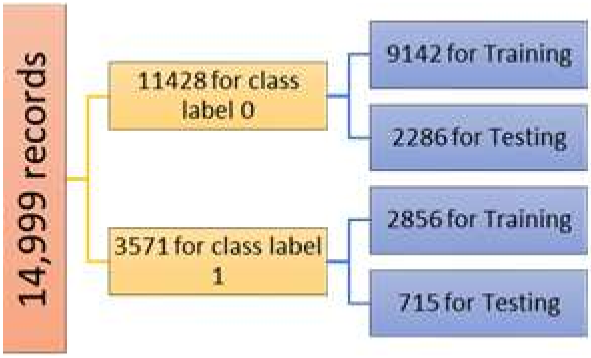

*Fig. 4 Data division for Training and testing*

### *C. Classification model*

With the appropriate training and testing division of the data, the next comes in picture is the choice of classification algorithm for development of model [11]. Since for our dataset taken under consideration is the clear case of imbalanced classification, an appropriate algorithm must be selected and the comparison with the other algorithms need to be made to select the best possible solution. The imbalanced dataset will be there for many other real life problems, so the approach for selection will depend upon the application under study [12]. The classifiers used are Decision tree (DT), Support Vector machine (SVM), K nearest neighbor (KNN), Naïve Bayes (NB) and Boosted tree ensemble are implemented and evaluated [13-16].

### *D. Evaluation method*

In evaluating the performance and comparing the classification algorithms a 5 fold cross validation scheme and ensuring non-overlapped data is utilized. The accuracy will be computed and comparison will be drawn.

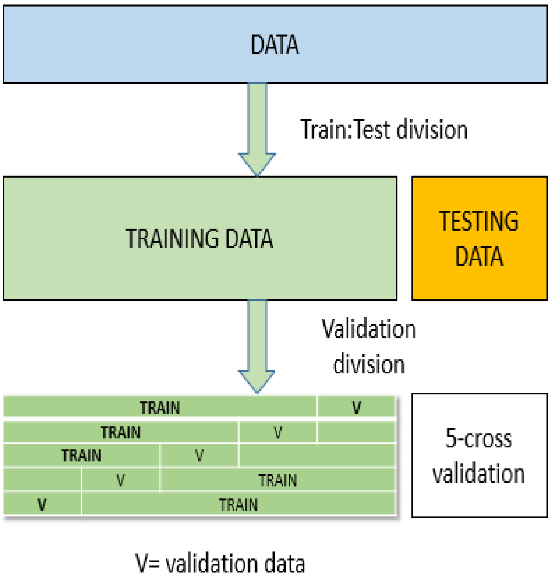

*Fig. 5 K cross validation scheme*

## III. RESULTS AND DISCUSSION

The simulation of this proposed workflow is done through MATLAB with a dedicated Classification Leaner application. The purpose of this work is to develop the model which can help HR mangers to predict the future retention of the employee based on the recorded data of previous employees.

*Table IV Accuracy Evaluation*

| Classifier | Training Vs. Testing data | Number of Predictors | Validation | Accuracy |
|---|---|---|---|---|
| NB | 80:20 | 8 | 5 | 90.3% |
| KNN | 80:20 | 8 | 5 | 93.6% |
| DT | 80:20 | 8 | 5 | 98.0% |
| SVM | 80:20 | 8 | 5 | 78.1% |
| Boosted tree Ensemble | 80:20 | 8 | 5 | 97.5% |

Authorized licensed use limited to: Birla Inst of Technology and Science Pilani Dubai. Downloaded on June 02,2023 at 14:02:30 UTC from IEEE Xplore. Restrictions apply.

Since these algorithms can handle high dimensional and complex data this can be savior for designing the new policies and strategies. The dataset as discussed in the previous section is pre-prepared and divided is the n passed onto classifier for prediction purposes.

Table IV displays the values of accuracies obtained for accuracy of various machine learning algorithms. The best performance is obtained from Decision tree and the worst case performance belongs to SVM for this dataset. The testing of trained model is done with the help of testing data divided. The predicted value from the classifier is matched with the actual response value from the dataset and the confusion matrix will be plotted on the basis of true positive rate and false positive rate.

Fig. 5 displays the confusion matrix plot for DT classifier as it is found to the optimum choice for this dataset.

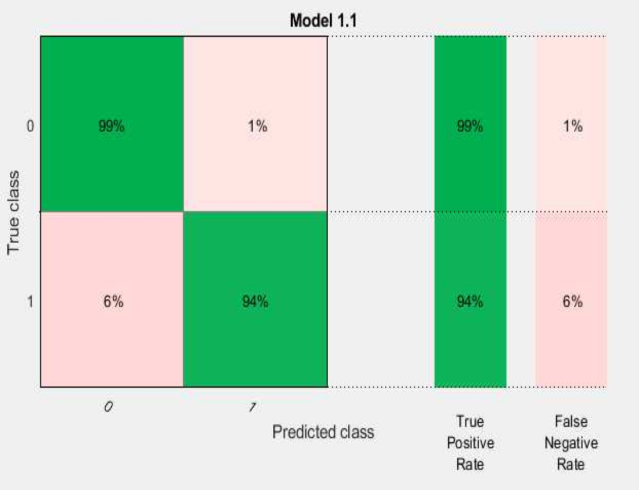

*Fig. 6 Confusion Matrix for DT model*

Fig. 6 shows the graphical comparison of various models implemented in this study. The accuracy of DT and Ensemble with Boosted tree is close enough as shown in Fig.7 and hence can be alternatively chosen for imbalance classification problems relating to real world.

## CONCLUSION

The Machine learning models are promising tool for data analysis and decision making process. One such data analytics problem i.e. analyzing employee's data to predict retention trends is taken up in this work.

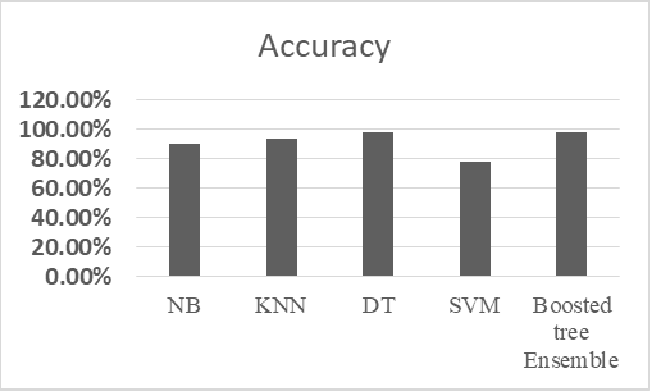

*Fig. 7 Accuracy comparison chart*

The machine learning based model will be a helping hand to the managers for drafting up the decisions to improve retention ratio by rectifying the root cause of the employee leaving the organization that too maintained in the interest of the organization. In future this work can be extended for the more detailed analysis comprising various attributes relating to employee retention and thereby selecting the most appropriate attributes for classification model.

## REFERENCES

[1] 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.

[2] Xu, Zhen, and Binheng Song. "A machine learning application for human resource data mining problem." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 847-856. Springer, Berlin, Heidelberg, 2006.

[3] Wood, Stephen. "Human resource management and performance." International journal of management reviews 1, no. 4 (1999): 367-413.

[4] http://www.thestaffingstream.com/ [Accessed on 28-3-20]

[5] Herman, Roger E. "HR managers as employee-retention specialists." *Employment Relations Today* 32, no. 2 (2005): 1.

[6] Jaiswal, Akanksha, C. Joe Arun, and Arup Varma. "Rebooting employees: Upskilling for artificial intelligence in multinational corporations." The International Journal of Human Resource Management 33, no. 6 (2022): 1179-1208.

[7] https://www.raconteur.net/ [Accessed on 28-3-20]

[8] https://www.kaggle.com/liujiaqi/hr-comma-sepcsv [Accessed on 1-4-20]

[9] 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.

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

[11] 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.

[12] Lanzi, Pier L. *Learning classifier systems: from foundations to applications*. No. 1813. Springer Science & Business Media, 2000.

[13] Kononenko, Igor. "Semi-naive Bayesian classifier." In *European Working Session on Learning*, pp. 206-219. Springer, Berlin, Heidelberg, 1991.

[14] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." *Expert Systems with Applications* 38, no. 3 (2011): 1999-2006.

[15] Rish, Irina. "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41-46. 2001.

[16] 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.

[17] Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." *algorithms* 4, no. 5 (2016): C5

Authorized licensed use limited to: Birla Inst of Technology and Science Pilani Dubai. Downloaded on June 02,2023 at 14:02:30 UTC from IEEE Xplore. Restrictions apply.