Prediction of the Staff Exhaustion with Machine Learning Classifier

[1] K Radha , [2] K Bhanu Prakash

[1] Assistant Professorr,[2] B.TECH-IV-CSE-B11

[1] Department of Computer Science and Engineering

[1] GITAM UNIVERSITY, Rudraram, Hyderabad, Telangana, India.

Abstract: In the past years, IT industry has been going through a problem of high attrition which results in economic loss and members of staff may leave the Organizations. Main objective of this paper is to evolve a prototype to predict member of staff attenuation and to determine the organizational chances to solve any issues by improving job security. Member of Staff track record is collected from HR databases from the three different Organizations in India. In the Results, it was shown that the accuracy logistic regression algorithm is 85%. It is performing better when predicting the member of staff attrition that who are leaving the organization than who are not leaving the organization. Member of staff Attrition is considered as the Number of Situations that loses the members of staff in the Organization due to retirement and resignation, or dismissal in company's interest. The objective of this paper is to find out if there any relation in between given dataset and attrition. If there exists any such relation, how they must take actions in order to make less member of staff attrition.

Index Terms **- Attrition, Logistic Regression, Attenuation.**

I. INTRODUCTION

An Organization can encounter throughout its lifecycle stating that Revenue of Member of Staff is the most significant problem. It is tedious to predict and revoked the enterprise's skilled workforce. Due to delay in timely delivery of services, it can be a failure and entire Enterprise productivity can become less significant .Hence, whenever the members of staff leaves the Organization Unexpectedly, Customer Loyalty can be dropped [3].Consequently, it is crucial that enterprise plan genuine recruitment, accession and retention plans and implementing operative techniques to control and reduce members of staff ROI by finding the main reasons [4], [5]. Ultimately, in the area of Computer Science, Intelligent Machine Learning Algorithms are developed. Strong number of methods to get the intuitions from the Enterprises. Distinct Machine Learning Approaches are used to give the accurate Results HR datasets. Earlier, different authors proved that the performance estimation of machine learning algorithms [6], [7], [8], [9]. Along with the recent developed algorithms, Punnoose ,Ajit[8] differentiated the seven types of algorithms. To predict the staff ROI can be estimated in Sikaroudi and co-researchers [9] with ten different data mining algorithms along with the different kinds of neural networks and induction rule methods, and also focused on prediction ability and classification ability, various researchers are reported that which attribute is the most effective and to get the awareness on that attribute, that is taken to predict the member of staff ROI with respective to work experience, age and compensation. These features are gaining better understanding of their significance and holds same value in the data mining applications.

Most of the researches are computing the impurity reduction by partitioning the node in decision trees [10], [11]. In addition to this, sensitivity analysis and updated genetic algorithms have been used to understand the relative feature significance. Various Researches are created the classification rules to layout the good insight and confidence in machine learning algorithms. Along with the above mentioned research results, the findings for the prediction of member of staff turnover by using the machine learning techniques. Initially, HR data is confidential [12], but it conducts in depth analysis on various data sets. Moreover, HR data contains missing information [13], [18], inconsistent and noisy. Accuracy is not feasible for imbalanced data sets [3], [4], [9], [14]. The employees who are staying in the Organization are smaller than those who are leaving, which leads to high risk in processing the high accuracy correlations. Due to accurate results in analysis, and often opting for relatively instead. With the proposed technique most affected feature results significance can be achieved. The analysis of feature significance in various researches [7],[11],[3],[16] ,algorithm performance can be predicted and improved the results. For example, some of the researches are using the decision trees [11], [16], to compute feature significance as part of the model building process. Anyhow, feature significance gives less accurate output because of the in-efficiency of decision tree classifier. To improve the decision trees performance by displaying their classification rules to improve the model. But decision trees results in low stability and high variance which results in less amount of modification in the data [17]. Objective of this paper is to layout the estimation and to show the supervised machine learning approaches and to predict the Member of Staff ROI. In the Current Research, For the small, medium and large scale industries, various supervised machine learning approaches are used. These algorithms will be give and performance and benefits of staff turnover. With the larger data sets also supervised machine learning techniques are authentic exploration for Human Resource datasets. In this , Numeral results for reproducing HR Data Sets representing the Large –Scale Industries suffering due to large size of staff members, it is experimented on , Gradient Boosting Trees , Support Vector Machine technique, Logistic Regression Analysis Decision Tree technique, Random Forest, ,LDA(Linear Discriminate Analysis ,Naïve Bayes Method ,Neural Networks, K-Nearest Neighbour Technique.

II. METHODOLOGY

In this Research, To Assess and Predict the Member of Staff Turnover, distinct Supervised Machine Learning Algorithms are used.

2.1 Logistic Regression

Logistic Regression is proposed in the year of 1958 by the Cox [18] to divide the input data into two parts.

2.2 Problem Statement

The dataset describe the member of staff attrition. Attrition is referred as a staff member is leaving the Enterprise on Various Reasons. The Reasons are might be personal, but it might impact Organizations that they may be terminated from their duties. logistic regression is used to gain the knowledge out of data.

III. EXPERIMENTAL DESIGN

3.1 Evaluation Metrics

In Staff Member Turnover analysis, members of staffs who left the Organization and who are staying can be taken into consideration and also imbalance of the individuals are considered. The staff member Turnover is always less than the 0.50, to make the more accuracy. To solve this problem, additional performance attributes are used. Distinct Performance attributes are used to estimate the performance, such as TP,FP,F1-score,ROC,Precison ,Recall, and Accuracy [20].

IV. MODEL BUILDING AND VALIDATION

On a single data set to compute the capability of technique, cross validation is used. Due to high complexity of the technique, cross validation is preventing the model from over fitting. It predicts with the test dataset within a predefined range [20], Grid Search can be applied. Once the most appropriate attribute is recognized, Accuracy, ROC_Curve, AUC_Score, Confusion Matrix, F1 and Recall, were computed with the 10-cross fold validation.

V. DATA COLLECTION

To gather and measure the information from a distinct sources Data Collection technique is used to get accurate results. The given dataset gives details about the corporate data sets which can affect the staff member attrition. The data was collected from unknown data sources which are related to anonymous enterprise.

The flow of this work is as follows

- Importing the data into python.
- Doing Exploratory data analysis
- By using various methods of visualizations
- If the given data consists of any missing values they must be filled with some strategies.
- The above step is known as missing value treatment.
- Mostly any data is non normal, thereby we must apply some normalization techniques and then we can use then data for further processes.
- The data normalization reduces variance in the data.
- Then apply the Logistic Regression on the data by splitting the data.
- There by we prepare a various kind of models in the whole modelling. So, we later on go on to choose the models with good accuracy and other metrics.

VI. EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis (EDA) is viewing the data is different dimensional views and the data analysis can be done by preparing some of the predefined models. This EDA has gained some of the insights staff attrition dataset, which is about the factors are effecting the staff member attrition among the training and test datasets. Staff Member Attrition Data Set contains the following Features such as Attrition, Age of the Following attributes are considered as staff member, Department they are working, business travelling, how much distance from the home, educational qualifications, Interested area, Education Field, Staff count, staff member satifaction,staff member number,Gender,Job Role, Job Involvement ,Number of years in the current designation, Number of years worked at Company, Standard Hours, Relationship satisfaction, work life balance, number of years

with the Current Manager Role, number of years with the last promotion, monthly income, marital status, salary hike, over time.

Table- 1: Employee Attrition towards their Age

	Age	Attrition
Age	1.000000	-0.159205
Attrition	-0.159205	1.000000
DailyRate	0.010661	-0.056652
DistanceFromHome	-0.001686	0.077924
Education	0.208034	-0.031373
EmployeeNumber	-0.010145	-0.010577
EnvironmentSatisfaction	0.010146	-0.103369
HourlyRate	0.024287	-0.006846
JobInvolvement	0.029820	-0.130016
JobLevel	0.509604	-0.169105
JobSatisfaction	-0.004892	-0.103481
MonthlyIncome	0.497855	-0.159840
MonthlyRate	0.028051	0.015170
NumCompaniesWorked	0.299635	0.043494
PercentSalaryHike	0.003634	-0.013478
PerformanceRating	0.001904	0.002889
RelationshipSatisfaction	0.053535	-0.045872
StockOptionLevel	0.037510	-0.137145

Table- 2: Years Spent in the Organization

TotalWorkingYears	0.680381	-0.171063
TrainingTimesLastYear	-0.019621	-0.059478
WorkLifeBalance	-0.021490	-0.063939
YearsAtCompany	0.311309	-0.134392
YearsInCurrentRole	0.212901	-0.160545
YearsSinceLastPromotion	0.216513	-0.033019
YearsWithCurrManager	0.202089	-0.156199

There is presence of categorical data columns, in order to apply Logistic Regression; we have to convert the data into numerical format. And there are no missing values in the data. And out of all those above-mentioned columns we have built various kinds of visualization charts in order to understand the given data set and then moving on we have got down to only 15 columns out of all the input columns. The columns that we choose after considering correlation values and by using visualization techniques like boxplots, distplots etc we have tried to understand the variance in data and the behaviors the different columns are exhibiting.

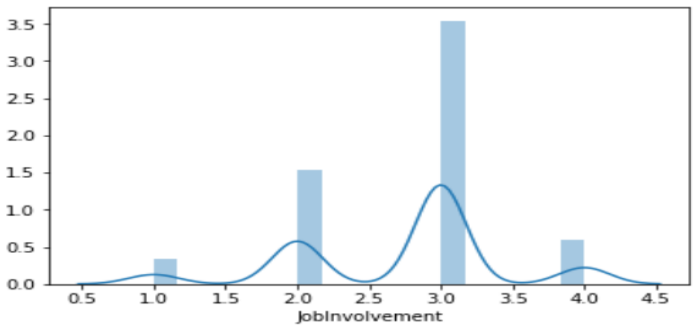


Fig.1 Correlation values of attrition with other inputs

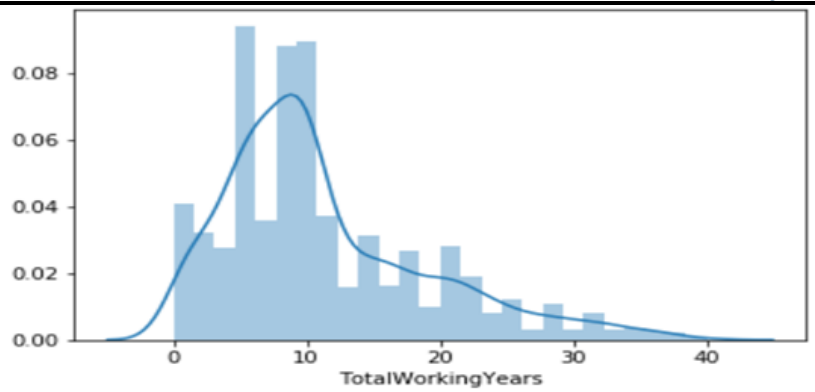


Fig.2. Distplots for few Selected Attribute

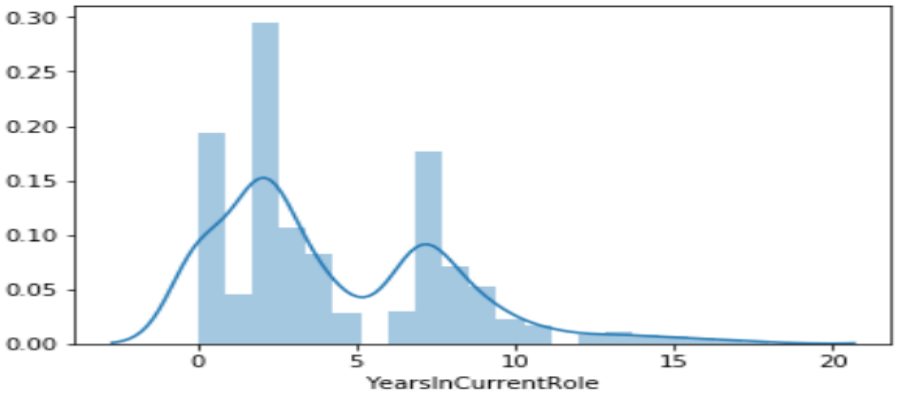


Fig.3. Boxplots for few selected input

```
Emp_data.boxplot(column='EnvironmentSatisfaction')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1df34baa710>
```

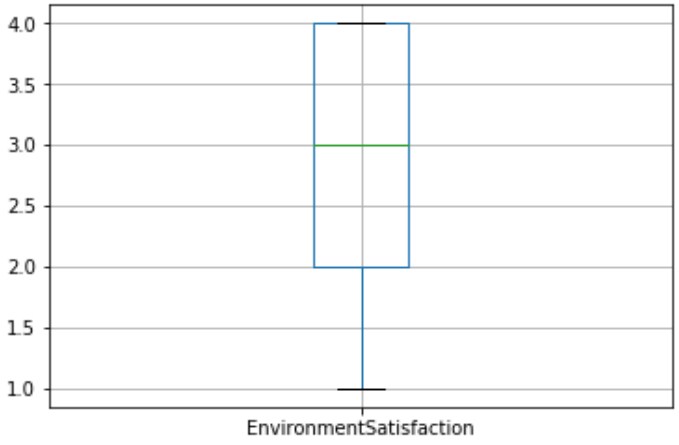


Fig.4 Staff Satisfaction

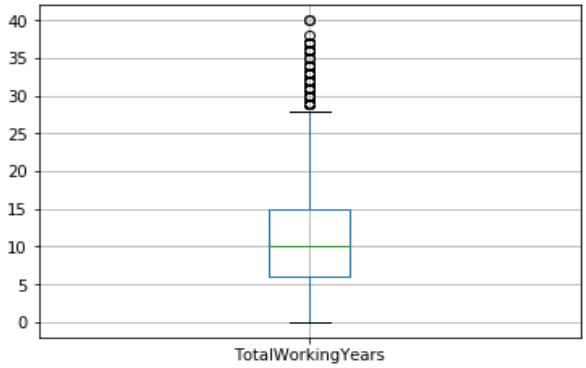


Fig.5. Staff Total Working Years

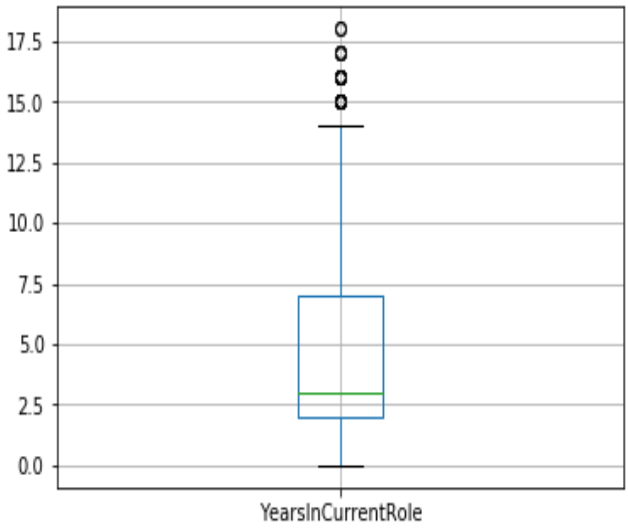


Fig.6. Number of Years spent in the Current Role

6.1 Data Normalization

Data Normalization can be applied for Data Pre-processing in machine Learning. The objective of this technique is to modify the values of numerical variables in the dataset to common scale irrespective of range values. In Machine learning Normalization is not required but it requires only when the attributes have different ranges. For Instance, Assume a dataset consists of two attributes such as income and age (x1, x2) where the age ranges can be taken as 0-100 years and income ranges can be considered as 0-20000 Rupees. Income is 1000 times larger than the age and ranges from 20,000-500,000. Hence these attributes are to be taken in different ranges. For the Further analysis, multivariate linear regression can be applied. We will apply the min-max normalization technique can be applied for the numerical data.

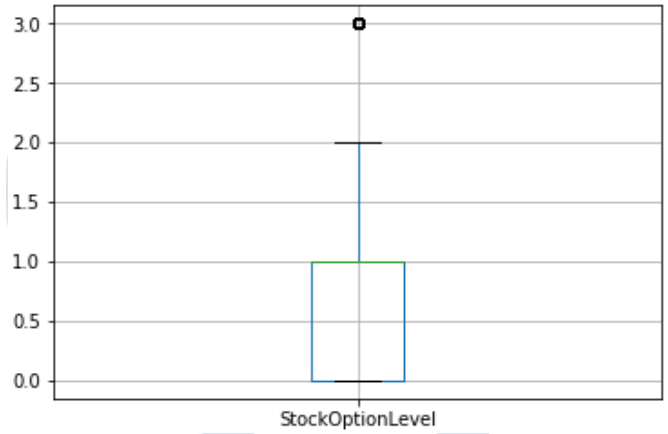


Fig.7. Stock-Option Level of the Staff Member

```
# Normalize the numeric variables
from sklearn import preprocessing
minmax=preprocessing.MinMaxScaler(feature_range=(0,1))
minmax.fit(X).transform(X)
```

```
C:\Users\rry59\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py:334: DataConversionWarning: Data with input dtype i
nt64 were all converted to float64 by MinMaxScaler.
  return self.partial_fit(X, y)
```

```
array([[0.54761905, 0.        , 0.71581961, ..., 0.22222222, 0.        ,
        0.29411765],
       [0.73809524, 0.5       , 0.12670007, ..., 0.38888889, 0.06666667,
        0.41176471],
       [0.45238095, 0.        , 0.9090673 , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.21428571, 0.        , 0.03793844, ..., 0.11111111, 0.        ,
        0.17647059],
       [0.73809524, 0.5       , 0.65926986, ..., 0.33333333, 0.        ,
        0.47058824],
       [0.38095238, 0.        , 0.37652112, ..., 0.16666667, 0.06666667,
        0.11764706]])
```

VII. LOGISTIC REGRESSION

Logistic Regression is most familiar technique. Linear Regression and Logistic Regression techniques are identical. Logistic Regression is used for performing the classification tasks and the Linear Regression Analysis can be done to predict the values. Logistic regression is made up on the basis of mean and it uses standard deviation rules. We have created the three folds to get the accuracy with the help of Train and test validations. With these, the accuracy of three folds is computed by cross validation.

```
lr_acc
```

```
array([0.86627907, 0.86588921, 0.86842105])
```

VIII. RESULT ANALYSIS WITH OTHER METRICS

From sklearn we have imported accuracy_score, roc_auc_score, confusion_matrix.

- The accuracy score talks about accuracy of model on test data
- The Recall score speaks about the true positive rate of the model
- The roc_auc_score is about the sensitivity and the specificity.

```
#Accuracy score on Test and Train
from sklearn.metrics import accuracy_score,recall_score,roc_auc_score,confusion_matrix
print("\nAccuracy score:%f" %(accuracy_score(y_test,y_predict)*100))
print("Recall score:%f" %(recall_score(y_test,y_predict)*100))
print("ROC score:%f" %(roc_auc_score(y_test,y_predict)*100))
print(confusion_matrix(y_test,y_predict))
```

```
Accuracy score:87.074830
Recall score:25.714286
ROC score:62.183288
[[366   5]
 [ 52  18]]
```

8.1 ROC Curve

Generally, the ROC curves of the model explain about the true and the false positive rate. If the value is close to 0 then model is said to be poor. If the value is close 0.5 model is said to be average. If the value is closer to 1 then model is performing well.

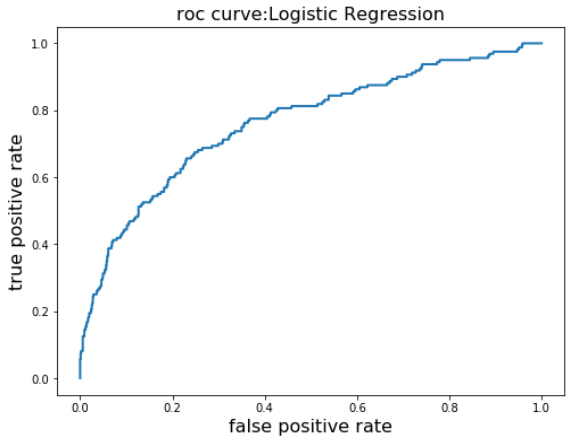
8.2. Models:

Single variant logistic regression accuracy, test size=0.3

Number of years worked in the present designation = 86.84
Number of years worked with current manager = 83.67
Environment satisfaction = 84.12
Job involvement = 82.31
Job satisfaction = 85.03
Total years' work = 81.85
Overtime =82.53
Stock option level = 81.52

```
roc_cur("Logistic Regression",y_train,lr1_scores)
```

```
AUC curve (Logistic Regression):0.76
```



Multi variant logistic regression accuracy, test size=0.3

environment satisfaction, job involvement = 82.53.Number of years worked in the present designation manager= 80.04 job satisfaction ,Number of years worked in the present designation = 82.76,the employee is worked for over time, stock option = 82.08 ,Number of years worked in the present designation manager =83.67 job satisfaction, over time, stock option level =83.44 years with current manager=84.58 ,overtime =83.4 , Number of years worked with manager designation = 84.35 years with current manager = 84.35and finally we have achieved the Complete data set accuracy = 85.94

Single variant logistic regression accuracy, test size = 0.2

Company Atomsphere satisfaction = 84.35
Job involvement = 79.93
Job satisfaction = 84.01
Stock option level= 84.35
Over time = 82.65
Years in current role =84.35
Years with current manager = 83.33
Total work years=81.29

Multi variant logistic regression accuracy, test size = 0.2

job satisfaction, Number of years in the present designation = 81.97 environment satisfaction, job involvement = 80.04 total number of working years with the current manager= 86.39 over time , stock option = 82.08 stock option level ,total working years ,years in current role ,Number of years with current manager designation =79.93 and job satisfaction, over time, total working years, stock option level =86.73. environmental satisfaction, job involvement, years in current role, years with current manager=81.29 environment satisfaction, job involvement, job satisfaction, overtime, stock option level, total working years, years in current role, years with current manager = 82.6 Department, distance from home, education field, environment satisfaction, job involvement, job satisfaction, number of companies worked, overtime, percent salary hike, stock option level, total working years, years in current role, years with current manager = 86.39
Complete data set accuracy = 85.71.

Best Performing models that we have chosen:
Single Variant Model Accuracy:

Job satisfaction ,test size(0.3)= 85.03
Stock option level, test size(0.2)= 84.35
Environment Satisfaction , test size(0.2) = 84.35

Years in current role , test size(0.3) = 86..84

IX. FINDINGS

There are few unwanted or un-influential columns present in the data

- The Years is current role has played crucial role in employee attrition according to our models .
- The data looks like it is noisy because of its outliers and the non normal form .
- In consideration to stat models and correlation data , we have selected 8 critical inputs , they are environment satisfaction , job involvement , years in current role , stock option level , over time , years with current manager
- All these inputs have effect on attrition, there I would predict that staff member attrition is mainly because of the staff personal reasons rather than company dismissing employee.
- In such case, I would suggest to improve few conditions and situations in the company based on total working years, years with current manager, test size (0.2) = 86.39 job satisfaction, years in current role = 82.76 job satisfaction, over time, total working years, stock option level =86.73 environmental satisfaction, job involvement, years in current role, years with current manager=84.58 current role, years with current manager = 84.35. These inputs will be effected and result in changing attrition.

X. CONCLUSION

An Organization can encounter throughout its life cycle stating that Revenue of staff attenuation is the most significant problem. It is tedious to predict and revoked the enterprise's skilled workforce. Due to delay in timely delivery of services, it can be a failure and entire Enterprise productivity can become less significant. Staff Member Attrition is considered as the Number of Situations

that loses the staff members in the Organization due to retirement and resignation, or dismissal in company's interest. The objective of this paper is to find out if there any relation in between given columns and attrition. If there exists any such relation how they must take actions in order to make less Staff Member attrition.

REFERENCES

[1] Shikha N. Khera,"Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques", Volume: 23 issue: 1, page(s): 12-21,March 5, 2019; Issue published: March 1, 2019 .
[2] Yue Zhao,et al, "Employee Turnover Prediction with Machine Learning: A Reliable Approach", Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2, January 2019.
[3] Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M.: Employee turnover: a neural network solution. Comput. Oper. Res. 32, 2635–2651 (2005).
[4] Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employees performance. Int. J. Adv. Comput. Sci. Appl. 3, 144–151 (2012).
[5] Li, Y.M., Lai, C.Y., Kao, C.P.: Building a qualitative recruitment system via SVM with MCDM approach. Appl. Intell. 35, 75–88 (2011).
[6] Nagadevara, V., Srinivasan, V., Valk, R.: Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. Res. Pract. Hum. Resour. Manag. 16, 81–97 (2008).
[7] Suceendran, K., Saravanan, R., Divya Ananthram, D.S., Kumar, R.K., Sarukesi, K.: Applying classifier algorithms to organizational memory to build an attrition predictor model.
[8] Punnoose, R., Ajit, P.: Prediction of employee turnover in organizations using machine learning algorithms. Int. J. Adv. Res. Artif. Intell. 5, 22–26 (2016)
[9] Sikaroudi, E., Mohammad, A., Ghousi, R., Sikaroudi, A.: A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). J. Ind. Syst. Eng. 8, 106–121 (2015)
[10] Jantan, H., Hamdan, A.R., Othman, Z.A.: Human talent prediction in HRM using C4.5 classification algorithm. Int. J. Comput. Sci. Eng. 2, 2526–2534 (2010).
[11] Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms.
[12] Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013).
[13] Quinn, A., Rycraft, J.R., Schoech, D.: Building a model to predict caseworker and supervisor turnover using a neural network and logistic regression. J. Technol. Hum. Serv.19, 65–85 (2002).
[14] Chien, C.F., Chen, L.F.: Data mining to improve personnel selection and enhance human
[15] capital: a case study in high-technology industry. Expert Syst. Appl. 34, 280–290 (2008).
[16] Tzeng, H.M., Hsieh, J.G., Lin, Y.L.: Predicting nurses' intention to quit with a support vector machine: a new approach to set up an early warning mechanism in human resource management. CIN: Comput. Inf. Nurs. 22, 232–242 (2004).
[17] Chang, H.Y.: Employee turnover: a novel prediction solution with effective feature selection.WSEAS Trans. Inf. Sci. Appl. 6, 417–426 (2009).
[18] Jantan, H., Hamdan, A.R., Othman, Z.A.: Human talent prediction in HRM using C4.5 classification algorithm. Int. J. Comput. Sci. Eng. 2, 2526–2534 (2010).
[19] Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer, New York (2001).
[20] Cox, D.R.: The regression analysis of binary sequences. J. Roy. Stat. Soc. B. Met., 215–242 (1958).
[21] Raschka, S.: Python Machine Learning. Packt Publishing Ltd, Birmingham (2015).