## A PROJECT REPORT ON,

**“Employee Turnover Prediction”**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

**SUBMITTED BY**

|  |  |
| --- | --- |
| Naik Ayush Ashutosh | Roll No: T5487 (Div 2) |
| Rakesh Mali | Roll No: T5522 (Div 2) |
| Rote Mitali Ravindra | Roll No: T5526 (Div 2) |
|  |  |



# DEPARTMENT OF COMPUTER ENGINEERING

## NBN SINHGAD SCHOOL OF ENGINEERING, PUNE-41

**SAVITRIBAI PHULE PUNE UNIVERSITY**

**2023 - 24**



# CERTIFICATE

This is to certify that the project report entitles

**“Employee Turnover Prediction”**

Submitted by

**Naik Ayush Ashutosh TE Div 2 Roll No: T5487**

**Rakesh Mali TE Div 2 Roll No: T5522**

**Rote Mitali Ravindra TE Div 2 Roll No: T5526**

is a bonafide work carried out by him under the supervision of **Prof. Snehal Rathod** and it is approved for the partial fulfillment of the requirement of University of Pune as a part of Laboratory Practice III work syllabus (Third year Computer Engineering).

|  |  |
| --- | --- |
| **(Prof. Snehal Rathod)** | **(Prof. Shailesh Bendale)** |
| Subject In-charge | Head, Department of Computer Engineering |

**ACKNOWLEDGEMENT**

It is indeed a great pleasure and moment of immense satisfaction for we to present a project report on “**Employee Turnover Prediction** " amongst a wide panorama that provided us inspiring guidance and encouragement, we take the opportunity to thanks those who gave us the indebted assistance. We wish to extend our cordial gratitude with profound thanks to our internal guide **Prof. Snehal Rathod** for her everlasting guidance. It was his inspiration and encouragement which helped us in completing our project.

Our sincere thanks and deep gratitude to Head of Department, **Prof. Shailesh Bendale** and other faculty member; but also, to all those individuals involved both directly and indirectly for their help in all aspect of the project.

At last, but not least we express our sincere thanks to our Institute’s Principal **Dr. Shivprasad Patil**, for providing us infrastructure and technical environment.

**Naik Ayush Ashutosh**

**Rakesh Mali**

**Rote Mitali Ravindra**

## ABSTRACT

Supervised machine learning methods are described, demonstrated and assessed for the prediction of employee turnover within an organization. In this study, numerical experiments for real and simulated human resources datasets representing organizations of small-, medium- and large-sized employee populations are performed using (1) a decision tree method; (2) a random forest method; (3) a gradient boosting trees method; (4) an extreme gradient boosting method; (5) a logistic regression method; (6) support vector machines; (7) neural networks; (8) linear discriminant analysis; (9) a Naïve Bayes method; and (10) a K-nearest neighbor method. Through a robust and comprehensive evaluation process, the performance of each of these supervised machine learning methods for predicting employee turnover is analyzed and established using statistical methods. Additionally, reliable guidelines are provided on the selection, use and interpretation of these methods for the analysis of human resources datasets of varying size and complexity

## CONTENTS

|  |  |  |  |
| --- | --- | --- | --- |
| **CHAPTER** | | **TITLE** | **PAGE NO** |
| **1** | | **Article I. Problem Statement Definition** | **6** |
| **2** | | **Article II. Software Requirements Specification** | **6** |
|  | **2.1** | **Introduction** | **6** |
|  | **2.2** | **Scope** | **6** |
|  | **2.3** | **Requirements** | **7** |
| **3** | | **System Implementation** | **8** |
|  | **3.1** | **Probability and Statistical Analysis** | **8** |
|  | **3.2** | **Model Building and Validation** | **8** |
| **4** | | **Project Snapshots** | **9** |
| **5** | | **Article III. Conclusion** | **15** |
| **6** | | **Article IV. References** | **16** |

1. **Problem Statement Definition**

Employee Turnover is the measurement of the total number of employees who leave an organization in a particular year. Employee Turnover Prediction means to predict whether an employee is going to leave the organization in the coming period.

A Company uses this predictive analysis to measure how many employees they will need if the potential employees will leave their organization. A company also uses this predictive analysis to make the workplace better for employees by understanding the core reasons for the high turnover ratio.

## Software Requirements and Specification

* 1. **Introduction**

Employee turnover is one of the most significant problems an organization can encounter throughout its lifecycle, as it is difficult to predict and often introduces noticeable voids in an organization’s skilled workforce [1]. Service firms recognize that the timely delivery of their services can become compromised, overall firm productivity can decrease significantly and, consequently, customer loyalty can decline when employees leave unexpectedly. As a result, it is imperative that organizations formulate proper recruitment, acquisition and retention strategies and implement effective mechanisms to prevent and diminish employee turnover, while understanding its underlying, root causes [2, 3].

## Scope

The aim of this project is to provide a comprehensive description, demonstration and assessment of supervised machine learning approaches for the prediction of employee turnover within organizations of varying size. In the present study, ten supervised machine learning methods are evaluated for organizations of small-, medium- and large-sized populations. Details of each supervised machine learning method are given and the benefits, capabilities and performance of each are provided in the context of predicting employee turnover. The effect of data size and data type, and how to get reliable feature importance and data visualization are also discussed. Lastly, general guidelines are provided on the selection, use and interpretation of these ten supervised machine learning methods for reliable analysis of HR datasets of varying size and complexity.

## Requirements

Software Requirements are:

* + - Windows 7 or higher (Server side)
    - Any OS Platform for the attacker
    - Any Web Browser (For E.g. Google Chrome, Mozilla Firefox, Internet Explorer)
    - MySQL Database (On the server side for SQL injection attack)
    - Jupiter Notebook
    - Libraries – numpy, pandas, matplotlib, seaborn.

## System Implementation

### Probability and Statistical Analysis

Non-parametric *Kruskal-Wallis* tests followed *by Dunn’s post-hoc* test was used to conduct multi-group comparison on classifier performances (e.g. data type, size and model selection). The Mann-Whitney U test was used to conduct pairwise comparisons between two groups. For these tests, the probability P<0:05 was considered significant while the remainder was considered non-significant (NS).

Probability and information theory methods were also used in this study to analyze data characteristics. In general, mutual information (MI) measures how much uncertainty is reduced about random variable (RV) Y after X is observed. MI between X and Y, II (X|Y); is given as follows, where p is the probability:

II (X|Y) X x X y p xð Þ; y log p xð Þ ; y p xð Þp yð Þ ð1Þ

In this study, features include both discrete and continuous RVs. However, MI is only feasible for a pair of discrete RVs, rather than continuous RVs. Therefore, the maximal information coefficient (MIC) was introduced to quantify the linear and nonlinear correlation between features and the predicted value. MIC could measure the MI between continuous and discrete RVs, ranging from 0 (no correlation) to 1 (fully correlated). The function m (x; y) is defined as the approximately maximized **MI** with various bin sizes and locations while discretizing a continuous random variable:

Where, ℊ (x, y) is the set of 2-dimensional grids with size x \* y. II(X (G)| Y (G)) is MI (Eq. 1) enumerated on ℊ(x, y). MIC is then given as:

MIC, maxx;y;xy<BmðÞ ð x; y 3Þ

Where, x and y are two RVs and B is a sample size dependent bound.

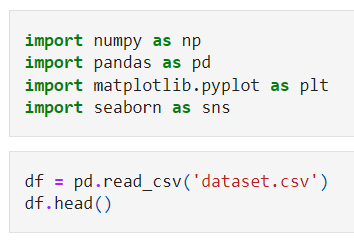
### Model Building and Validation

Cross validation is used to assess the generalization ability of an algorithm on an independent dataset. It can prevent a model from overfitting that is possibly caused by 746 Y. Zhao et al. the high complexity of the model. Grid search is a parameter searching algorithm that is used to automatically find the most optimal parameters within a predefined range. All of the datasets listed in Table 2 were run against the ten algorithms introduced in Sect. 2 with data preprocessing methods. In total, there were 10 datasets, 10 algorithms and 3 data formats (raw, normalized, standardized), yielding a total of 300 numerical experiments performed in this research. For each numerical experiment, the optimal algorithm parameters were defined by the Grid Search technique within a predefined range using Grid Search CV package. Once the optimal parameter was found, the accuracy, precision, recall, F1 and ROC values were calculated using 10- fold cross validation.

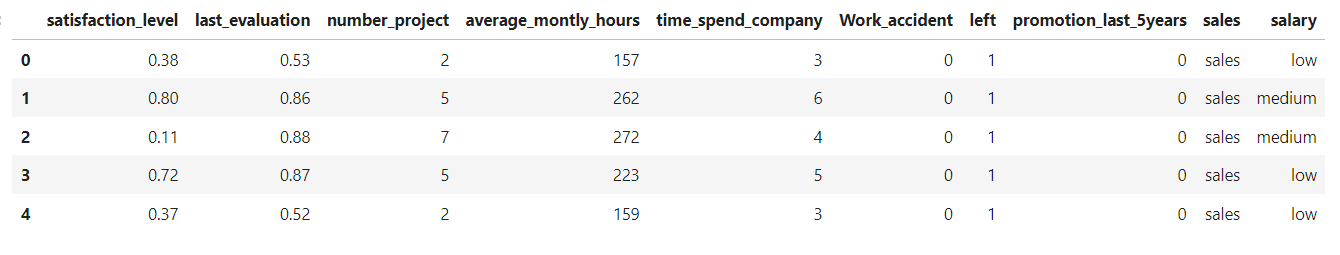
## Project Snapshots

### Import Library and dataset

1. Import Libraries and adding dataset for Data Processing

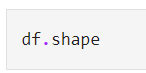


1. Output after adding data

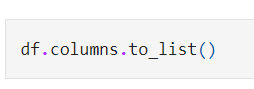


* 1. **Data Processing**

**4.2.1 Data info**

****

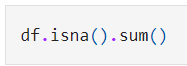
**4.2.2 Column List**

****

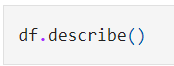
**4.2.3 Dataframe Info**

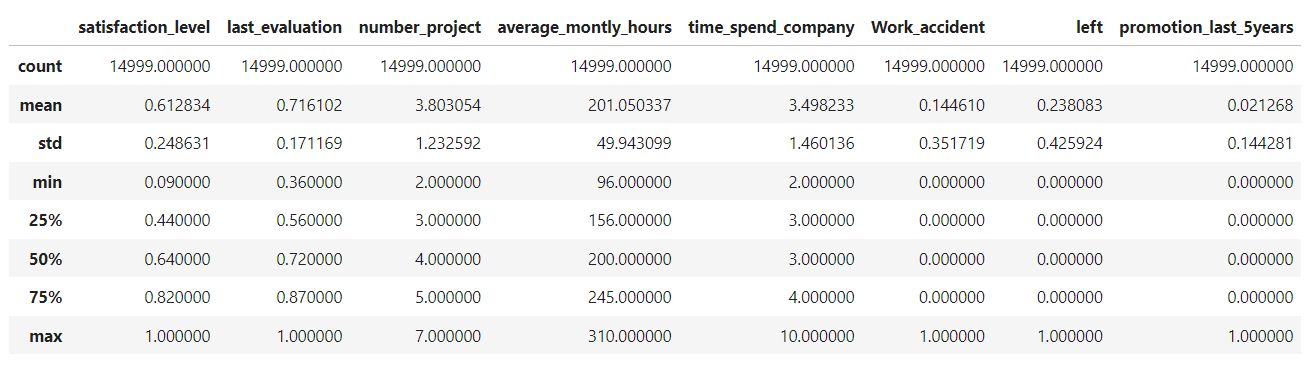
****

**4.2.4 Missing Value Handling**

****

**4.2.5 Data Statistics**

****

****

* 1. **Categories to Numerical – Feature Transformation**

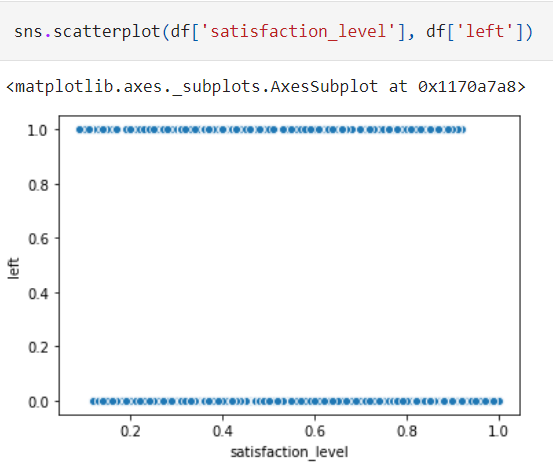
**4.3.1 Label Encoder**

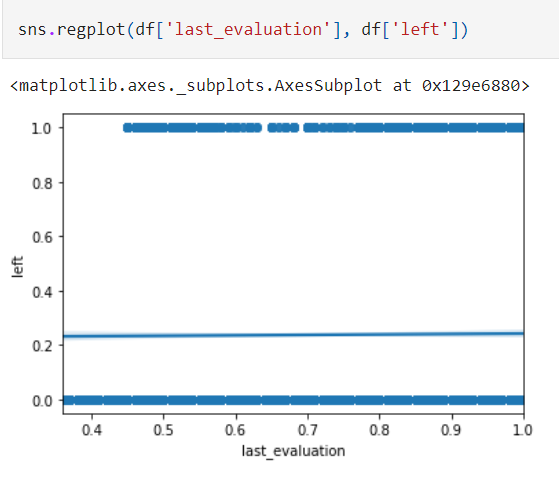
****

****

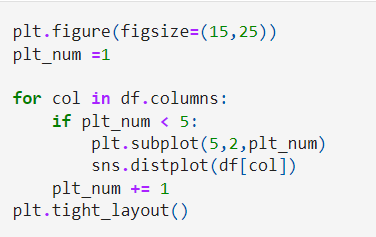
* 1. **Data Visualization**

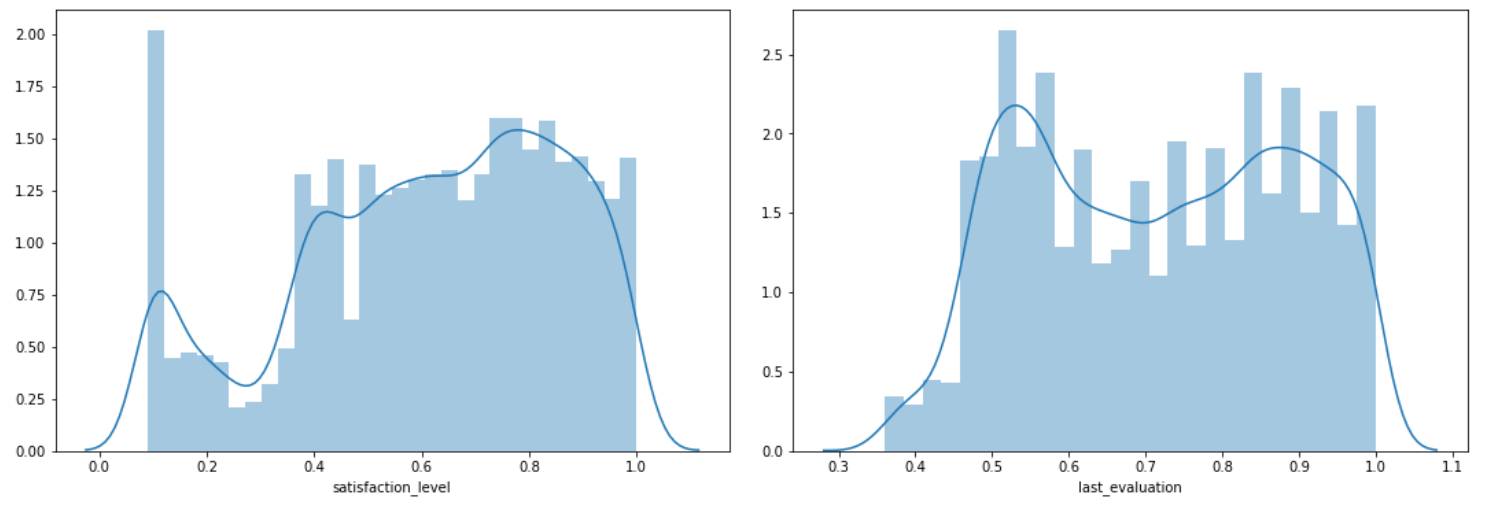
**4.4.1 Get Insights**

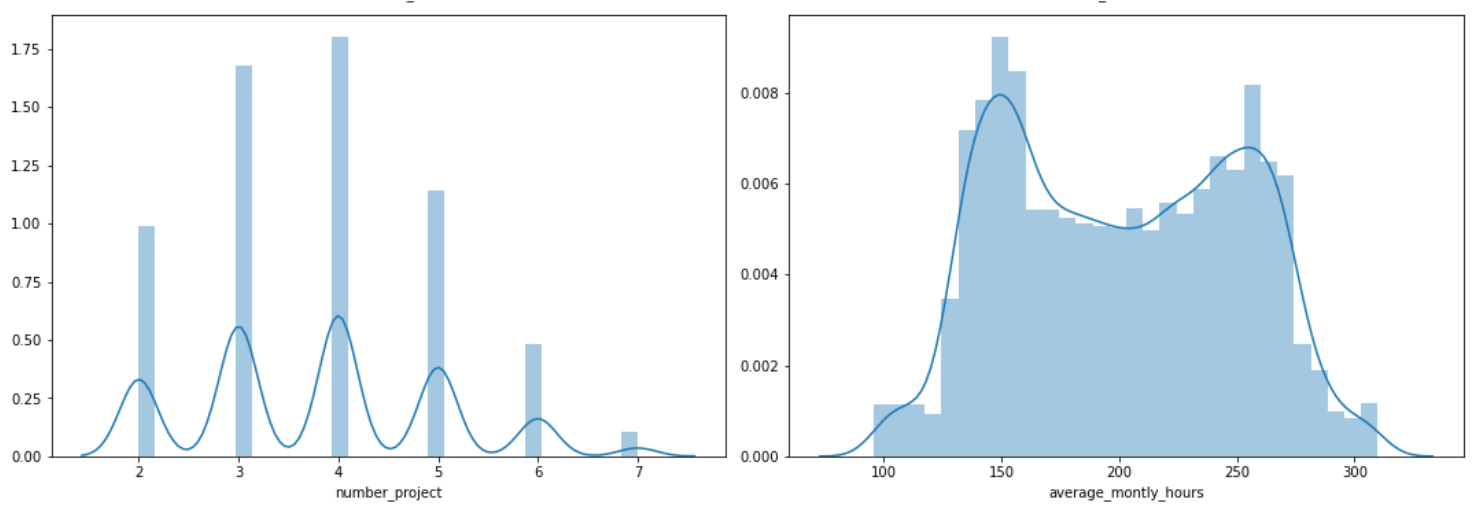


****

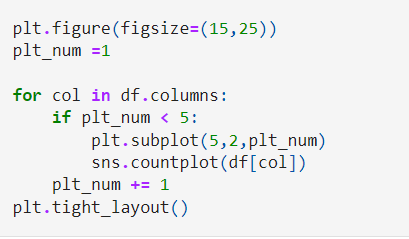
**4.4.2 Distribuion**

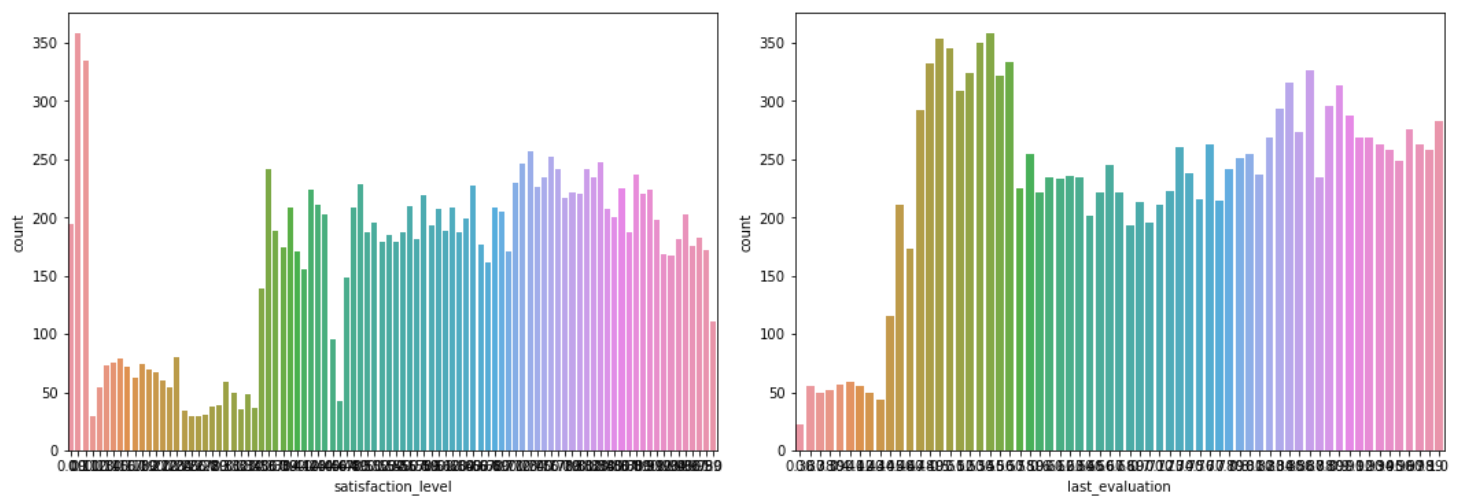
****

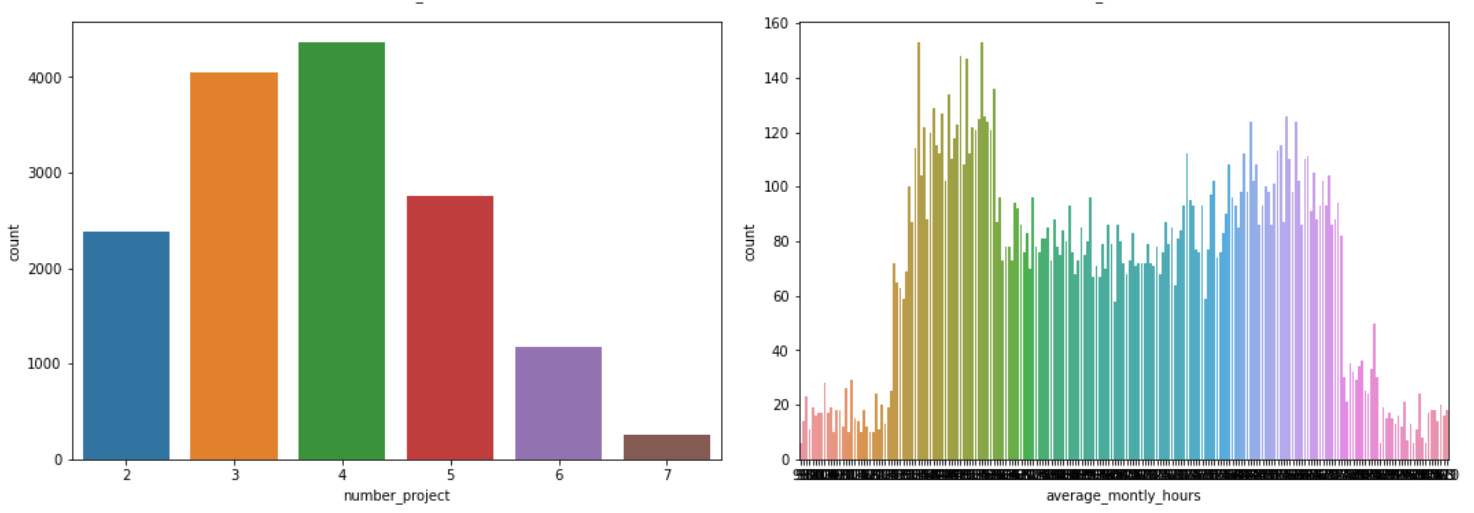
****

****

**4.2.3 Outlier Handling**

****

****

****

## Conclusion

Employee turnover has been identified as a pivotal factor to curb the growth of organizations. In this research, the performance of ten supervised machine learning methods was evaluated on various HR datasets. In addition to statistical analysis, a number of data mining techniques were introduced and used in this study, including data scaling, parameter searching and cross validation. To enhance the interpretability of employee turnover model, the examples of feature importance ranking and classifier visualization, and suggestions on how to use them appropriately.

The numerical experiment results indicate that for small HR datasets, the key is to try different algorithms as Hughes phenomenon may result in overoptimistic results. If there are more HR datasets available, extreme gradient boosting is recommended to use as the most reliable algorithm. It requires minimal data preprocessing, has decent predictive power, and ranks the feature importance automatically and reliably. However, due to the complexity of employee turnover prediction, one should try to find the classifier that best fits the underlying data before taking this approach.

## References

* 1. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013).
  2. Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employee’s performance. Int. J. Adv. Comput. Sci. Appl. 3, 144–151 (2012)
  3. Chang, H.Y.: Employee turnover: a novel prediction solution with effective feature selection. WSEAS Trans. Inf. Sci. Appl. 6, 417–426 (2009)
  4. Chien, C.F., Chen, L.F.: Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. Expert Syst. Appl. 34, 280–290 (2008)
  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. 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)
  8. Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M.: Employee turnover: a neural network solution. Comput. Oper. Res. 32, 2635–2651 (2005)
  9. 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
  10. 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)