IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Special Issue 1, March 2024

1st International Conference on Machine Learning, Optimization and Data Science

Organized by

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, India

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

Attrition Analytics: Unveiling the Best Model for Predicting Employee Retention

Divya Pandey[1], Zeba Vishwakarma[2], Mallika Dwivedi[3]

Assistant Professor, Department of CSE, BGIEM, Jabalpur, MP, India[1, 2, 3]

ABSTRACT: Organizations in all sectors are very concerned about employee attrition as it affects output, morale, and general performance. Keeping an even and productive workforce requires anticipating and controlling attrition. In this work, we investigate how well different machine learning models predict attrition among employees. We evaluate the effectiveness of Decision Tree, Random Forest, Logistic Regression, and Naive Bayes classifiers on a dataset that includes performance, job-related, and demographic characteristics of employees. GridSearchCV and other hyper parameter tweaking approaches are used to maximize model performance. Our findings provide important new information on how well various machine learning methods predict attrition. The study's conclusions extend the field of attrition analytics and offer insightful advice to businesses looking to improve retention tactics and reduce employee attrition.

KEYWORDS: employee attrition; employee turnover; machine learning; attrition rate, data analytics, data visualization

I. INTRODUCTION

For firms, employee turnover is a major problem that frequently results in interrupted operations and lower output. To lessen the effects of turnover, firms must create efficient plans for hiring, retaining, and comprehending the fundamental reasons behind it [1]. Turnover rates, which are determined by counting the number of workers who leave a company over a certain time period, are influenced by a range of employee attrition factors, including internal, external, and voluntary attrition [2].

Statistics showing that a sizable percentage of new hires quit within the first six months of work [3] demonstrate the severity of employee turnover. An attrition rate of 57.3% is a result of millions of workers quitting their employment each month in the United States alone [4].

Although a 90% retention rate is optimal, attrition rates in several businesses are as low as 19% [5]. Hiring new staff comes at a high expense; the typical hire costs USD 4129 [5]. A subset of artificial intelligence called machine learning is essential for solving problems with employee attrition prediction. Machine learning models are able to understand patterns and make predictions by using past data, which allows them to outperform humans in decision-making [6]. Applications of machine learning may be found in many different fields, such as speech recognition, picture identification, traffic prediction, and staff attrition prediction [6–10].

II. RELATED WORK

The performance evaluation of machine learning algorithms has also been studied previously by various researchers [11, 12, 13, 14]. Notably, Punnoose and Ajit [13] compared the predictive capabilities of seven different machine learning algorithms, including recently developed algorithms, like Extreme Gradient Boosting [15], on employee turnover. Similarly, Sikaroudi and co-researchers [14] conducted simulations to predict employee turnover using ten different data mining algorithms, including tests on various types of neural networks and induction rule methods. In addition to placing focus on classification and prediction ability, many researchers have also made substantial efforts to better understand which features (e.g. compensation, age, work experience, etc.) are most influential in predicting employee turnover [1–4, 8, 9, 14]. These features seldom carry equal value in data mining applications, so it is useful to gain a better understanding of their importance [16]. For instance, many of the studies using tree-based quantified feature importance by calculating the impurity reduction by node split in decision trees [1, 17]. Moreover, modified

genetic algorithms [8] and sensitivity analysis [6] have been used to understand relative feature importance as well. Numerous studies have also generated classification rules or visualized the classification procedure to provide further insight and confidence in using machine learning methods [2, 6, 17]. Despite the breadth of research outcomes mentioned above, the findings for predicting employee turnover that stem from using machine learning methods are often problem-specific and difficult to generalize. First and foremost, this is primarily because HR data is confidential [7], which inherently impedes conducting in-depth analyses on multiple datasets. In addition, HR data is often noisy, inconsistent and contains missing information [4, 13], a problem that is exacerbated by the small proportion of employee turnover that typically exists within a given set of HR data. Secondly, gaps tend to persist in model performance evaluation. Specifically, previous research on the assessment of machine learning algorithms has generally focused on a narrow evaluation of metrics across various models

III. PROPOSED ALGORITHM

To predict employee attrition, we use sophisticated machine learning algorithms in our suggested research study, including Decision Tree Classifier (DTC), Random Forest (RF), Logistic Regression (LR), and Naïve Bayes Classifier (NB).

Decision Tree

A supervised learning technique called a decision tree creates regression or classification models in a structure like a tree. Since its initial introduction by Morgan and Sonquist in 1963, the approach has gained a solid reputation. The decision tree approach has a number of benefits, including the ability to automatically choose variables, handle missing values and mixed features, be robust conceptually, and be intuitive to read. Its forecast accuracy might not always be very competitive, though. High model variance also makes decision trees unstable, and even slight modifications to the input data can have a big impact on the tree's structure.

Random Forest

According to Breiman's research, random forests use an ensemble strategy that improves the fundamental decision tree structure by combining a collection of weak learners into a stronger learner. Divide and conquer tactics are used in this ensemble approach to improve algorithm performance. Multiple decision trees, also known as weak learners, are built using bootstrapped training sets in random forests. A random subset of predictors is selected as split candidates from the entire collection of predictors. The impact of any one predictor is reduced by restricting the total number of predictors taken into account for each tree (where m is smaller than P). As a result, there is a decreased chance that a small number of significant predictors will control the individual trees. The variance is reduced by random forests by averaging these uncorrelated trees.

Logistic Regression

Cox first presented the concept of Logistic Regression in 1958. It is a traditional classification approach that uses linear discriminants. It determines the likelihood that a given input point belongs to a particular class, and then draws a linear boundary to divide the input space into different sections depending on this likelihood. Because of its popularity and ability to handle classes that may be divided into linear segments, logistic regression is a popular tool for a wide range of applications.

Naïve Bayes

The Bayes Theorem is used by the probabilistic Naïve Bayes method to calculate the likelihood of an occurrence given past knowledge of associated features. The assumption of conditional independence of features, which states that the existence of one feature does not affect the presence of others, is a fundamental component of the Naïve Bayes algorithm. Based on this supposition, naïve Bayes classifiers learn the joint probability distribution of inputs and then apply Bayes Theorem to get the greatest posterior probability for a particular input.

Gathering of Data: Compile a thorough dataset with pertinent aspects including attrition status, job-related information, personnel demographics, and performance indicators. Make sure your data is consistent and of high quality by doing extensive pretreatment.
Analyzing exploratory data (EDA): To learn more about the distribution, correlation, and trends in the dataset, do exploratory data analysis (EDA). Investigate the connections between various factors to find probable indicators of staff attrition.
Engineering Features: Utilize feature engineering to take the raw data and turn it into useful information. This might involve scaling numerical features, encoding categorical variables, and generating new features by transforming or interacting with existing ones.
Model Choice: Choose a variety of machine learning models, such as Naive Bayes, Random Forest, Decision Trees, and Logistic Regression that are appropriate for categorization jobs. Take into account the features of the dataset and the presumptions of each model.
Adjusting Hyperparameters: Make use of methods such as GridSearchCV or RandomizedSearchCV to improve each model's hyperparameters. The method for finding the combination of hyperparameter values that optimizes model performance is to iteratively search across a range of values.
Assessment of the Model: Use relevant measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), to assess each modified model's performance. Use cross-validation to reduce overfitting and evaluate the models' capacity for generalization.
Comparing and Choosing: Select the best-performing model or models) for employee attrition prediction by comparing the performance of several models using assessment measures. Think about the trade-offs between performance, interpretability, and model complexity.
Verification and Interpretation: To make sure the chosen model(s) are reliable and resilient, validate them using a separate test dataset. Analyze the model's projections and pinpoint the main causes of employee attrition.
Record-keeping and Reporting: All phases of the technique, such as the data pretreatment, model training, hyperparameter tweaking, and assessment processes, should be documented. Write thorough reports outlining the conclusions, revelations, and suggestions for interested parties.
Repeated Fine Tuning: Iterate and improve the suggested process continuously in response to new information, feedback, and changing company needs. Increase prediction accuracy and usefulness by using more data sources or sophisticated modeling approaches.

This is a suggested machine learning method for forecasting staff attrition:

Data Input:
Employee dataset that includes attributes including attrition status, job-related details, performance indicators, and demographics as input.
Prior to processing:
To deal with missing numbers, outliers, and inconsistencies, clean up the dataset.
Use methods such as label encoding or one-hot encoding to encode categorical information.
In order to guarantee consistency and enhance model convergence, scale numerical characteristics.
To evaluate the model, divide the dataset into training and testing sets.
Model Choice:
Choose a variety of machine learning models, such as Naive Bayes, Random Forest, Decision Trees, and Logistic Regression that are appropriate for categorization jobs.
Take into account the computing resources, model assumptions, and dataset properties.
Adjusting Hyperparameters:
Make use of methods such as GridSearchCV or RandomizedSearchCV to improve each model's hyperparameters.
Investigate various hyperparameter combinations to optimize model performance while preventing overfitting.
Training Models:
The training dataset is used to train each tailored model. Modify the model's parameters in light of the optimization procedure after fitting the model to the data.
Assessment of the Model:
Use suitable assessment measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), to assess each trained model's performance. To evaluate the models' capacity for generalization and guarantee resilience, use cross-validation.

Comparing and Choosing:
Select the best-performing model or models) for employee attrition prediction by comparing the performance of several models using assessment measures. Think about the trade-offs between performance, interpretability, and model complexity.
Verification and Interpretation:
To guarantee generalizability and dependability, validate the chosen model or models using a separate test dataset. Analyze the model's projections and pinpoint the main causes of employee attrition, such as tenure, performance reviews, and work satisfaction.
Implementation:
Install the finished model or models in production settings to anticipate staff turnover in real time. In order to aid in decision-making and retention tactics, integrate the model or models with the current HR systems.
Observation and upkeep:
Recalibrate deployed models as necessary and keep an eye on their performance over time. Update the model(s) often with fresh information and stakeholder input to increase predicted relevance and accuracy.

IV. PSEUDO CODE

This pseudo code includes brief passages of Python code to demonstrate each phase of the process in an organized manner. It provides a concise synopsis of the tasks completed at every machine learning pipeline level. The data collection has been from Kaggle. These variables are crucial for managing the model's performance and complexity. The algorithm's initial step is to take the features (X) and labels (Y) out of the data set. After that, it divides the data in half, 80 to 20, into training and test sets.
The proposed algorithm consists of the following steps:

1. Input: Employee dataset containing features such as demographics, job-related information, performance metrics, and attrition status.

2. Preprocessing:

   - Clean the dataset to handle missing values, outliers, and inconsistencies.

   - Encode categorical variables using techniques like one-hot encoding or label encoding.

   - Scale numerical features to ensure uniformity and improve model convergence.

   - Split the dataset into training and testing sets for model evaluation.

3. Model Selection:

   - Select models: Logistic Regression, Decision Tree, Random Forest, Naive Bayes.

4. Hyper parameter Tuning:

   - For each model:

     - Define hyper parameter grid.

     - Perform hyper parameter tuning using techniques like GridSearchCV or RandomizedSearchCV.

5. Model Training:

   - For each tuned model:

     - Train the model using the training dataset.

6. Model Evaluation:

- For each trained model:

- Evaluate model performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score).

- Employ cross-validation to assess generalization ability and mitigate overfitting.

7. Comparison and Selection:

- Compare models based on evaluation metrics.

- Select top-performing model(s) for predicting employee attrition.

8. Validation and Interpretation:

- Validate selected model(s) using independent test dataset.

- Interpret model predictions and identify key factors contributing to employee attrition.

9. Deployment:

- Deploy final model(s) into production environments for real-time prediction.

- Integrate model(s) with existing HR systems for decision-making and retention strategies.

10. Monitoring and Maintenance:

- Monitor model performance over time.

- Recalibrate models as needed based on new data and feedback.

V. SIMULATION RESULTS

In conclusion, out of all the classifiers examined, Logistic Regression showed the greatest performance. The training and testing accuracy results are

```
[Logistic Regression] training data accuracy is: 0.884078
[Logistic Regression] test data accuracy is: 0.865922
[Decision Tree] training data accuracy is: 1.000000
[Decision Tree] test data accuracy is: 0.854749
[Random Forest] training data accuracy is: 1.000000
[Random Forest] test data accuracy is: 0.877095
[Naive Bayes] training data accuracy is: 0.870112
[Naive Bayes] test data accuracy is: 0.826816
```

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

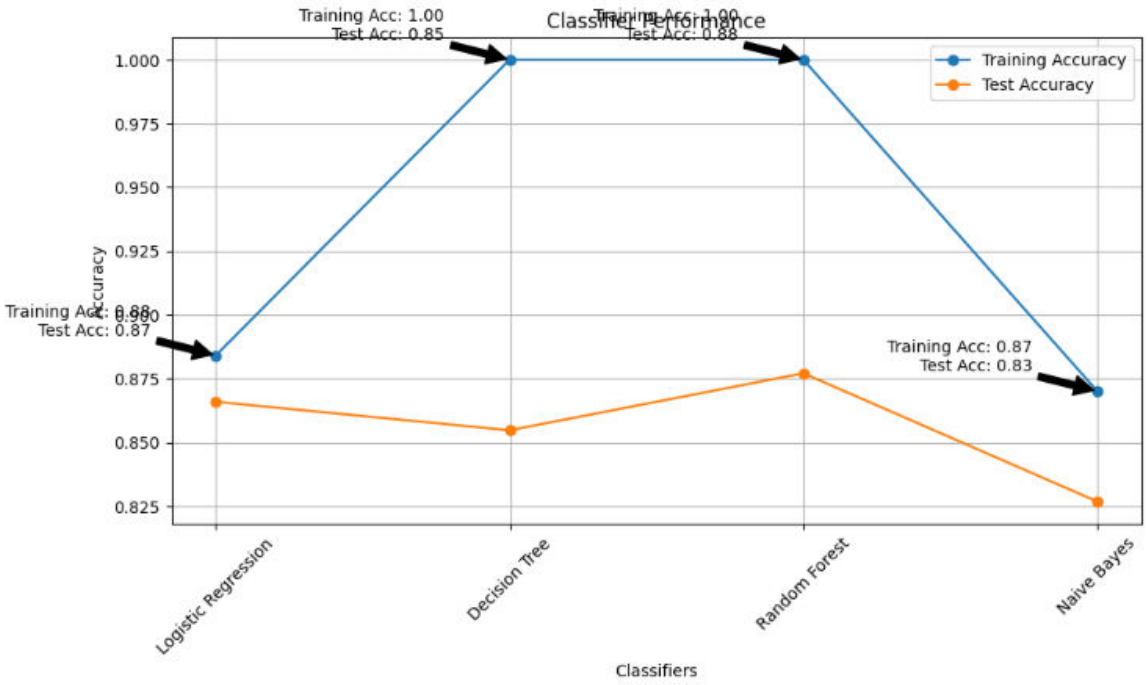

Fig1. Training and Testing Accuracy of various classifiers used

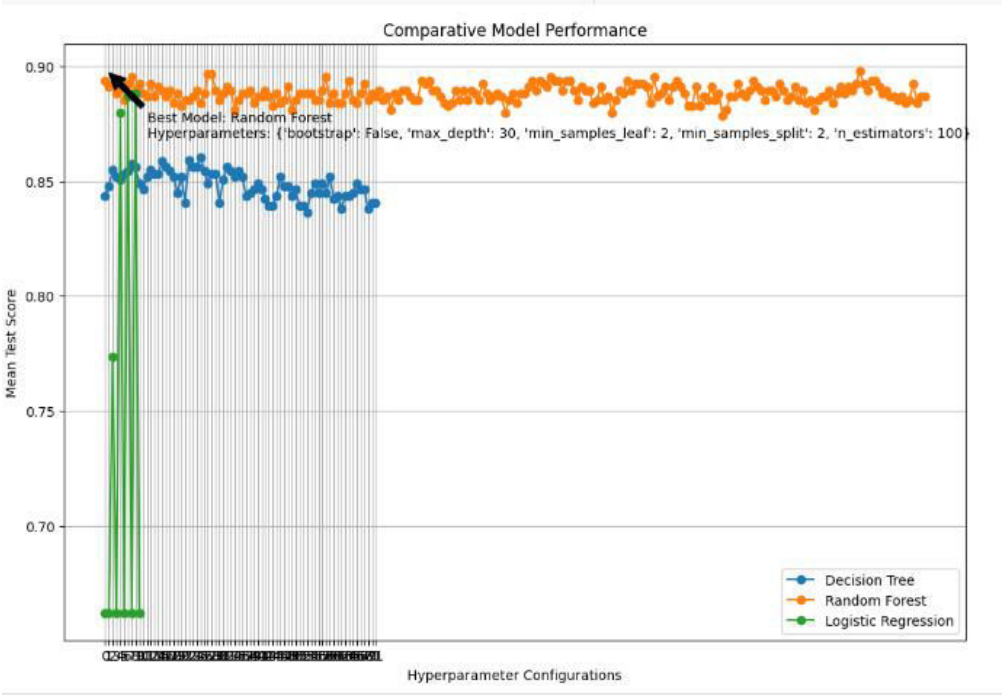

Fig2. Comparative model performance

With AUC values of 0.93 for both Random Forest and Logistic Regression, these classifiers demonstrate strong discriminative ability—that is, they can reliably discriminate between departing and remaining employees based on the provided data. When compared to Random Forest and Logistic Regression, Decision Tree and Naive Bayes exhibit somewhat weaker discriminative capacity, but overall good performance, with AUC values of 0.84.

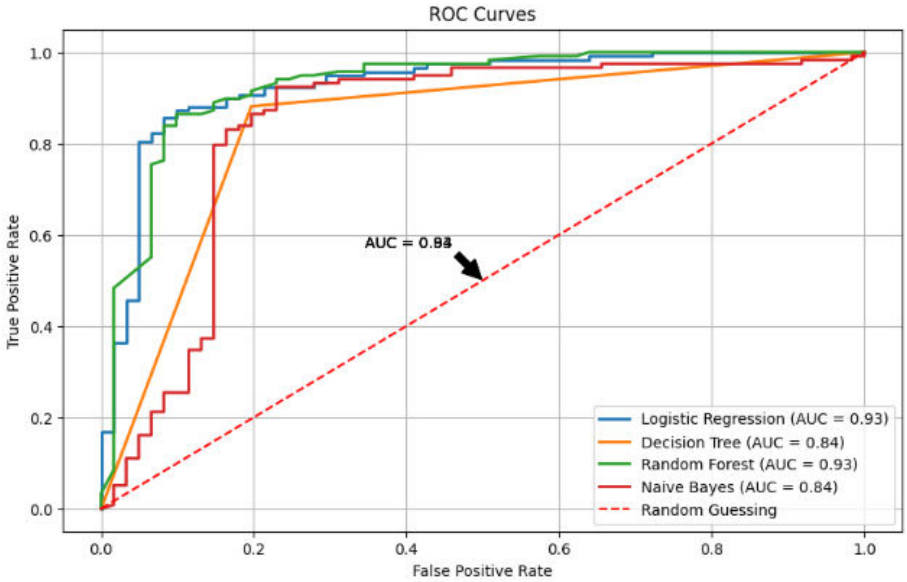

Fig3. AUC values of ML models used

VI. CONCLUSION AND FUTURE WORK

Workers leaving is seen as a major impediment to organizational expansion. Four supervised machine learning approaches are evaluated in this study. Several important conclusions and possible directions for further investigation are revealed by the comparison of the Random Forest, Decision Tree, Naïve Bayes and Logistic Regression algorithms.

Conclusion:

Of all the classifiers, Logistic Regression produced the best results, with an estimated training accuracy of 88.41% and an approximate test accuracy of 86.59% on both the training and test datasets. This suggests that, given the provided features, Logistic Regression is doing a good job of forecasting employee attrition.

The Random Forest and Decision Tree classifiers both reached 100% accuracy on the training dataset, suggesting that the training data may have been overfitted. They did, however, perform somewhat worse than Logistic Regression on the test dataset, with accuracies of around 85.47% and 87.71%, respectively.

On the training dataset, the Naive Bayes classifier obtained an accuracy of around 87.01%, while on the test dataset, it acquired an accuracy of 82.68%. It was not as good as Random Forest and Logistic Regression, but it was still not too bad.

Future Scope:

Additional hyperparameter fine-tuning: By lowering overfitting and boosting generalization to new data, hyperparameter tweaking may help Decision Tree and Random Forest classifiers perform better.

Feature engineering: By investigating new features or improving current ones, more pertinent data for forecasting staff turnover may be obtained, which might improve model performance.

Ensemble techniques: To combine the predictions of several classifiers and enhance performance, research ensemble techniques like bagging, boosting, or stacking.

Deep learning techniques: In order to capture the rich patterns and correlations present in the dataset, deep learning models like neural networks may be investigated, given the complexity and non-linearity of the data.

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]19 Employee Retention Statistics That Will Surprise You. 2022. Available online: https://www.apollotechnical.com/employee-retention-statistics/ (accessed on 6 May 2022).
[3]Chang, H.Y.: Employee turnover: a novel prediction solution with effective feature selection. WSEAS Trans. Inf. Sci. Appl. 6, 417–426 (2009)
[4]Here's What Your Turnover and Retention Rates Should Look Like. Available online: https://www.ceridian.com/blog/turnover-and-retention-rates-benchmark (accessed on 6 May 2022).
[5]SHRM Survey: Average Cost Per Hire Is $4129. Available online: https://www.businessmanagementdaily.com/46997/shrm-survey-average-cost-per-hire-is-4129/ (accessed on 6 May 2022).
[6]Gandomi, A.H.; Chen, F.; Abualigah, L. Machine Learning Technologies for Big Data Analytics. *Electronics* **2022**, *11*, 421. [Google Scholar] [CrossRef]
[7] Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.W.; Zou, B.; Hou, D. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. *Environ. Pollut.* **2021**, *270*, 116281. [Google Scholar] [CrossRef] [PubMed]
[8] Reshma Ramchandra, N.; Rajabhushanam, C. Machine learning algorithms performance evaluation in traffic flow prediction. *Mater. Today Proc.* **2022**, *51*, 1046–1050. [Google Scholar] [CrossRef]
[9] Aljedani, N.; Alotaibi, R.; Taileb, M. HMATC: Hierarchical multi-label Arabic text classification model using machine learning. *Egypt. Inform. J.* **2021**, *22*, 225–237. [Google Scholar] [CrossRef]
[10] Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; et al. Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data. *Diagnostics* **2022**, *12*, 203. [Google Scholar] [CrossRef] [PubMed]
[11] 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)
[12] 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
[13] Punnoose, R., Ajit, P.: Prediction of employee turnover in organizations using machine learning algorithms. Int. J. Adv. Res. Artif. Intell. 5, 22–26 (2016)
[14] 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)
[15] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, ACM (2016)
[16] Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer, New York (2001)
[17] 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)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details