JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2025**, 20, 069-075

DỰ ĐOÁN TÌNH TRẠNG NGHỈ VIỆC CỦA NHÂN VIÊN BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY

Lương Tiến Vinh[1], Phan Thị Ngân[2*]
[1]*Trường Đại học Quang Trung, số 327 Đào Tấn, phường Nhơn Phú, thành phố Quy Nhơn, tỉnh Bình Định, Việt Nam*
[2]*Hong Bang International University, Ho Chi Minh City, Viet Nam*
*Tác giả liên hệ: *p.thungan87@gmail.com*

THÔNG TIN BÀI BÁO

Ngày nhận: 11/1/2025
Ngày hoàn thiện: 4/2/2025
Ngày chấp nhận: 21/2/2025
Ngày đăng: 15/3/2025

TỪ KHÓA

Employee Exploratory Data Analysis;
IBM HR Employee Attrition;
Support Vector Machine;
Support Vector Machine;
Decision Tree Classifier;
Extra Trees Classifier.

TÓM TẮT

Tình trạng nhân viên nghỉ việc đặt ra thách thức nghiêm trọng đối với các tổ chức, cả về chi phí tài chính lẫn tính liên tục trong vận hành, với chi phí thay thế trung bình cho mỗi nhân sự ước tính là 4.129 USD và tỷ lệ nghỉ việc được báo cáo lên đến 57,3% vào năm 2021. Nghiên cứu này ứng dụng các kỹ thuật học máy để dự đoán tình trạng nghỉ việc của nhân viên và xác định các yếu tố tổ chức chính gây ra hiện tượng. Phân tích đánh giá bốn mô hình học có giám sát như Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), và Extra Trees Classifier (ETC), trong đó mô hình ETC sau khi được tối ưu hóa đã đạt độ chính xác dự đoán cao nhất là 93%, vượt qua các phương pháp tiên tiến hiện có. Phân tích khám phá dữ liệu nhân sự (EEDA) cho thấy thu nhập hàng tháng, mức lương theo giờ, cấp bậc công việc và độ tuổi là những yếu tố quan trọng ảnh hưởng đến tình trạng nghỉ việc. Các kết quả nghiên cứu nhấn mạnh tính hiệu quả của các phương pháp dựa trên trí tuệ nhân tạo trong phân tích lực lượng lao động, đồng thời cung cấp những thông tin có giá trị cho các nhà quản lý trong việc nâng cao hiệu quả giữ chân nhân viên thông qua các chiến lược dựa trên dữ liệu.

PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING APPROACHES

Luong Tien Vinh[1], Phan Thi Ngan[2*]
[1]*Quang Trung University, No. 327 Dao Tan Street, Nhon Phu Ward, Quy Nhon City, Binh Dinh Province, Vietnam*
[2]*Hong Bang International University, Ho Chi Minh City, Viet Nam*
*Corresponding Author: *p.thungan87@gmail.com*

ARTICLE INFO

Received: Jan 11st, 2025
Revised: Feb 4th, 2025
Accepted: Feb 21st, 2025
Published: Mar 15th, 2025

KEYWORDS

Employee Exploratory Data Analysis;
IBM HR Employee Attrition;
Support Vector Machine;
Support Vector Machine;
Decision Tree Classifier;
Extra Trees Classifier.

ABSTRACT

Employee attrition poses a critical challenge to organizations, both in terms of financial costs and operational continuity, with the average replacement cost per hire estimated at USD 4,129 and a reported attrition rate of 57.3% in 2021. This study applies machine learning techniques to predict employee attrition and identify its primary organizational drivers. Four supervised learning models were evaluated, Support Vector Machine (SVM), Support Vector Machine (LR), Decision Tree Classifier (DTC), and Extra Trees Classifier (ETC), in which the optimized ETC achieving the highest prediction accuracy of 93%, surpassing existing state-of-the-art methods. An Employee Exploratory Data Analysis (EEDA) revealed that monthly income, hourly rate, job level, and age are key factors influencing attrition. These findings highlight the effectiveness of AI-driven approaches in workforce analytics and provide actionable insights for organizational leaders aiming to improve retention through data-informed strategies.

Doi: https://doi.org/10.61591/jslhu.20.717
Available online at: https://js.lhu.edu.vn/index.php/lachong

1. INTRODUCTION

Employee attrition refers to the natural reduction in an organization's workforce due to various reasons such as resignations, retirements, or terminations. It is a common and ongoing process, yet when attrition occurs at a faster pace than hiring, it can lead to vacant positions, decreased productivity, and increased operational costs [1]. Measuring and monitoring the attrition rate provides valuable insight into an organization's health and workforce stability. A high attrition rate typically signals frequent employee turnover, which can hinder organizational progress and lead to significant losses [2]. Attrition can be voluntary, involuntary, external, or internal. Voluntary attrition occurs when employees leave by choice, while involuntary is initiated by the employer. External attrition involves moving to other organizations, and internal refers to role changes within the same company. Recognizing these types helps organizations identify and address the causes of turnover.

Employee attrition refers to the gradual reduction in workforce due to resignations, retirements, or terminations. When attrition outpaces hiring, it leads to unfilled roles, increased costs, and reduced productivity. It can be voluntary, involuntary, external, or internal, and measuring its rate helps assess organizational health. In 2021, the U.S. attrition rate reached 57.3%, with 3 to 4.5 million employees leaving jobs monthly according to the Job Openings and Labor Turnover Survey (JOLTS) [2]. High turnover poses significant challenges, especially given the average hiring cost of USD 4,129 as estimated by SHRM [4]. Maintaining a low attrition rate is essential for workforce stability and long-term organizational success [1], [3].

Machine learning (ML) in Artificial Intelligence (AI) enables machines to learn from historical data and make future predictions, becoming an essential part of data science. The objective of ML techniques is to outperform human accuracy, and these models are widely used in decision-making processes, where machines automatically learn from refined data to predict outcomes for new data. The primary aim is to identify patterns and gain insights from data. Machine learning applications are rapidly expanding, addressing real-world problems across various domains such as image recognition [6], traffic prediction [7], speech recognition, text classification [8], social analysis, stock market trading, healthcare [9], and agriculture. In the context of employee attrition, machine learning models have been used to predict workforce turnover [10].

Additionaly, the study evaluated various data balancing techniques to address the common challenge of imbalanced datasets in classification tasks, such as fraud or intrusion detection [11]. Techniques including SMOTE, Hybrid Sampling, and Clustering-Based Under Sampling were tested, and results showed that while data balancing improved model performance, no single method consistently outperformed the others. Another study explored the application of neural networks in real-world problem solving, emphasizing their advantages in handling large-scale data through parallel processing and high computational speed [12]. It was found that feedforward and feedback propagation networks performed well on large datasets, particularly due to their scalability, fault tolerance, and accuracy. These studies provide foundational insights into improving predictive models and underscore the importance of model selection and data preprocessing in attrition prediction.

This study contributes by applying four machine learning techniques: Extra Trees Classifier, Support Vector Machine, Logistic Regression, and Decision Tree Classifier to predict employee attrition. A comparative analysis was conducted to evaluate the models' accuracy, with the optimized Extra Trees Classifier achieving the highest performance, surpassing other techniques and state-of-the-art methods. Employee Exploratory Data Analysis was used to identify key factors influencing attrition, while the Synthetic Minority Oversampling Technique was applied to balance the dataset, improving model accuracy and reducing prediction complexity. Additionally, K-Fold cross-validation was performed to assess the robustness of the models.

The structure of this paper is as follows: Section 2 reviews related literature, Section 3 presents the methodology, Section 4 discusses the proposed approaches for employee attrition prediction, Section 5 covers the results and evaluations, and Section 6 concludes the study.

2. RELATED WORK

This section reviews recent literature focused on predicting employee attrition, highlighting key methodologies and findings from previous studies. Several recent studies have explored the prediction of employee attrition using machine learning techniques. One study applied classification algorithms such as K-Nearest Neighbors, Extreme Gradient Boosting, AdaBoost, Decision Tree, Neural Networks, and Random Forest to HR data from Kaggle, achieving an accuracy of 88% after regularization tuning [13], [14]. Another study utilized the IBM HR dataset and implemented models like AdaBoost, Random Forest Regressor, and Logistic Regression, with Decision Tree and Logistic Regression models achieving 86% accuracy in predicting attrition to support organizational decision-making [15], [16]. A three-stage framework using preprocessing, feature selection via max-out, and Logistic Regression reported an accuracy of 81% [17].

Comparative studies using the IBM HR dataset applied various models, with Random Forest achieving 85% accuracy and highlighting social, financial, and professional factors as key contributors to attrition [18], [19]. Another approach using gain ratio analysis and bootstrapping for data balancing identified the top influencing factors and reached an 80% accuracy rate [20]. Gradient boosting and ensemble learning, combined with hyperparameter tuning and k-fold validation, formed the basis of a machine learning pipeline that delivered state-of-the-art performance [21], [22]. Additionally, one study assessed emotional factors through survey data, applying Decision Tree, Random Forest, and SVM, achieving an

86% accuracy score [23]. Lastly, a systematic flow using multiple classifiers in Python identified Random Forest as the best performer with 83% accuracy and highlighted key causes of attrition through data analysis [24].

These studies collectively demonstrate the effectiveness of machine learning in predicting employee attrition and underscore the importance of data preprocessing, model selection, and feature importance analysis.

3. METHODOLOGY

The methodological framework of this research is illustrated in Figure 1. The IBM HR Employee Attrition dataset served as the foundation for model development and evaluation. To uncover meaningful patterns and influential factors related to employee attrition, EEDA was conducted. Feature engineering was employed to identify relevant attributes through correlation analysis and to perform necessary encoding procedures. The dataset was found to be imbalanced, so the SMOTE was applied to address class imbalance and enhance prediction performance.

Following preprocessing, the dataset was divided into training and testing subsets using an 85:15 split. Four machine learning algorithms were trained on the 85% training data and evaluated using the 15% test data. Each model underwent hyperparameter tuning to optimize performance. The finalized, generalized model was then capable of predicting employee attrition outcomes based on input employee information.

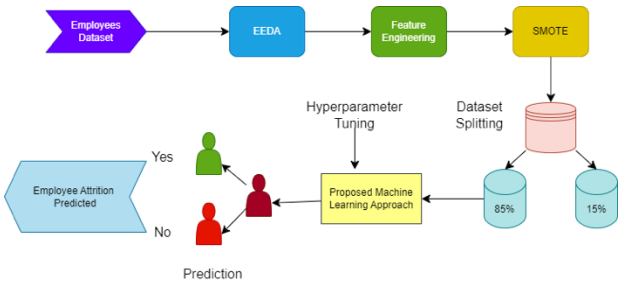


Figure 1. *The methodological analysis*

3.1 Data set

This study employed the IBM HR Employee Attrition dataset [26], created by IBM data scientists, to develop a generalized machine learning model for predicting employee attrition. The dataset includes 1,470 employee records and 35 features, occupying approximately 402 KB of memory. It was analyzed to identify key factors contributing to employee turnover, with a summary of feature details presented in Table 1.

Feature engineering played a critical role in optimizing model performance by transforming the data and improving predictive accuracy. Correlation analysis was conducted to identify and remove features with low or negative impact on the prediction task. As a result, 14 features—such as *DailyRate*, *DistanceFromHome*, *EmployeeCount*, and *StandardHours*—were excluded from the dataset. One-hot encoding was then applied to convert categorical variables into a machine-readable format [27]. These preprocessing steps significantly enhanced the model's ability to learn from the data and contributed to achieving high accuracy in predicting employee attrition.

Table 1. *The summary of feature details*

Column	*Data Type*	Column	*Data Type*
Age	Int64	StandardHours	Int64
Attrition	object	StockOptionLevel	Int64
BusinessTravel	Object	TotalWorkingYears	Int64
DailyRate	Int64	TrainingTimesLastYear	Int64
Department	Object	WorkLifeBalance	Int64
DistanceFromHome	Int64	YearsAtCompany	Int64
Education	Int64	YearsInCurrentRole	Int64
EducationField	Object	YearsSinceLastPromotion	Int64
EmployeeCount	Int64	YearsWithCurrManager	Int64
EmployeeNumber	Int64	Over18	Int64
EnvironmentSatisfaction	Int64	OverTime	Int64
Gender	Object	PercentSalaryHike	Int64
HourlyRate	Int64	PerformanceRating	Object
JobInvolvement	Int64	RelationshipSatisfaction	Object
JobLevel	Int64	MonthlyIncome	Int64
JobRole	Object	MonthlyRate	Int64
JobSatisfaction	Int64	NumCompaniesWorked	Int64
MaritalStatus	Object		

3.2 Employee Exploratory Data Analysis (EEDA)

The EEDA was instrumental in uncovering patterns and identifying key factors influencing attrition. As illustrated in Figure 2, analysis of age and monthly income distributions revealed that attrition is highest among employees aged 10 to 25, with a noticeable decline as age increases. Similarly, employees earning between USD

1,000 and 5,000 exhibited a higher attrition rate, highlighting the combined impact of age and income on turnover. Figure 5 presents a pair plot of key features, indicating elevated attrition rates among employees at job levels one and two, particularly within the first year of employment. Additionally, tenure with the current manager and overall years at the company were influential factors. These findings support the identification of critical drivers of attrition through data visualization and pattern analysis.

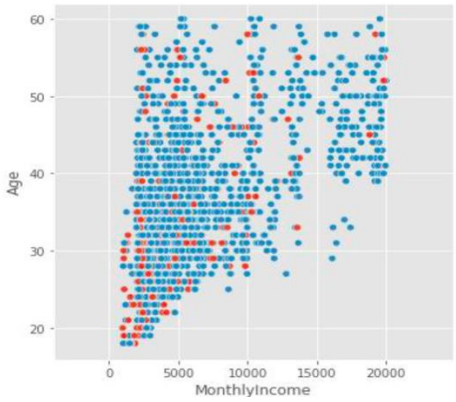
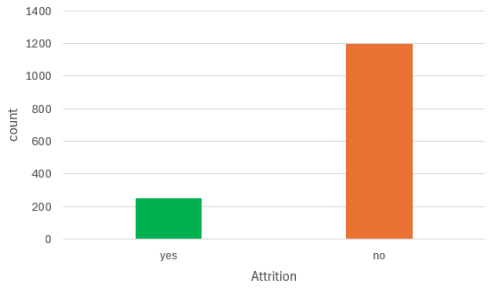


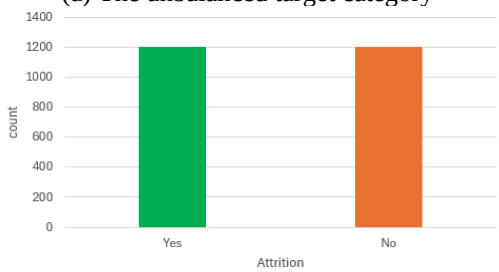
Figure 2. *The age and monthly income distributions*

3.3 Data Resampling

To address class imbalance in the dataset, the SMOTE was applied [28], resulting in a more balanced target distribution. This resampling helped reduce model complexity and improve accuracy by ensuring the learning algorithm was trained on an equal representation of classes. Figure 3 illustrates the data distribution before and after resampling.



(a) The unbalanced target category



(b) The target category after balancing the dataset.

Figure 3. *The dataset resampling using SMOTE technique*

4. PROPOSED MACHINE LEARNING APPROACHES

In this study, four advanced machine learning algorithms were employed for employee attrition prediction: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), and Extra Trees Classifier (ETC). Among these, the ETC model is proposed as the primary approach due to its superior performance.

A. Support Vector Machine (SVM)

The SVM technique [29, 30], a supervised learning algorithm, constructs a decision boundary referred to as a hyperplane, that effectively separates data points in an n-dimensional space. This hyperplane is formed by identifying support vectors, which are the critical data points closest to the boundary. The optimal hyperplane is iteratively refined to minimize classification error, as expressed mathematically as

$$\overrightarrow{W_1 x} + \vec{x} + b = 0, \qquad (1)$$

where *w* is the weight vector, *x* the input features, and *b* the bias term.

B. Logistic Regression (LR)

The LR is a supervised machine learning technique commonly used for binary classification tasks. It models the relationship between a dependent variable and one or more independent variables by employing a logistic (sigmoid) function [31]. This S-shaped curve maps predicted values to probabilities ranging between 0 and 1, making it suitable for classification problems. The LR model predicts the likelihood of a class label based on a linear combination of input features, as shown

$$y = \frac{e^{(b_0 + b_1 * x)}}{(1 + e^{(b_0 + b_1 * x)})},$$

where *y* represents the predicted output, b_0 is the bias term, and b_1 is the coefficient associated with input *x*.

C. Decision Tree Classifier (DTC),

DTC is a hierarchical structure where internal nodes represent attributes, branches denote decision rules, and leaf nodes correspond to target class labels. DTC operates by learning decision rules from training data to predict class outcomes. Its interpretability and similarity to human decision-making processes make it particularly useful. The model selects optimal attributes using metrics such as Information Gain and the Gini Index, the latter of which is calculated as:

$$\text{Gini Index} = 1 - \sum_{(j)} P_j^2 \qquad (3),$$

where P_j represents the probability of class *j*.

D. Extra Trees Classifier (ETC),

ETC is an advanced ensemble learning method that builds upon the principles of bagged decision trees. While conceptually similar to the Random Forest algorithm, ETC differs in its construction approach by introducing greater randomness in feature selection and decision rule formulation. Multiple de-correlated trees are generated through random splits on the training dataset, and their outputs are aggregated via majority voting to enhance

prediction performance. The entropy measure used for decision tree splitting is defined in Equation (4):

$$\text{Entropy}(S) = -\sum_i p_i \log_2(p_i). \quad (4)$$

During training and evaluation, the proposed ETC model achieved an accuracy of 93% on unseen data, demonstrating strong generalization capabilities. Accuracy is computed using Equation (5), based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

To further evaluate model performance, precision and recall were both calculated at 93%, using Equations (6) and (7), respectively. These metrics assess the model's ability to correctly classify positive instances. The F1-score, a harmonic mean of precision and recall, also reached 93%, indicating balanced performance across both measures.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

Additionally, the model's log loss, which quantifies prediction error in probabilistic classification, was calculated at 2.33, as expressed in Equation (8).

$$\text{F1-score} = \frac{2^* \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (8)$$

5. RESULTS AND DISCUSSIONS

The experiments were conducted on a system with an Intel® Xeon® CPU, 13 GB RAM, and AMD EPYC 7B12 processor. We evaluated four models, such as LR, SVM, DTC, and ETC using accuracy, precision, recall, and F1-score.

In Table 2 shows LR achieved the lowest accuracy (72%), followed by DTC (83%) and SVM (87%). The proposed ETC model outperformed the others with 93% accuracy, precision, recall, and F1-score. Dataset balancing using SMOTE improved performance, and ETC proved most effective for predicting employee attrition.

Table 2 *The accuracy performance*

Models	*Accuracy Score %*
ETC	93
SVM	87
LR	72
DTC	83

Table 3 presents the classification report for all machine learning models based on precision, recall, F1-score, and support—both category-wise and average. The proposed ETC outperformed others with the highest scores across all metrics.

Table 3. *The employed approaches classification*

Category	*Precision%*	*Recall%*	*F1 Score%*	*Support Score*
ETC				
0	92	95	94	193
1	93	92	93	177
Average	92	93	93	370
SVM				
0	90	84	86	185
1	85	90	87	186
Average	88	88	86	373
LR				
0	77	69	71	195
1	70	75	72	173
Average	74	73	72	375
DTC				
0	86	82	85	192
1	81	83	84	180
Average	83	82	83	377

To further validate performance, 10-fold cross-validation was applied to all models. The k-fold results, shown in Table 4 and Figure 4, demonstrate improved model accuracy, with ETC consistently achieving superior results.

Table 4. *The K-Fold cross-validation comparative analysis among the employed approaches*

Models	*K-Fold*	*Accuracy Score%*
ETC	10	93
SVM	10	88
LR	10	74
DTC	10	84

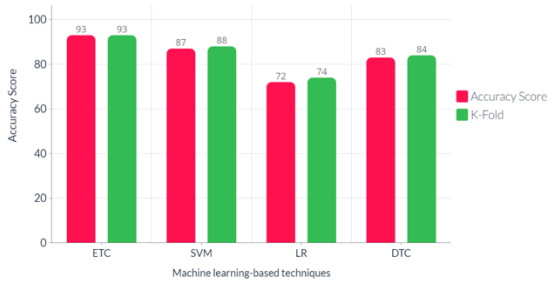


Figure 4. *The comparative performance analysis of employed approaches with K-Fold validation*

Comparative analysis in Tables 5 and 6 shows that the proposed ETC model surpasses previously applied state-of-the-art approaches, establishing it as the most effective method for predicting employee attrition.

Table 5. *The proposed ETC approach performance*

Accuracy Score%	Precision%	Recall %
10	SVM	88
10	LR	74
10	DTC	84

Table 6. *The performance validation comparative analysis*

Literature	*Technique*	*Accuracy Score%*
[13]	Decision Tree	88
[15]	Decision Tree + Logistic Regression	86
[17]	Logistic Regression	81
[18]	Random Forest	85
[20]	Support Vector machines	80
Proposed	ETC	93

6. CONCLUSION

This study employed four advanced machine learning techniques, such as ETC, SVM, LR, and DTC to predict employee attrition. The ETC model outperformed the others, achieving 93% accuracy, precision, recall, F1 score, and ROC-AUC. In comparison, SVM, LR, and DTC achieved accuracy scores of 87%, 72%, and 83%, respectively. After applying SMOTE for data balancing and validating results through 10-fold cross-validation, ETC consistently maintained the highest performance, while SVM, LR, and DTC achieved 88%, 74%, and 84%, respectively. The EEDA identified key attrition factors, including monthly income, hourly rate, job level, and age. These findings provide valuable insights for organizations to address employee turnover. As a future direction, the study proposes the integration of deep learning models and enhanced feature engineering to further improve prediction accuracy.

7. REFERENCES

[1] S. Singh and P. Gupta, “Comparative study ID3, cart and C4 . 5 Decision tree algorithm: a survey,” Int. J. Adv. Inf. Sci. Technol., 2014.

DOI: 10.15693/ijaist/2014.v3i7.47-52.

[2] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

DOI: 10.1023/A:1010933404324.

[3] H. Aydadenta and Adiwijaya, “A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest,” J. Inf. Process. Syst., vol. 14, no. 5, pp. 1167–1175, 2018.

DOI: 10.3745/JIPS.04.0087.

[4] G. Esteves and J. Mendes-Moreira, “Churn perdiction in the telecom business,” in 2016 11th International Conference on Digital Information Management, ICDIM 2016, 2016.

DOI: 10.1109/ICDIM.2016.7829775.

[5] A. Sonak and R. A. Patankar, “A Survey on Methods to Handle Imbalance Dataset,” Int. J. Comput. Sci. Mob. Comput., vol. 4, no. 11, pp. 338–343, 2015, available at : Google Scholar.

[6] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” Int. J. Adv. Soft Comput. its Appl., vol. 7, no. 3, pp. 176-203, 2015, available at: http://home.ijasca.com/data/ documents/13IJASCA-070301_Pg176-204_Classification-with-class-imbalance-problem_A-Review.pdf .

[7] S. Du, F. Zhang, and X. Zhang, “Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach,” ISPRS J. Photogramm. Remote Sens., 2015.

DOI: 10.1016/j.isprsjprs.2015.03.011.

[8] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, “An Ensemble Random Forest Algorithm for Insurance Big Data Analysis,” in Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017, 2017.

DOI: 10.1109/CSE-EUC.2017.99.

[9] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” BMC Med. Inform. Decis. Mak., 2011.

DOI: 10.1186/1472-6947-11-51.

[10] V. Effendy and Z. K. a. Baizal, “Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest,” 2014 2nd Int. Conf. Inf. Commun. Technol., 2014.

DOI: 10.1109/ICoICT.2014.6914086.

[11] E. Dwiyanti, Adiwijaya, and A. Ardiyanti, “Handling imbalanced data in churn prediction using RUSBoost and feature selection (Case study: PT. Telekomunikasi Indonesia regional 7),” in Advances in Intelligent Systems and Computing, 2017.

DOI: 10.1007/978-3-319-51281-5_38.

[12] Ł. Kobyliński and A. Przepiórkowski, "Definition extraction with balanced random forests," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008.

DOI: 10.1007/978-3-540-85287-2_23.

[13] S. Ghosh and S. Kumar, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," Int. J. Adv. Comput. Sci. Appl., 2013.

DOI: 10.14569/IJACSA.2013.040406.

[14] S. Venkateswara and V. Swamy, "A Survey : Spectral Clustering Applications and its Enhancements," Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 1, pp. 185–189, 2015, available at: Google Scholar.

[15] A. Y. Shelestov, "Using the agglomerative method of hierarchical clustering as a data mining tool in capital market," Int. J. "Information Theor. Appl., vol. 15, no. 1, pp. 382–386, 2018, available at: http://hdl.handle.net/10525/80.

[16] K. Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm-A Review," Int. J. Sci. Res. Publ., 2013.

DOI: 10.1016/S0090-3019(03)00579-2.

[17] W. Tian, Y. Zheng, R. Yang, S. Ji, and J. Wang, "A Survey on Clustering based Meteorological Data Mining," Int. J. Grid Distrib. Comput., vol. 7, no. 6, pp. 229–240, 2014, available at: Google Scholar.

[18] A. Chowdhary, "Community Detection: Hierarchical clustering Algorithms," Int. J. Creat. Res. Thoughts, vol. 5, no. 4, pp. 2320–2882, 2017, available at: http://ijcrt.org/papers/IJCRT1704418.pdf.

[19] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Univ. California, Berkeley, 2004, available at: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

[20] D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: a Review," Int. J. Comput. Bus. Res., vol. 5, no. 4, 2014, available at: http://www.researchmanuscripts.com/July2014/2.pdf.

[21] S. Sardari, M. Eftekhari, and F. Afsari, "Hesitant fuzzy decision tree approach for highly imbalanced data classification," Appl. Soft Comput. J., 2017.

DOI: 10.1016/j.asoc.2017.08.052.

[22] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," Glob. J. Technol. Optim., 2018.

DOI: 10.4172/2229-8711.s1111.

[23] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," J. Inf. Eng. Appl., vol. 3, no. 10, pp. 27-38, 2013, available at: Google Scholar.

[24] C. G. Weng and J. Poon, "A new evaluation measure for imbalanced datasets," Proceedings of the 7th Australasian Data Mining Conference., vol. 87, no. 6, pp. 27-32, 2008, available at: http://dl.acm.org/citation.cfm?id=2449288.2449295.

[25] J. S. Akosa, "Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data," SAS Glob. Forum, 2017, available at: Google Scholar.

[26] Y. Zhang and D. Wang, "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets," Abstr. Appl. Anal., 2013

DOI: 10.1155/2013/196256.

[27] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., 2006.

DOI: 10.1016/j.patrec.2005.10.010.

[28] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, 2015.

DOI: 10.5121/ijdkp.2015.5201.

[29] A. K. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," IJCSI Int. J. Comput. Sci. Issues, 2012, available at: Google Scholar.

[30] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," Int. J. Geogr. Inf. Sci., 2017.

DOI: 10.1080/13658816.2017.1346255.