A comparative study on machine learning algorithms for employee attrition prediction

P M Usha¹ and Dr N V Balaji²

¹Research Scholar, Department of CA, CS & IT, Karpagam Academy of Higher Education, Coimbatore. E mail Id – pmusha.72@gmail.com
²Dean, Faculty of Arts, Science and Humanities, Karpagam Academy of Higher Education, Coimbatore. E mail Id – balajinv@karpagam.com

Abstract: The fourth industrial revolution introduces a wide range of technologies for an effective functioning of organizations with optimal usage of all resources including human resource. Machine learning is one of the driving technologies implemented in fourth Industrial Revolution. Machine learning can be utilised for developing models that can predict the retention or attrition of employees. The study is using machine learning algorithms like classification and clustering for preparing the prediction models. A comparison of these algorithms is done based on its performance. The performance is measured using parameters like accuracy, precision, recall, F Measure and time taken to build the model. The study is also finding the correlation between variables used in the work to the decision of staying back in the organization. The study is using the open-source tool Weka and also python for doing the same.

1. Introduction

The performance of an organization depends on the availability of adequate resources at its disposal. Human resources are considered as the most coveted resources of any organization. A replacement for a human resource is not as easy as replacement of any other resource. An organization can perform well only with right number of people with right skills to perform the right task. Stating the same reason, the attrition of employees is considered as a great menace for organizations.

The fourth industrial revolution introduces a wide range of technologies where human intelligence is emulated in performing various tasks. Machine learning is one of the driving technologies implemented in fourth Industrial Revolution. Machine Learning can develop technology to help smart industries [1]. It leads to an effective functioning of organizations with optimal usage of all resources including human resource. Machine learning can be utilised for developing models that can predict the retention or attrition of employees. Reducing the attrition rate is very important because losing an employee can be very costly for the organization as it is difficult to get an exact replacement. Selection and training of newly recruited employee can also cost heavy for the organization [2]. The study is using machine learning algorithms like classification and clustering for preparing the prediction models [3]. The algorithms used in this study are Naïve Bayes, J48, Decision Table and KMeans. A comparison of these algorithms is done based on its performance. The performance is measured using parameters like accuracy, precision, recall, F Measure and time taken to build the model. The dataset used in this study is collected in real time using a well-structured questionnaire from the employees of an organization working in IT sector. The study is also finding the correlation between variables used in the work to the
decision of staying back in the organization. The study is using the open-source tool Weka [4] and also python for doing the same. Weka has got readymade tools for carrying out certain classification and clustering algorithms [5].

The organization of the paper is as follows. Next section explains methodology and background of work. Section 3 explains the results and discussion followed by a conclusion section.

2. Methodology
The researcher is using a real time data set for conducting the study. The data is collected using a well-structured questionnaire, which includes variables that influence the decision of an employee to leave the organization. Data is collected from an organization which is functioning in IT sector. The demographic features of employees like age, qualification, gender, experience, marital status, number of companies worked, qualification, number of dependents, monthly income and also the satisfaction level of employees in various aspects of organization like compensation, freedom to work, recognition given by superiors, promotion policy, growth opportunities, clarity of role, relationship with superiors, relationship with colleagues, training, work life balance, office timing and the like were collected using a well-structured questionnaire which was distributed to the employees using google form. The collected data was saved as a comma separated file (CSV).

The study is using python to understand the correlation between the variables used in the study to the target variable planstocontinue which denotes whether the employee has plans to continue in the organization for a longer period. This helps in identifying the variables that influences the decision of an employee leaving the organization the most.

Employee attrition prediction models are built using Weka and a comparison is done between the models based on accuracy, prediction, F Measure and time taken for building the model.

2.1 Correlation of variables in the data set
The dataset was given as input to a program written in python using the Anaconda Navigator and Jupyter Notebook to calculate the correlation between the listed variables and the target variable planstocontinue. Anaconda Navigator is a Python and R Distributor. It has editing tool, python interpreter and python packages. Jupyter notebook is opened from Anaconda and is used for creating and editing of python code. The correlation between variables are found by calculating the correlation coefficient using corr() function in python. An output of 1 indicates perfect positive correlation and if the output is -1 it indicates perfect negative correlation between the variables selected. If the output is near to 1 it indicates good correlation which denotes that the increase in one variable will result in increase of the other and vice versa. The correlation of the target variable and other variables were calculated [6].

2.2 Comparison of machine learning algorithms for building attrition prediction model
Organizations end up in a dangerous state of affairs when productive employees resign and leave the organization. If a machine learning model could be developed to predict the attrition of employees, necessary precautions could be taken to avoid that situation [7]. In this study, a comparison of the performance of various algorithms in constructing a model for the collected data set is executed. The models where developed using Weka which is an open-source software. Weka contains tools which can do data pre-processing, clustering, classification and the like [4]. The algorithms used for comparison in this study are Naïve Bayes, Decision table, Random Forest, J48 and K-Means. The parameters used to compare the performance of the algorithms are accuracy, precision, recall and F1 score. The following parameters are calculated using TP, TN, FP and FN. TP stands for True Positive which denotes the number of observations which are positive and also predicted as positive whereas FP stands for False Positive where the observation is negative, but is predicted as positive. TN stands for True Negative where the observation is negative and predicted as negative and FN is where the observation is positive where it was predicted as negative [7].
Accuracy stands for the ratio of correct predictions = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)  

(1) 

Precision = The ratio of correct positive predictions to all positive predictions = TP / (TP + FP)  

(2) 

Recall = The ratio of correct positive predictions to actual number of positive case = TP / (TP + FN)  

(3) 

F1score = (2*Recall*Precision) / (Recall + Precision)  

(4) 

2.2.1 Naive Bayes. Naïve Bayes is a classification algorithm which is dependent on Bayes Theorem. In Naïve Bayes classification we assume that each event is independent of the other [8]. Bayes theorem is based on the mathematical formula, 

\[ P(Y/X) = P(X/Y) \cdot P(Y)/P(X) \]

where X represents attribute, Y represents class, P(Y/X) represents the probability of occurrence of event Y given that event X has occurred, P(X/Y) represents the probability of occurrence of event X given that event Y has occurred, P(Y) is the probability of occurrence of event Y and P(X) denotes the probability of occurrence of X.

2.2.2 Decision Table. Decision table is a machine learning algorithm coming under the category of classification algorithms. Weka provides a tool for doing classification using decision table. A decision table is a visual representation of decisions taken under different sets of conditions. In decision table rows can represent objects, columns represent attributes of objects and a special column can mention the decision.

2.2.3 J48. J48 is a decision tree algorithm. It is a classification algorithm. Weka has a readymade tool to perform J48 classification. In decision tree, branches are formed based on the values of the node. This helps in forming an appropriate classification. The attribute that has highest information gain and can classify the instances clearly is identified for the formation of the tree.

2.2.4 Random Forest. Random Forest is a classification algorithm which creates many decision trees. Each tree predicts from samples and the best answer is selected by way of voting. Different subsets of the features in the dataset are used for training separate decision trees. The decisions of each tree are aggregated to form the end decision. This process is called as bagging.

2.2.5 Simplekmeans. K-Means algorithm comes under the category of clustering algorithms. In k-means the centroids are randomly selected and the distance from each data point to the cluster centres is calculated. The datapoint is assigned to the cluster with minimum distance to the centre. Based on the new assignments, the centroids are again calculated. The iteration continues until there are no more changes. The algorithm will give k clusters as output.

3. Results and discussions

3.1 Correlation

After executing the corr() function in python with plans to continue as the target variable, the following results were obtained. The variable showing higher correlations are given below.
Table 1: List of fields having higher correlation with plans to continue.

| Name of the field | Correlation coefficient | Name of the field | Correlation coefficient |
|-------------------|-------------------------|-------------------|-------------------------|
| Compensation      | 0.463996                | Wlb               | 0.540620                |
| Feedback          | 0.474139                | Promotionpolicy   | 0.571525                |
| Superiorrelation  | 0.455532                | Officetiming      | 0.475550                |
| Recognition       | 0.552022                | Teamwork          | 0.508224                |
| Growthopp         | 0.525373                | Jobsecurity       | 0.579873                |
| Training          | 0.459784                | proud             | 0.653296                |
| Opptouseskills    | 0.457808                |                   |                         |

The above data shows that some variables are having a positive correlation to the decision of continuing in the organization for a longer period. For example, the variables like proud to work in organization (proud), job security (jobsecurity), promotion policy (promotionpolicy), work life balance (wlb), recognition and growth opportunity (growthopp) are having high positive correlation to the plan to continue in the organization. These results can be visually represented using a graph using the countplot() function in python. The following figures give a visual representation of the correlations present in the dataset.

Figure 1: The chart representing the relation between the variables proud and plans to continue.

In the above figure, 1 represents “No” with a meaning that there are no plans to continue and 2 represents “Yes” which denotes that there are plans to continue. The chart above shows that majority of people who strongly agree and agree that they are proud to work in the organization have plans to continue in the organization. People who are neutral and disagree have no plans to continue in the organization.

Figure 2: The chart representing the relation between the variables job security and plans to continue.
In the above figure, 1 represents “No” meaning there are no plans to continue and 2 represents “Yes” which denotes that there are plans to continue. In the x-axis 5 denotes highly satisfied, 4 denotes satisfied, 3 is neutral, 2 is dissatisfied and 1 is highly dissatisfied. The chart above shows that majority of people who are highly satisfied and satisfied with job security have plans to continue in the organization. People who are dissatisfied and highly dissatisfied have no plans to continue in the organization.

![Chart](chart.png)

**Figure 3:** The chart representing the relation between the variables promotion policy and plans to continue.

In the above figure, 1 represents “No” meaning there are no plans to continue and 2 represents “Yes” which denotes that there are plans to continue. In the x-axis 5 denotes highly satisfied, 4 denotes satisfied, 3 is neutral, 2 is dissatisfied and 1 is highly dissatisfied. The chart above shows that majority of people who are highly satisfied and satisfied with the promotion policy have plans to continue in the organization. People who are dissatisfied and highly dissatisfied have no plans to continue in the organization.

The study has found the correlation between the target variable “planstocontinue” with other variables. It was found that the variables proud, jobsecurity, promotionpolicy, wlb, recognition and growthopp were having more correlation with the target variable planstocontinue. This illustrates that the satisfaction level of employees regarding the factors like pride to work in the organization, job security, promotion policy of the organization, work life balance, recognition given for work by superiors and the growth opportunities in the organization have great impact on the employee’s decision to stay back in the organization.

3.2 A mathematical comparison of the results after implementation of selected algorithms

The study is also comparing the performance of various algorithms in predicting the decision of an employee to continue in the organization.

3.2.1 Naïve Bayes. The Naïve Bayes model gave the following result after computation.  
\[ TP = 58, \; TN=34, \; FP = 6, \; FN = 9 \]  
\[ \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{58+34}{107} = 0.8598 \]  
\[ \text{Precision} = \frac{TP}{TP+FP} = \frac{58}{58+6} = 0.906 \]  
\[ \text{Recall} = \frac{TP}{TP+FN} = \frac{58}{58+9} = 0.8657 \]  
\[ F\text{-Measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} = 0.885 \]  
\[ \text{Time taken to build model:} \; 0 \text{ seconds} \]

3.2.2 Decision table. The results of implementation of decision table classifier are as follows.  
\[ TP = 58, \; TN=23, \; FP = 17, \; FN = 9 \]  
\[ \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{58+23}{107} = 0.757009 \]  
\[ \text{Precision} = \frac{TP}{TP+FP} = \frac{58}{58+17} = 0.773 \]
Recall = TP/(TP+FN) = 58/ (58+9) = 58/67 = 0.8657
F-Measure = (2*Recall*Precision)/(Recall+Precision) = 0.8167
Time taken to build model: 0.05 seconds

3.2.3 J48. The results after the execution of J48 are as follows.
TP = 63, TN=28, FP = 12, FN = 4
Accuracy = (TP+TN)/(TP+TN+FP+FN) = (63+28)/107=0.8505
Precision = TP/(TP+FP) = 63/ (63+12) = 0.84
Recall = TP/(TP+FN) = 63/ (63+4) = 63/67=0.9403
F-Measure = (2*Recall*Precision)/(Recall+Precision) = 0.8873
Time taken to build model: 0.03 seconds

3.2.4 Random Forest. The random forest resulted in the performance as mentioned below.
TP = 63, TN = 28, FP = 12, FN = 4
Accuracy = (TP+TN)/(TP+TN+FP+FN) = (63+28)/107=0.8505
Precision = TP/(TP+FP) = 63/ (63+12) = 0.84
Recall = TP/(TP+FN) = 63/ (63+4) = 63/67=0.9403
F-Measure = (2*Recall*Precision)/(Recall+Precision) = 0.8873
Time taken to build model: 0.06 seconds

3.2.5 Simplekmeans. TP = 45, TN = 32, FP = 8, FN = 22
Accuracy = (TP+TN)/(TP+TN+FP+FN) = (45+32)/107=0.7196
Precision = TP/(TP+FP) = 45/ (45+8) = 0.8491
Recall = TP/(TP+FN) = 45/ (45+22) = 0.6716
F-Measure = (2*Recall*Precision)/(Recall+Precision) = 0.75
Time taken to build model: 0.01 seconds

| Algorithm    | Accuracy | Precision | Recall | F1score | Time in seconds |
|--------------|----------|-----------|--------|---------|-----------------|
| Naïve Bayes  | 85.98    | 90.6      | 86.57  | 88.5    | 0               |
| Decision Tree| 75.7     | 77.3      | 86.57  | 81.67   | 0.05            |
| J48          | 85.05    | 84.0      | 94.03  | 88.73   | 0.03            |
| Random Forest| 85.05    | 84.0      | 94.03  | 88.73   | 0.06            |
| KMeans       | 71.96    | 84.91     | 67.16  | 75      | 0.01            |

The above table shows that Naïve Bayes is performing better than other algorithms considering accuracy as well as time followed by J48 and Random Forest. Considering the time for building the model, J48 is exhibiting a better performance. Clustering gives lesser accuracy compared to the classification algorithms. Considering F1 score, J48 and Random Forest are showing the best output. F1 score is considered as a better parameter for performance comparison in case of datasets which are imbalanced. The dataset used here is a balanced dataset.
Figure 4 showing the comparison of various algorithms

4. Conclusion
The retention of employees is very much crucial as far as organizations are concerned. The employee’s attrition can be reduced to an extent if we are able to identify the factors leading to attrition and also the factors leading to retention [9]. Machine learning can be identified as a very good tool for developing models for predicting attrition as well as retention [10]. The study has used python to find correlation of variables and has also compared the performance of various classification and clustering algorithms using the software Weka which is open-source software. The correlation coefficient calculated using python is giving the correlation between the variables like satisfaction level of employees regarding salary, training, job security and the like with the target variable plans to continue. Hence the study has revealed that the factors like pride to work in the organization, job security, promotion policy of the organization, work life balance, recognition given for work by superiors and the growth opportunities in the organization have great impact on the employee’s decision to stay back in the organization which in turn will help the organization to keep things in order. The efficiency of attrition prediction models is compared and it is found that for the dataset used in the study, maximum efficiency is achieved by Naïve Bayes method.

References
[1] Pedro Larrañaga, David Atienza, Javier Diaz-Rozo 2018 “Industrial Applications of machine learning”, CRC Press, London
[2] Jain P K, Jain M & Pamula R 2020 Explaining and predicting employees’ attrition: a machine learning approach. SN Appl. Sci. 2, 757
[3] https://tdwi.org/articles/2019/04/05/adv-all-predictive-analytics-and-employee-retention.aspx
[4] https://www.cs.waikato.ac.nz/ml/weka/
[5] Monika Lambsonge, Girish Talmale, Poonam Prasad February-2020 Review paper on: WEKA Tool With different Data Mining Technique, International Journal of Emerging Technologies and Innovative Research, Vol.7, Issue 2, page no.891-897
[6] Kothari C R 2019 “Research Methodology Methods and Techniques”, New Age International (P) Limited
[7] Alduayj Sarah & Raijpoor, Kashif 2018, Predicting Employee Attrition using Machine Learning, 2018 13th Int. Conf. on Innovations in Information Technology (IIT)
[8] Nafizatus Salmi and Zuherman Rustam 2019, Naïve Bayes Classifier Models for Predicting the Colon Cancer, IOP Conf. Series: Materials Science and Engineering
[9] Sri Harsha B, A. Jithendra Varaprasad, L V N Pavan Sai Sujith March 2020, “Early prediction of employee attrition”, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY
[10] Mohbey K K, (2020) Employee's Attrition Prediction Using Machine Learning Approaches, Machine Learning and Deep Learning in Real Time Applications, pp 121-128