INTENSIF, Vol.9 No.1 February 2025
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

# Student Dropout Prediction Using Random Forest and XGBoost Method

**Received:**
8 June 2024
**Accepted:**
23 February 2025
**Published:**
28 February 2025

**[1*]Lalu Ganda Rady Putra, [2]Didik Dwi Prasetya, [3]Mayadi**
*[1,2]Teknik Elektro dan Informatika Universitas Negeri Malang*
*[3]Kolej Pengajian Pengkomputeran, Informatik dan Media, Universiti Teknologi Mara*
*E-mail: [1]lalu.ganda.2305349@students.um.ac.id, [2]didikdwi@um.ac.id, [3]2023336871@student.uitm.edu.my*
*Corresponding Author

**Abstract**—**Background:** The increasing dropout rate in Indonesia poses significant challenges to the education system, particularly as students advance through higher education levels. Predicting student attrition accurately can help institutions implement timely interventions to improve retention. **Objective:** This study aims to evaluate the effectiveness of the Random Forest and XGBoost algorithms in predicting student attrition based on demographic, socioeconomic, and academic performance factors. **Methods:** A quantitative study was conducted using a dataset of 4,424 instances with 34 attributes, categorized into Dropout, Graduate, and Enrolled. The performance of Random Forest and XGBoost was compared based on accuracy, specificity, and sensitivity. **Results:** Random Forest achieved the highest accuracy at 80.56%, with a specificity of 76.41% and sensitivity of 72.42%, outperforming XGBoost. While XGBoost was slightly less accurate, it remained a competitive approach for student attrition prediction. **Conclusion:** The findings highlight Random Forest's robustness in handling extensive datasets with diverse attributes, making it a reliable tool for identifying at-risk students. This study underscores the potential of machine learning in addressing educational challenges. Future research should explore advanced ensemble techniques, such as the Ensemble Voting Classifier, or deep learning models to further enhance prediction accuracy and scalability.

**Keywords**— Student Dropout; Prediction; Random Forest; XGBoost



***Corresponding Author:***

Lalu Ganda Rady Putra,
Teknik Elektro dan Informatika,
Universitas Negeri Malang,
Email: lalu.ganda.2305349@students.um.ac.id,
Orchid ID: https://orcid.org/0000-0002-7596-5734

INTENSIF, Vol.9 No.1 February 2025
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

## I. INTRODUCTION

Ensuring access to high-quality education is a fundamental entitlement for all individuals. The attainment of education exhibits a strong positive correlation with an individual's prospective achievements, owing to its influence on occupational opportunities and an enhanced standard of living [1][2]. The prioritization of enabling students to successfully complete their academic pursuits is of utmost significance for educational establishments [3]. The school dropout rate in Indonesia has an upward trend in correspondence with the progression of educational levels [4]. The factors contributing to students' inability to successfully complete their studies encompass a range of elements, such as inadequate academic aptitude, entry age, grades, and other related issues [5][6]. The precise and timely identification of students at risk is a crucial factor in mitigating school dropout rates [7]. Hence, it is imperative for educational institutions to proactively address the issue of student attrition by including a predictive system.

The application of machine learning techniques enables the effective management of educational data through the process of categorising data into discernible information and generating valuable insights to support decision-making [8]. Machine learning is a methodology that educational institutions can employ to proactively identify students who are at risk of discontinuing their studies [9]. Numerous scholarly investigations have delved into this subject matter [10][11]. A study was conducted to investigate the phenomenon of student attrition at different educational levels, encompassing both secondary education (specifically high school) [12] and tertiary education (specifically university) [13]. The sources of data utilised in this study exhibit variation, encompassing student performance on Massive Open Online Courses (MOOCs) [14], electronic courses (e-courses) [15], and data from institutional academic systems [16]. Previous research on student dropout in Germany was undertaken by Kemper [17]. In Kemper’s study, two techniques, namely logistic regression and decision trees, were employed to predict student dropout. The researchers discovered that the most significant variables in forecasting student attrition were performance on the examination (passing or failing) and the grade point average. In a previous study, Cannistra [18] found that the initial academic attainment and performance were the key factors influencing the outcomes.

These diverse research offer varying perspectives on the prediction of student attrition. This study will utilise a dataset comprising information from several organisations encompassing a range of disciplines including agronomy, design, education, nursing, journalism, management, social services, and technology. In addition to this, the present research diverges from other studies in its use of algorithms. The algorithms employed in this study encompass Random Forest

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

and XGBoost. The objective of this study is to evaluate the efficacy of the Random Forest and XGBoost algorithms in forecasting dropping students.

## II. RESEARCH METHOD

This study aims to investigate the correlation between demographic data, socioeconomic characteristics, and academic performance information by employing a classification system. Subsequently, a model was devised with the capability to discern student risk factors associated with dropout and implement timely interventions aimed at enhancing student retention rates. The research process diagram stage is displayed in Figure 1

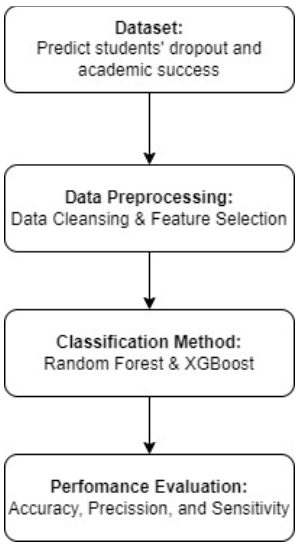



**Fig 1.** Research Stage

A. Data Collection

This study utilizes the “Predict students' dropout and academic success” dataset obtained from Kaggle[19]. The number of instances in the dataset is 4424 instances, 34 input attributes and one output attribute with three categories: Dropout, Graduate, and Enrolled (See Table 1). The attributes owned by the dataset are Marital status, Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Nacionality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment, International, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved),

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations), Unemployment rate, Inflation rate, GDP, Target.

**Table 1.** Sample Data [19]

| Marital status | Application mode | Application order | Course | … | Previous qualification | Target |
|---|---|---|---|---|---|---|
| 1 | 8 | 5 | 2 | … | 1 | Dropout |
| 1 | 6 | 1 | 11 | … | 1 | Graduate |
| 1 | 1 | 5 | 5 | … | 1 | Dropout |
| 1 | 8 | 2 | 15 | … | 1 | Graduate |
| 2 | 12 | 1 | 3 | … | 1 | Graduate |

B. Data Pre-Processing

1. Data Cleansing

Data cleansing and transformation are essential components of the data pre-processing phase. During this step, the process involves the removal of undesirable data and the rectification of missing values, or NA values[20]. Next, we proceed to eliminate a small number of erroneous and outlier data points that have the potential to introduce errors in our prediction models.

2. Feature Selection

A correlation matrix is used to identify the most relevant attributes by analyzing the relationships between features and the target variable. Features with low correlation to the target variable are considered less impactful and are removed from the dataset[21]. This process helps eliminate redundant or irrelevant attributes, improving model performance and reducing computational complexity. By focusing on highly correlated features, the predictive accuracy of the model is enhanced. As a result, the number of features is reduced to 25, retaining only the most significant variables for student dropout prediction.

C. Classification Method

This study uses the Random Forest and XGBoost classification techniques to identify students who indicate greater chances of discontinuing their education. The Random Forest algorithm is a powerful ensemble technique that can effectively handle both regression and classification tasks. It achieves this by utilizing several decision trees and employing a method known as Bootstrap and Aggregation, which is generally referred to as bagging. The fundamental concept underlying this approach is to aggregate many decision trees to decide the ultimate output, as opposed to depending solely on individual decision trees[22].

The Random Forest algorithm utilizes a collection of decision trees as its basis learning models. Row sampling and feature sampling are performed randomly on the dataset to create sample datasets for each model. The algorithm under consideration exhibits a high level of user-friendliness and robustness about the training data, particularly when compared to the decision

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

tree. The fundamental concept behind this approach is to aggregate the outputs of many decision trees to decide the final prediction, rather than relying on a single tree. The final classification in Random Forest is determined through majority voting, as given in **Equation (1)**:

$$\hat{y} = \underset{mode}{} \{h_1(x), h_2(x), \dots, h_n(x)\} \quad (1)$$

where:

- $\hat{y}$ is the predicted class,
- $h_1(x)$ represents the individual decision tree predictions,
- $n$ is the total number of decision trees in the forest.

For regression tasks, the final prediction is obtained by averaging the outputs of individual trees, as shown in **Equation (2)**:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} h_i(x) \quad (2)$$

Additionally, feature importance in Random Forest is often determined using Gini Importance, which is calculated as follows (**Equation (3)**):

$$I_G(f) = \sum_{t \in T} p_t \cdot \Delta G_t(f) \quad (3)$$

where:

- $I_G(f)$ represents the importance of feature $(f)$,
- $p_t$ is the probability of reaching node $t$,
- $\Delta G_t(f)$ is the decrease in Gini impurity due to the feature $(f)$.

Meanwhile, the XGBoost method is a highly optimized and distributed gradient boosting library that has been specifically created to facilitate efficient and scalable training of machine learning models. The ensemble learning technique under consideration is a methodology that integrates the predictions of numerous weak models to get a more robust and accurate forecast. XGBoost, short for "Extreme Gradient Boosting," has gained significant popularity and widespread adoption as a machine learning algorithm. This can be attributed to its capability to effectively handle extensive datasets and its remarkable performance in various machine learning tasks, including classification and regression, often surpassing existing benchmarks.

One of the primary advantages of XGBoost is its effective management of missing values, enabling it to handle real-world datasets containing missing values without necessitating extensive preprocessing steps. Moreover, XGBoost possesses inherent capabilities for parallel

INTENSIF, Vol.9 No.1 February 2025
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

processing, hence enabling the training of models on extensive datasets within a feasible timeframe[23].

The objective function in XGBoost consists of a loss function and a regularization term, as defined in **Equation (4)**:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \boldsymbol{\Omega}(T_k) \quad (4)$$

where:

- $l(y_i, \hat{y}_i)$ is the loss function measuring the difference between actual and predicted values,
- $\Omega(T_k)$ is the regularization term to control model complexity,
- $K$ represents the total number of trees.

The model is updated using gradient boosting, where the weight of each tree is determined by minimizing the second-order approximation, as shown in Equation (5):

$$g_i = \frac{\partial l(y_i, \hat{y_i})}{\partial \hat{y_i}}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y_i})}{\partial \hat{y_i}^2} \quad (5)$$

where:

$g_i$ and $h_i$ are the first and second-order gradients of the loss function, respectively.

One of the key advantages of XGBoost is its ability to handle missing values by learning the optimal split direction for missing data, which significantly reduces the need for extensive preprocessing. Furthermore, XGBoost supports parallel processing, allowing for efficient training on large datasets within a feasible timeframe.

D. Performance Evaluation

The models are evaluated using a confusion matrix, assessing accuracy, sensitivity, and specificity. The dataset is split into training (80%) and testing (20%) sets to prevent overfitting. To assess the performance of a model, the confusion matrix combines predicted values from the model with actual values extracted from observed data. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four values comprising the confusion matrix[24][25]. The applied model in this study utilizes a number of performance measurement metrics, including precision, sensitivity, and specificity. By calculating the ratio of accurate predictions (true positives and true negatives) to the total number of samples, one can determine the accuracy. The sensitivity metric is employed to assess the degree of precision in true positives, or positive class predictions. When attempting to reduce the occurrence of false negatives

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

(positive cases that are erroneously classified as negative), this metric is crucial. In the interim, specificity quantifies the degree of precision exhibited by predictions of negative classes (true negatives). The formula utilized to determine accuracy, sensitivity, and specificity is denoted as (6), (7), dan (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

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

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

Where:

TP (True Positive) – The number of correctly classified positive instances.

FN (False Negative) – The number of actual positive instances incorrectly classified as negative.

FP (False Positive) – The number of actual negative instances incorrectly classified as positive.

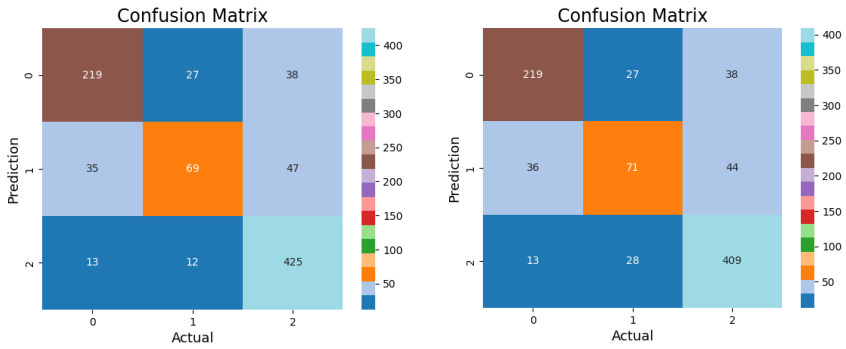
TN (True Negative) – The number of correctly classified negative instances.

## III. RESULT AND DISCUSSION

In this research, we created a prediction model using the RF and XGBoost methods. Before making predictions, we carry out feature selection on the “Predict students' dropout and academic success” dataset. Feature selection is carried out using a correlation matrix to determine the relationship between features and targets[21]. With the correlation matrix, you can find out the correlation value between attributes in the “Predict students' dropout and academic success” dataset. The correlation value is used to determine which columns to delete based on low correlation with the target variable ('Target').

After eliminating several variables that had low correlation, 25 attributes were obtained that met the criteria. Next, a data-splitting step is carried out before applying the RF and XGBoost algorithms. Data splitting is intended to divide training data and test data. Data splitting uses an 80:20 composition, with 80% training data and 20% test data[26]. We use a ratio of 80:20 to avoid overfitting in the model used.

INTENSIF, Vol.9 No.1 February 2025
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191




(a) (b)

**Fig 2.** Confusion Matrix (a) Random Forest; (b) XGBoost

This research uses the RF and XGBoost methods to predict student dropout. The prediction results from the two methods are then evaluated using a confusion matrix. The results of the confusion matrix can be seen in Figure 1. Figure 1 (a) shows that the RF algorithm was able to predict dropout for 219 out of 284 instances, Enrolled for 69 out of 151 instances, and Graduate for 425 out of 450 instances. Meanwhile, Figure 1 (b) is the result of the confusion matrix from XGBoost. It can be seen that from 284 instances dropout was successfully predicted according to 219 instances. Enrolled and Graduated respectively produce 71 of 151 and 409 of 450 instances.

Based on the confusion matrix values obtained, accuracy, sensitivity, and specificity can be calculated using formulas (6), (7), and (8). After feature selection, the dataset was processed using Random Forest and XGBoost. The confusion matrix results indicate that Random Forest performs better (see Table 2). It can be seen that the accuracy results from RF are higher compared to XGboost, 80.56% for RF and 78.98 % for XGBoost. Random Forest exhibits higher accuracy and specificity, making it the superior method for predicting student dropout.

**Table 2.** Experimental Result

| Method | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Random Forest | 80.56 % | 72.42 % | 76.41 % |
| XGBoost | 78.98 % | 71.67 % | 73.79 % |

Furthermore, this study is in line with previous research conducted by Sushma[27], Xu[28], and Quevedo[29], who also concluded that Random Forest outperforms XGBoost in handling certain attributes. Their findings support the results of this study, which show that Random Forest's ability to effectively process categorical and numerical features gives it an edge over XGBoost in predicting student dropout rates.

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v9i1.21191

## IV. CONCLUSION

This study demonstrates that machine learning techniques can effectively predict student dropout. Among the evaluated methods, Random Forest outperformed XGBoost with an accuracy of 80.56%, making it a more reliable option for student attrition prediction. The results indicate that Random Forest provides better generalization due to its ability to handle diverse datasets, while XGBoost remains a competitive alternative with its optimization capabilities. The evaluation metrics, including accuracy, sensitivity, and specificity, reinforce the importance of using ensemble learning methods for predicting dropout rates.

**Author Contributions:** *Lalu Ganda Rady Putra*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Didik Dwi Prasetya*: Software, Investigation, Data Curation, Writing - Original Draft. *Mayadi*: Investigation, Data Curation.

All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no specific grant from any funding agency.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Data Availability:** You can contact author's email whenever you need the explanation for this.

**Informed Consent:** There were no human subjects.

**Animal Subjects:** There were no animal subjects.

**ORCID**:
Lalu Ganda Rady Putra: https://orcid.org/0000-0002-7596-5734
Didik Dwi Prasetya: https://orcid.org/0000-0002-3540-2961
Mayadi: https://orcid.org/0000-0002-4585-9973

## REFERENCES

[1] R. W. Rumberger, "The economics of high school dropouts," in *The Economics of Education*, Elsevier, 2020, pp. 149–158. **doi:** 10.1016/B978-0-12-815391-8.00012-4.

[2] T. D. Snyder, C. de Brey, and S. A. Dillow, *Digest of Education Statistics*, vol. 51, no. 10. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, 2014. **doi:** 10.5860/choice.51-5366.

[3] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 233, p. 120933, Dec. 2023, **doi:** 10.1016/j.eswa.2023.120933.

[4] J. J. Lanawaang and R. Mesra, "Faktor Penyebab Anak Putus Sekolah di Kelurahaan Tuutu Analisis Pasal 31 Ayat 1, 2, dan 3 UUD 1945," *Jurnal Ilmiah Mandala Education*, vol. 9, no. 2, Apr. 2023, **doi:** 10.58258/jime.v9i2.5103.

[5] C. Marquez-Vera, C. R. Morales, and S. V. Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7–14, Feb. 2013, **doi:** 10.1109/RITA.2013.2244695.

[6] J. E. Nieuwoudt and M. L. Pedler, "Student Retention in Higher Education: Why Students Choose to Remain at University," *Journal of College Student Retention: Research, Theory & Practice*, vol. 25, no. 2, pp. 326–349, Aug. 2023, **doi:** 10.1177/1521025120985228.

[7] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950–965, Nov. 2009, **doi:** 10.1016/j.compedu.2009.05.010.

[8] H. Luan and C.-C. Tsai, "A Review of Using Machine Learning Approaches for Precision Education," *Educational Technology & Society*, vol. 24, no. 1, pp. 250–266, Nov. 2021, [Online]. Available: https://www.jstor.org/stable/26977871

[9] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6--10, 2020, Proceedings, Part I 21*, 2020, pp. 129–140.

[10] N. Y. L. Gaol, "Prediksi Mahasiswa Berpotensi Non Aktif Menggunakan Data Mining dalam Decision Tree dan Algoritma C4.5," *Jurnal Informasi & Teknologi*, pp. 23–29, Mar. 2020, **doi:** 10.37034/jidt.v2i1.22.

[11] M. T. Anwar and D. R. A. Permana, "Perbandingan Performa Model Data Mining untuk Prediksi Dropout Mahasiswa," *Jurnal Teknologi dan Manajemen*, vol. 19, no. 2, pp. 33–40, Aug. 2021, **doi:** 10.52330/jtm.v19i2.34.

[12] J. Y. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," *Children and Youth Services Review*, vol. 96, pp. 346–353, Jan. 2019, **doi:** 10.1016/j.childyouth.2018.11.030.

[13] Y. Zheng, Z. Shao, M. Deng, Z. Gao, and Q. Fu, "MOOC dropout prediction using a fusion deep model based on behaviour features," *Computers and Electrical Engineering*, vol. 104, p. 108409, Dec. 2022, **doi:** 10.1016/j.compeleceng.2022.108409.

[14] H. Aldowah, H. Al-Samarraie, A. I. Alzahrani, and N. Alalwan, "Factors affecting student dropout in MOOCs: a cause and effect decision-making model," *Journal of Computing in Higher Education*, vol. 32, no. 2, pp. 429–454, Aug. 2020, **doi:** 10.1007/s12528-019-09241-y.

[15] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model," *Decision Support Systems*, vol. 135, p. 113325, Aug. 2020, **doi:** 10.1016/j.dss.2020.113325.

[16] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 289–295, 2023, **doi:** 10.18178/ijiet.2023.13.2.1806.

[17] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, Jan. 2020, **doi:** 10.1080/21568235.2020.1718520.

[18] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni, "Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques," *Studies in Higher Education*, vol. 47, no. 9, pp. 1935–1956, Sep. 2022, **doi:** 10.1080/03075079.2021.2018415.

[19] V. Realinho, J. Machado, L. Baptista, and M. V Martins, "Predict students' dropout and academic success." Zenodo, Dec. 2021. **doi:** 10.5281/zenodo.5777340.

[20] N. McKelvey, K. Curran, and L. Toland, "The Challenges of Data Cleansing with Data Warehouses," pp. 77–82. **doi:** 10.4018/978-1-5225-0182-4.ch005.

[21] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature Selection: A Review and Comparative Study," *E3S Web of Conferences*, vol. 351, p. 01046, May 2022, **doi:** 10.1051/e3sconf/202235101046.

[22] P. Dangeti, *Statistics for Machine Learning*. Packt Publishing, 2017. [Online]. Available: https://books.google.co.id/books?id=C-dDDwAAQBAJ

[23] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. **doi:** 10.1145/2939672.2939785.

[24] I. Düntsch and G. Gediga, "Indices for rough set approximation and the application to confusion matrices," *International Journal of Approximate Reasoning*, vol. 118, pp. 155–172, Mar. 2020, **doi:** 10.1016/j.ijar.2019.12.008.

[25] J. Görtler *et al.*, "Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels," in *CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2022, pp. 1–13. **doi:** 10.1145/3491102.3501823.

[26] N. M. Kebonye, "Exploring the novel support points-based split method on a soil dataset," *Measurement*, vol. 186, p. 110131, Dec. 2021, **doi:** 10.1016/j.measurement.2021.110131.

[27] T. Sushma and V. Ramakrishnan, "Comparison of random forest classifier with XG boost classifier to classify the accuracy of flight delays," 2023, p. 020040. **doi:** 10.1063/5.0178976.

[28] Z. Xu, Y. Zhu, G. Li, and J. Yang, "Diabetes risk prediction model based on random forest and Xgboost," in *International Conference on Electronic Information Engineering and Computer Science (EIECS 2022)*, Y. Yue, Ed., SPIE, Apr. 2023, p. 22. **doi:** 10.1117/12.2668038.

[29] R. P. Quevedo *et al.*, "Consideration of spatial heterogeneity in landslide susceptibility mapping using geographical random forest model," *Geocarto International*, vol. 37, no. 25, pp. 8190–8213, Dec. 2022, **doi:** 10.1080/10106049.2021.1996637.