**International Journal of Commerce and Management Research**
**ISSN: 2455-1627; Impact Factor: RJIF 5.22**
**Received: 28-04-2020; Accepted: 18-05-2020; Published: 31-05-2020**
**www.managejournal.com**
**Volume 6; Issue 3; 2020; Page No. 80-85**

# Employee attrition prediction using neural network cross validation method

**Shawni Dutta[1], Samir Kumar Bandyopadhyay[2*]**
[1] Lecturer, Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India
[2] Academic Advisor, The Bhawanipur Education Society College, Kolkata, India

**Abstract**
Any organization or company is strongly aware of the significance of employees in gaining and upholding competitive advantage. While putting concentration on earning maximized profit, employee attrition rates should be considered as an interfering factor. This paper emphasizes on predicting attrition probabilities beforehand by implementing an automated tool. The proposed system implements feed-forward neural network along with 10-fold cross validation procedure under a single platform for predicting employee attrition. This proposed method is evaluated as well as compared with six classifiers such as Support Vector Machine, k-Nearest Neighbor, naïve bayes, Decision Tree, Adaboost, and Random Forest classifiers. Experimental analysis concludes that proposed method outperforms well over aforementioned classifiers in terms of performance measure metrics.

**Keywords:** Employee attrition, predictive model, IT sector, neural network, cross-validation

## 1. Introduction
Employee Attrition denotes the continuous decrease in the number of employees by the process of retirement, resignation or death. However, the attrition rates may vary from one sector to other depending and following their standards, terms and policies. Attrition rates are quite serious addressable problem for any industry and companies which drive them for effective retention of talent [1]. For reaching maximized revenue, an organization must put concentration on employee attrition rates. Organization policies may include retention schemes to maintain the resource of employees. A transparent system is introduced in this paper that automatically recommends the attrition possibilities of employees. Early attrition prediction of a particular employee may help them to increase the efficiency and dedication for company.
This paper focuses on implementing an automatically captures considering related factors that accelerate employee attrition process. Data mining and knowledge discovery process can assist this automated tool for identifying the interfering factors and relationships among them. Given a set of messages, Machine Learning methods can acquire information and later can use the acquired information to classify unknown new messages. Employee attrition likelihood can be predicted by employing supervised machine learning approaches. Analyzing the past and existing employee information is exploited while designing the predictive model. To address the problem of aforementioned prediction, classification techniques are implemented that maps input variable to target classes by considering training data. The input variables are the interfering factors that include salary structure, work life balance, job satisfaction, comfort in working environment, relationship with supervisor and many more. All these data turn out to be good predictors while identifying employees having probable erosion. This prediction will in turn help the employees by alarming attrition in advance, thus allowing them take informed decisions and act accordingly. Prediction results will even assist the organization also for identifying employees having higher attrition rates and concentrating on them for survival.
A deep learning based feed-forward neural network with 10-fold cross-validation procedure is proposed in this paper that is dedicated for improving the efficiency in determining employee attrition likelihoods. Deep learning is an evolving area which exploits artificial intelligence as well as machine learning to learn features directly from the data, combining several nonlinear processing layers. A set of six classifiers are also implemented in this paper that serve as benchmark for comparing the proposed model in terms of prediction. The benchmark classifiers include Support Vector Machine (SVM) [2], Naïve Bayes (NB) [3], k-Nearest Neighbor (k-NN) [4], Decision Tree Classifier [5], and ensemble classifiers such as Random Forest (RF) [6] and Adaboost Classifiers [7]. This study has established that the proposed model is quite superior over other specified classifiers.

### Related Work
Using k-Nearest Neighbors algorithm prediction related to whether an employee will leave a company or not is evaluated in [8]. Employee Performance, average monthly hours at work and number of years spent in the company and among others are considered as features. Experimental result established a comparative study between Naïve Bayes, Logistic Regression, and Multi-layer Perceptron (MLP) Classifier and proved that k-Nearest Neighbors outperforms well over its peer algorithms. For preventing employee attrition, a couple of well-known classifiers such as Decision tree, Logistic Regression, SVM, KNN, Random Forest, Naive bayes methods on the human resource data are implemented in [9]. Feature selection methods are implemented on the data and analysis has shown results in order to prevent employee attrition.
Using logistic regression methods employee attrition prevention mechanism is proposed in [10]. In this context, demographic data of divided as well as present employees

are gathered. A risk equation was prepared and finally cluster of high risk employees was generated. For stopping those employees from attrition, organization must provide attention on that cluster. In [11], prediction of employee attrition using various data mining techniques such as Random Forest, Support Vector Machines (SVM), Gradient Boosted Classifier and Logistic Regression are discussed. Experimental analysis has shown that Extreme Gradient Boosting is superior over classifier related to attrition prediction task.

Numerous data mining techniques are applied for prediction of employee turnover in [12]. Historical and personal data of the employee are utilised for this purpose. Considering HR related data, data mining techniques are applied from prediction to classification of related features [12]. In [13], employee churn prediction is focused which has close relation to customer churn prediction. In this study, a several well-known classification methods including, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes methods are applied on the HR data. Obtained results are analysed by calculating the accuracy, precision, recall, and F-measure values of the results. A feature selection method on the data is implemented and the results with previous ones are analysed.

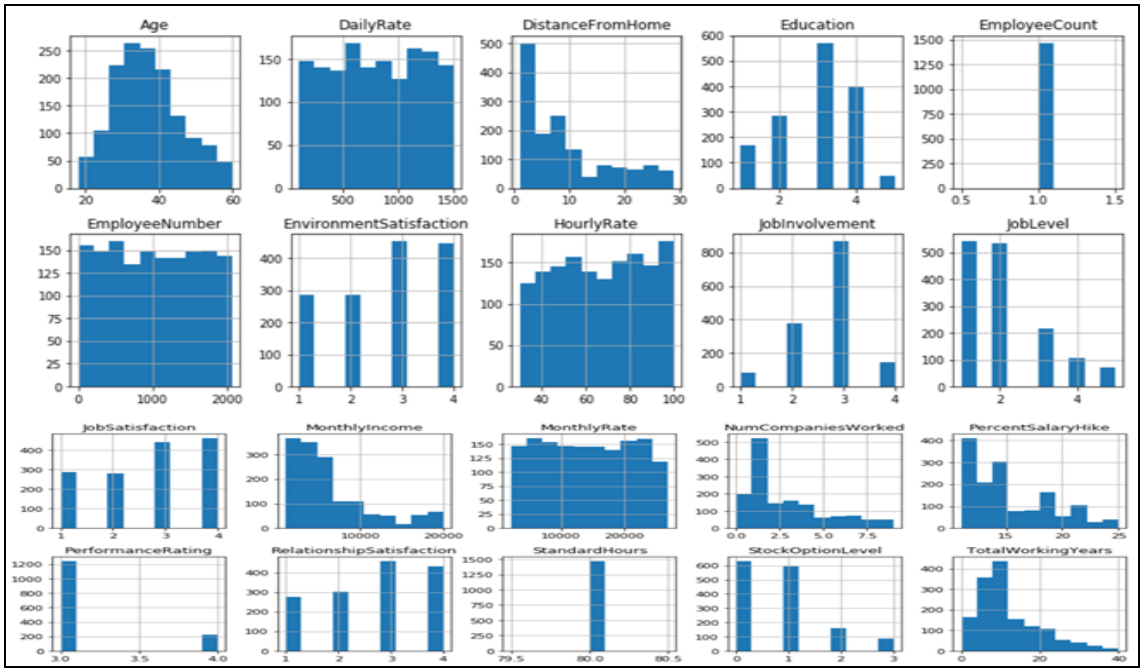
Coussement and Van den Poel in [14] implemented SVM method for predicting customer churns using two parameter-selection techniques. Both techniques are based on grid search and cross-validation. Application of support vector machines provides good results as generalization performance. A model is developed in [15] to predict employee attrition. This in turn will provide the organizations opportunities to address any issue and improve retention. Supervised machine learning algorithm support vector machine (SVM) is applied on Archival employee data (comprising of 22 input features) were collected from Human Resource databases. The database records for three IT companies in India, including their employment status (response variable) at the time of collection. Results analysis reported that the SVM model has an accuracy of 85% [15].

**Proposed Methodology**

Data mining techniques are applied in this paper for the purpose of employee attrition prediction. In brief, the data mining process is consisting of two major processes such as data pre-processing and classification/clustering. Data pre-processing (or attribute selection) is one of the most important steps in the data mining process. Pre-processing techniques filter out redundant or irrelevant information from the original data. After this, the classification or clustering step is performed for the task of prediction, estimation, etc. [16]. Following section elaborates application of data mining techniques for attrition prediction.

**Dataset Collection and Preprocessing-**

In this context, dataset related to employee attrition are obtained from kaggle [17]. The dataset consists of 1470 number of sample records and each of which can be framed as collection of attributes that include several criterions for detecting employee attrition tendency. The attribute list includes Employee's age, daily rate, distance from home, education, employee count, employee number, environment satisfaction, hourly rate, job involvement, job level, job satisfaction, monthly income, monthly rate, number of companies worked, hike in salary percentage, performance rating, relationship satisfaction, stock market option, standard hours, total working years, training times since last year, balance in work life, years spent in company, years spent in current role, years since last promotion, relationship with current manager in terms of years. Attribute 'attrition' is used as target variable of classification procedure. Following diagram Fig.1 shows overall understanding of the dataset. Fig.2. shows distribution of attrition variables in terms of positive and negative cases. For obtaining a balanced dataset, preprocessing techniques such as checking and handling missing value, scaling some attributes are performed. Some of the attributes such as employee number, employee count are eliminated since they are not fruitful in prediction. Applying theses pre-processing techniques will yield a transformed dataset that can be fitted to classifier.



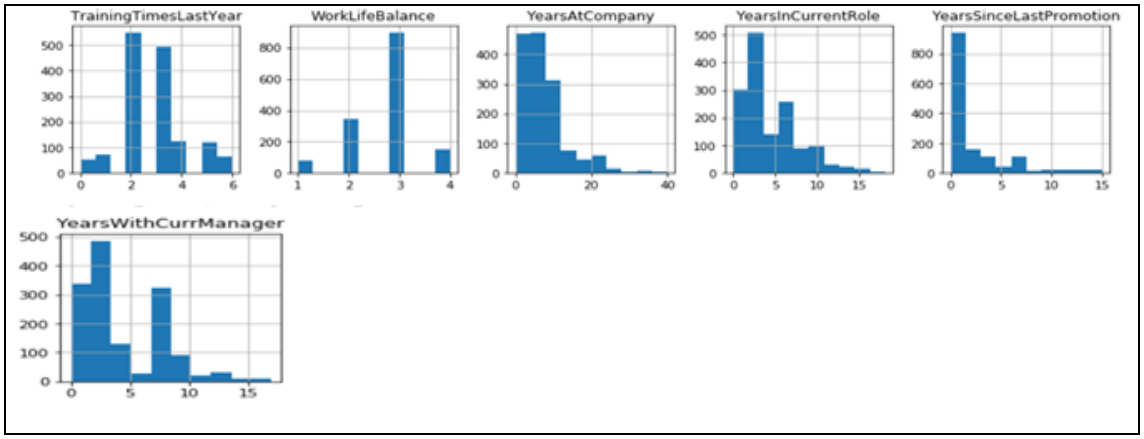



**Fig 1:** Histogram interpretation of collected dataset

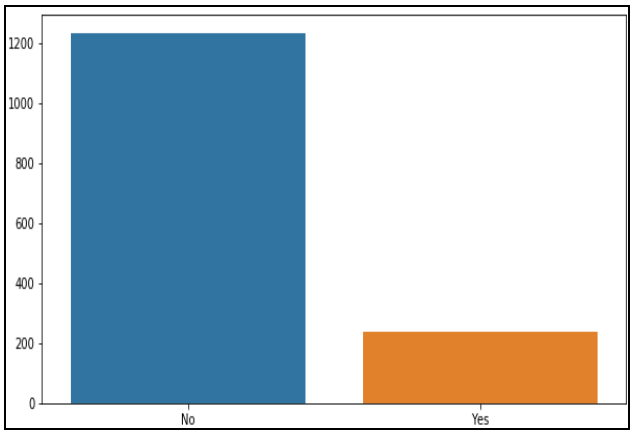



**Fig 2:** Distribution of target variable of collected dataset

## Methodology and Implementation

A classifier model maps input variable to target classes after learning from training data. The main objective of proposed classifier is to predict whether an employee may face attrition process or not using deep learning techniques. The essential constituent of deep learning [18] is the multi-layered hierarchical data representation typically in the form of a neural network by incorporating more than two layers. In this context, classifier Neural network is used that follows deep learning technique by coalescing multiple layers with linear or non-linear activation functions those are trained together for achieving complex problem solving approach. Activation functions [19] are suitable to execute diverse computations and produce outputs within a definite range. In other words, activation function is a step that maps input signal into output signal. 'relu' and 'sigmoid' are two popular activation functions that are explained as follows-

- Sigmoid activation function [19] transforms input data in the range of 0 to 1 and it is shown in equation (1).

$$f(x) = 1/(1 + exp^{-x}) \quad (1)$$

- Relu activation function [19] is a faster learning Activation function which is the most successful and widely used function. It performs a threshold operation to each input element where values less than zero are set to zero whereas the values greater or equal to zeros kept as intact and it is shown in equation (2).

$$f(x) = \max(0, x) = \begin{cases} Xi, & \text{if } Xi \geq 0 \\ 0, & \text{if } Xi < 1 \end{cases} \quad (2)$$

The feed-forward neural network proprosed in this paper consists of input layer, hidden layer and output layers. Neural Network is advantageous because of enhanced efficiency due to presence of huge number of well-interconnected processing components. These components work in harmony to achieve solution related to a specified problem [20]. As the data is passed through the nodes present in each layer, perspective and relationship among the data is realised and the output is generated. After configuring this neural model, training process is executed. The training process goes through one cycle known as an epoch where the dataset is partitioned into smaller sections. An iterative process is executed through a couple of batch size that considers subsections of training dataset for completing epoch execution.

## Implementation

In this framework, this neural network is implemented by considering three layers consisting of 32, 16, 1 number of nodes respectively. Each layer uses either 'relu' or 'sigmoid' as the activation function. Finally these layers are compiled using 30 epochs and with a batch size of 10. A popular optimizer known as 'adam' and 'binary crossentropy' function are also used for compilation. This model receives 1,601 number of parameters for training purpose. All aforementioned hypermeters are used as the best hyperparameters in order to achieve maximised performance. The model is summarised in figure 3.

After configuring the neural network model, it is followed by 10-fold cross validation method [21] for estimating the skill of the model. It is a resampling approach where the dataset is segregated into 10 groups and during each iteration one group is considered as the test data and the remaining nine folds are considered as training data. The above mentioned model is fitted into the training dataset and it is evaluated against the test dataset. Later evaluation scores for each of these iterations are accumulated and mean score is calculated.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense_31 (Dense) | (None, 32) | 1056 |
| dense_32 (Dense) | (None, 16) | 528 |
| dense_33 (Dense) | (None, 1) | 17 |

**Fig 3:** Summary of Neural Network Model.

This neural network structure along with 10-fold cross validation procedure is applied on employee attrition dataset. Implementation of this model is evaluated as well as compared with other benchmark classifiers such as SVM, Naïve Bayes Classifier, K-NN, Decision Tree, Adaboost, and Random. These six classifier models are considered as baseline for comparing the proposed method.

**Other classifiers and Implementation**

Classification is a supervised machine learning technique that analyses specified set of features and identifies data as belonging to a particular class. Different classification algorithms such as support vector machines, decision trees, K-nearest neighbour classifier, Adaboost, Gradient Boost are used to predict the target class. For these classifier models the pre-processed and transformed data are partitioned into training and testing dataset with the ratio of 7:3. Training dataset is fitted to the classifier and later predictions are obtained using testing dataset. Brief description of the classifiers are provided as follows-

*Support Vector machine (SVMs)* [2] belongs to the category of linear classifiers. It identifies different classes by separating samples with the help of decision boundary known as hyperplane. Both linear as well as non-linear data can be classified with the help of SVMs. It is also known as Maximum margin classifier since it can minimize the empirical classification error and maximize the geometric margin simultaneously [2].

The *Naive Bayes classifier* [3] is a supervised classification tool that exploits the concept of Bayes Theorem [3] of Conditional Probability. The decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate. This classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy.

*K-nearest neighbour (k-NN)* [4] is often considered as lazy learner which considers instances during classification process. It is known as lazy learners because during training phase it just stores training samples. This identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k.

A *Decision Tree (DT)* [5] is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Classification results are obtained by starting from the beginning at the root this tree are going through it until a leaf node is reached. Obtaining and training a DT model can forecast the value of a goal variable centered on a number of input variables [4].

Ensemble approach facilitates several machine learning algorithms to perform together to obtain higher accuracy of the entire system. Random forest (RF) [6] exploits the concept of ensemble learning approach and regression technique applicable for classification based problems. This classifier assimilates multiple tree-like classifiers which are applied on various sub-samples of the dataset and each tree casts its vote to the most appropriate class for the input.

For improving the accuracy of classification, multiple unstable learners are accommodated into a single learner using an efficient technique known as Boosting. Classification algorithms are applied to the reweighted versions of the training data and the weighted majority vote of the sequence of classifiers are chosen. AdaBoost [7] is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate.

**Implementation**

The above specified classifiers are implemented by considering and adjusting appropriate hyper-parameters for obtaining the maximised performance. The SVM classifier is implemented with 'rbf' kernel and regularization parameter C=1. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. For naïve bayes classifier, multinomial naïve bayes classifier is employed. The decision tree classifier implemented in this paper uses Gini index while choosing objects from dataset. The nodes of the decision tree are expanded until all leaves are pure or until all leaves contain less than minimum number of samples. In this case, minimum number of samples is assigned a value as 2. On the other hand, ensemble classifiers, such as, AdaBoost and Random classifiers are built based on 500 numbers of estimators on which the boosting is terminated.

**Performance Measure Metrics-**

In order to evaluate performance of a model, performance measure metrics are used. Evaluating any model with respect to these metrics will justify the performance of the model. Following are the metrics those are used for performance assessment-

1. *Accuracy* [22] is a metric that ascertains the ratio of true predictions over the total number of instances considered.
2. *Mean Squared Error* (MSE) [22] is another evaluating metric which is used for measuring absolute differences between the prediction and actual observation of the test samples.

Mathematically, the aforementioned metrics can be defined as follows-

$$\text{Accuracy} = TP+TN/(TP+FP+TN+TP) \quad (3)$$

with given True Positive, True Negative, False Positive, False Negative as TP,TN,FP,FN respectively-

$$\text{MSE} = \left(\sum_{i=1}^{N}(Xi - Xi')^2 / \mathbf{N}\right) \quad (4)$$

With given Xi is the actual value and Xi' is the predicted value.

A model having higher values of accuracy and lower MSE value indicate a better performing model.

**Experimental Results**

The proposed model is evaluated in terms of performance measure metrics. The results are summarized along with specified baseline classifiers such as Naïve Bayes, SVM, K-

NN, Decision Trees, Adaboost, and Random Forest Classifier. The summarized results of the baseline classifiers are shown in Table1 and Table 2 shows the overall performance of proposed neural network model followed by 10-fold cross-validation method.

**Table 1:** Performance comparison of all specified baseline classifiers

| Performance Measure Metrics | SVM | K-NN | Naïve Bayes | Decision Tree | Adaboost | Random Forest |
|---|---|---|---|---|---|---|
| Accuracy | 85.6% | 83.74% | 83.74% | 78.4% | 85.8% | 85.8% |
| MSE | 0.144 | 0.1626 | 0.16 | 0.216 | 0.14 | 0.142 |

**Table 2:** Performance of Proposed Methodology

| Performance Measure Metrics | Accuracy | MSE |
|---|---|---|
| Neural Network with Cross-Validation | 87.01% | 0.1299 |

The proposed neural network with 10-fold cross validation achieves maximum performance over all specified linear as well as ensemble classifiers. Figure 3, 4 indicates overall performance of all the classifiers with respect to accuracy and MSE. Figure 5 depicts the performance of proposed model obtained during all iterations in terms of accuracy and MSE.

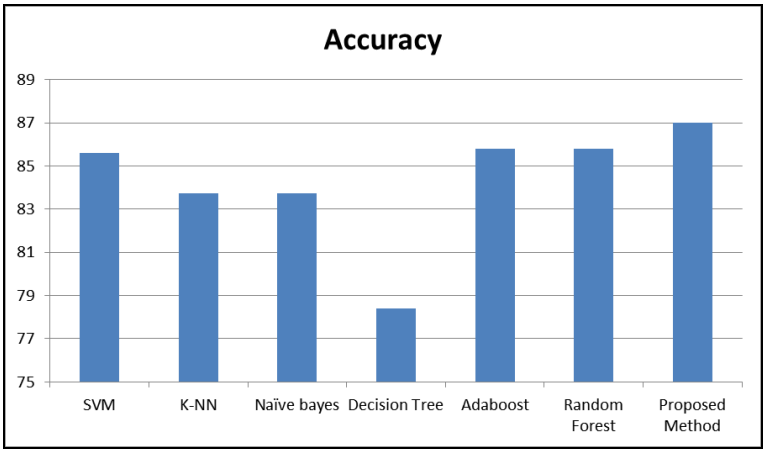



**Fig 4:** Overall Performance of the classifiers with respect to Accuracy

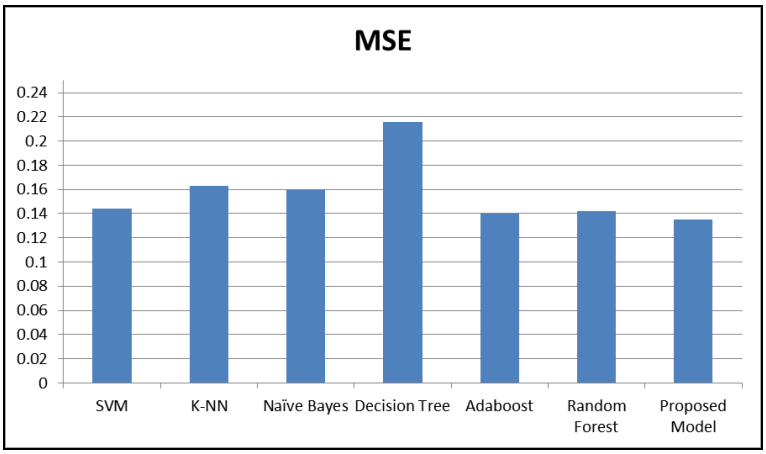



**Fig 5:** Overall Performance of the classifiers with respect to MSE

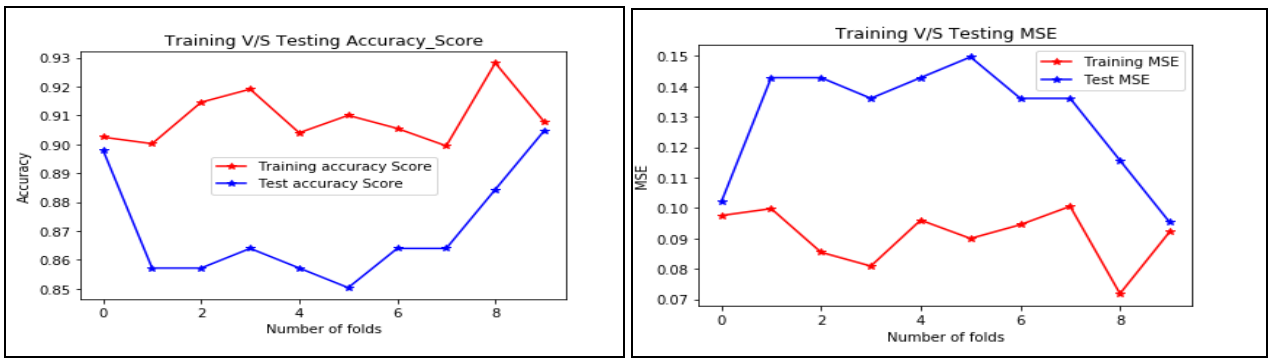



**Fig 6:** Training and Testing accuracy and MSE shown in each iteration of cross-validation.

## Conclusion

The objective of this study is to detect the feasibility of utilising related parameters and determine the probability of being affected by attrition process. Early detection of attrition may help the employee to rectify himself and put more concentration towards his/her work responsibilities. A feed-forward neural network and 10-fold cross validation procedure is provided under a single platform that can determine the attrition process beforehand. The proposed method achieves promising result with an accuracy of 87.01% and optimised MSE value as 0.1299.

## References


1. Anne J, Talapatra K, Rungta S, Jagadeesh A. Employee Attrition and Strategic Retention Challenges in Indian Manufacturing Industries : a Case Employee Attrition and Strategic Retention Challenges in Indian Manufacturing Industries : a Case Study. VSRD Int. J Bus. Manag. Res. 2016; 6:8.
2. Osuna E, Platt J. Support vector machines.
3. Rish I. An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier. 2001, 41-46.
4. Cunningham P, Delany SJ. K -Nearest Neighbour Classifiers, Mult. Classif. Syst., 2007, 1-17. Doi: 10.1016/S0031-3203(00)00099-6.
5. Sharma H, Kumar S. A Survey on Decision Tree Algorithms of Classification in Data Mining. Int. J Sci. Res. 2016; 5(4):2094-2097. doi: 10.21275/v5i4.nov162954.
6. Breiman L. Random_Forest, Mach. Learn. 2001; 45(1):5-32. doi: 10.1017/CBO9781107415324.004.
7. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting, Ann. Stat. 2000; 28(2):337-407. doi: 10.1214/aos/1016218223.
8. Yedida R, Reddy R, Vahi R, Jana R, GV A, Kulkarni D. Employee Attrition Prediction, 2018.
9. Shankar RS, Rajanikanth J, Sivaramaraju VV, Vssr Murthy K. Prediction of Employee Attrition Using Datamining, IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCA, 2018, 1-8. doi: 10.1109/ICSCAN.2018.8541242.
10. Khare R, Kaloya D, Choudhary CK, Gupta G. Employee Attrition Risk Assessment using Logistic Regression Analysis. Int. Conf. Adv. Data Anal. Bus. Anal. Intell. 2011, 1-33.
11. Yadav S, Jain A, Singh D. Early Prediction of Employee Attrition using Data Mining Techniques. Proc. 8th Int. Adv. Comput. Conf. IACC 2018; 8(2882):349-354. doi: 10.1109/IADCC.2018.8692137.
12. Mohammad Esmaieeli Sikaroudi A, Esmaieeli Sikaroudi A. A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). J Ind. Syst. Eng. 2015; 8(4):106-121.
13. Yiğit IO, Shourabizadeh H. An approach for predicting employee churn by using data mining, IDAP 2017 - Int. Artif. Intell. Data Process. Symp. 2017. doi: 10.1109/IDAP.2017.8090324.
14. Coussement K, Van den Poel D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Syst. Appl. 2008; 34(1):313-327. doi: 10.1016/j.eswa.2006.09.038.
15. Khera SN, Divya. Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. Vision, 2019; 23(1):12-21. doi: 10.1177/0972262918821221.
16. Yang J, Olafsson S. Optimization-based feature selection with adaptive instance sampling, *Comput. Oper. Res.*, 2006; 33(11):3088-3106. doi: 10.1016/j.cor.2005.01.021.
17. Pavansubhash. IBM HR Analytics Employee Attrition & Performance, Version 1, 2017. Retrieved on April 30,2020 from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
18. Liu J, Liu J, Du W, Li D. Performance analysis and characterization of training deep learning models on mobile device. Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS, 2019, 506-515. doi: 10.1109/ICPADS47876.2019.00077.
19. Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning, 2018, 1-20.
20. Schmidhuber J. Deep Learning in neural networks: An overview. Neural Networks. 2015; 61:85-117. doi: 10.1016/j.neunet.2014.09.003.
21. Kirschen RH, O'Higgins EA, Lee RT. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Am. J Orthod. Dentofac. Orthop. 2000; 118(4):456-461. doi: 10.1067/mod.2000.109032.
22. HM, SMN. A Review on Evaluation Metrics for Data Classification Evaluations, Int. J. Data Min. Knowl. Manag. Process, 2015; 5(2):01-11. doi: 10.5121/ijdkp.2015.5201.