Performance analysis of random forest on quartile classification journal

Cornaldo Beliarding Sucahyo [a,1], Fajriwati Qoyyum Rizqini [a,2], Ayyub Naufal [a,3], Hengky Yandratama[a,4], Jabar Ash Shiddiqy [a,5], Agung Bella Putra Utama [a,6], Nastiti Susetyo Fanany Putri [b,7], Aji Prasetya Wibawa [a,8,*]

[a] Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang 65145, Indonesia
[b] Department of Information Science and Engineering, Faculty of Science and Engineering, Saga University, Saga 8408502, Japan
[1]cornaldobeliarding@gmail.com; [2] fajriwati.qoyyum.1905356@students.um.ac.id; [3] ayyub.naufal.1905356@students.um.ac.id; [4]hengky.yandratama.1905356@students.um.ac.id; [5] jabar.ash.1905356@students.um.ac.id; [6] agungbpu02@gmail.com; [7]23951801@edu.cc.saga-u.ac.jp; [8] aji.prasetya.ft@um.ac.id
* corresponding author

ARTICLE INFO

Article history
Received December 10, 2023
Revised January 14, 2024
Accepted February 10, 2024
Available online March 22, 2024

Keywords
Journal
Random forest
SCImago journal rank
CRISP-DM

ABSTRACT

Journals play a pivotal role in disseminating scientific knowledge, housing a multitude of valuable research articles. In this digital age, the evaluation of journals and their quality is essential. The SCImago Journal Rank (SJR) stands as one of the prominent platforms for ranking journals, categorizing them into five index classes: Q1, Q2, Q3, Q4, and NQ. Determining these index classes often relies on classification methodologies. This research, drawing inspiration from the Cross-Industry Standard Process for Data Mining (CRISP-DM), seeks to employ the Random Forest method to classify journals, thus contributing to the refinement of journal ranking processes. Random Forest stands out as a robust choice due to its remarkable ability to mitigate overfitting, a common challenge in machine learning classification tasks. In the context of approximating SJR index classes, Random Forest, when utilizing the Gini index, exhibits promise, albeit with an initial accuracy rate of 62.12%. The Gini index, an impurity measure, enables Random Forest to make informed decisions while classifying journals into their respective SJR index classes. However, it is worth noting that this accuracy rate represents a starting point, and further refinement and feature engineering may enhance the model's performance. This research underscores the significance of machine learning techniques in the domain of journal classification and journal-ranking systems. By harnessing the power of Random Forest, this study aims to facilitate more accurate and efficient categorization of journals, thereby aiding researchers, academics, and institutions in identifying and accessing high-quality scientific literature.

1. Introduction

Scientific journals have long served as the cornerstone of academic discourse, housing the cumulative knowledge, observations, and discoveries of researchers [1]. These journals are more than mere repositories; they are a testament to the relentless pursuit of knowledge and the rigorous scientific methodology [2]. They not only serve as references but also play a pivotal role in shaping the future course of scientific exploration [3]. In today's research landscape, the significance of high-quality journals is exemplified by their inclusion in esteemed databases like Scopus, with SCImago Journal Rank (SJR)

being a prominent player in journal quality assessment [4]. The methodology underpinning SJR quartile classification holds profound implications for the evaluation of scholarly publications.

Previous research in the field of journal evaluation has highlighted several limitations and challenges that warrant further investigation [5], [6]. One significant gap in knowledge pertains to the inability of traditional methods to effectively capture the nuanced relationships and interactions among various journal metrics [7]. These methods often rely on linear models or heuristic approaches that may oversimplify the intricate dynamics of scholarly communication. Additionally, existing classification techniques may struggle to handle the heterogeneous nature of journal data, which encompasses diverse metrics such as citations, publications, and journal prestige.

In this dynamic era of data-driven decision-making, the fusion of data mining and journal quartile classification emerges as a novel and promising approach. Classification, a fundamental task within data mining, entails predicting the category to which a given dataset belongs [8]. While data mining-driven classification [9], has permeated various domains, encompassing manufacturing [10], agriculture [11], economics [12], education [13], and healthcare [14], the adaptation of these techniques to journal quartile classification remains an underexplored frontier. The landscape of classification models presents a rich tapestry, including K-Nearest Neighbors (KNN) [15], Support Vector Machines (SVM) [16], Naïve Bayes [17], Multi-Layer Perceptron (MLP) [18], and Random Forest [19].

Compared to alternative classification methods such as SVM or Naïve Bayes, Random Forest offers several distinct advantages in the context of journal classification [20], [21]. SVM, while powerful in high-dimensional spaces, may struggle with large-scale datasets and require careful tuning of hyperparameters. Naïve Bayes, on the other hand, assumes independence among features, which may not hold true for complex journal metrics characterized by interdependencies and correlations. In contrast, Random Forest's ensemble approach and flexibility in handling diverse data types make it a robust and practical choice for SJR classification.

The selection of the Random Forest method as the primary approach for journal quartile classification stems from its unique blend of versatility, robustness, and interpretability. Unlike traditional statistical approaches that assume linear relationships or simplistic models, Random Forest excels in capturing complex interactions and nonlinear patterns within multidimensional datasets. Its ensemble learning framework, which combines multiple decision trees trained on bootstrapped subsets of the data, mitigates the risk of overfitting and enhances generalization performance [22], [23]. Moreover, Random Forest's ability to handle categorical, numerical, and mixed data types, as well as its inherent feature selection capabilities, make it well-suited for the heterogeneous nature of journal metrics and attributes [24]. Notably, each tree in the Random Forest ensemble provides an estimation error known as out-of-bag (OOB) error [25], affording valuable insights into model performance.

By leveraging the Random Forest method, this research aims to surpass the limitations of traditional journal classification approaches and unlock hidden insights within journal metrics. Through rigorous analysis and experimentation, we seek to demonstrate the superiority of Random Forest in capturing the multidimensional complexities of journal data, ultimately advancing the theoretical foundations of journal classification and laying the groundwork for future research in this domain. What sets this research apart is its unwavering commitment to scrutinize the Random Forest method's performance across distinct criterion parameters, including the Gini Index, Information Gain, and Gain Ratio [26]. This endeavor is marked by its focus on the universality of SJR classification, transcending disciplinary boundaries. The anticipated results of this study hold significant implications for the field of scholarly publishing and academic evaluation. We expect to provide fresh insights into the strengths and limitations of data mining methodologies in the context of journal classification, paving the way for more accurate and nuanced approaches to evaluating scholarly impact. Moreover, by elucidating the potential of Random Forest in SJR classification, our research may inspire further exploration into the application of machine learning techniques in academic assessment. Ultimately, our findings have the potential to inform and shape future research directions in this area, contributing to a deeper understanding of the complex dynamics underlying scientific communication and dissemination.

2. Related works

2.1. Classification

Classification, a pivotal aspect of supervised learning within data mining, offers a systematic approach to organizing diverse datasets into predefined classes [27]. It serves as a guiding principle across various domains, akin to a taxonomist's role in categorizing species based on distinct attributes. While classification empowers machines to make informed decisions by extracting hidden patterns from data, critical analysis unveils potential challenges. One such challenge lies in the dependence on labeled training data, which may not always capture the full complexity of real-world scenarios. This reliance raises questions about the scalability and adaptability of classification algorithms in dynamic environments where labeled data may be scarce or outdated.

The journey of classification unfolds through the interplay between training and testing data, where models learn from past experiences and refine their predictive capabilities [28]. While classification finds extensive applications in email filtering, medical diagnosis, and fraud detection, its effectiveness hinges on the quality and representativeness of the training data [29]. Synthesizing existing literature reveals ongoing debates regarding the interpretability of classification models. While these models excel in accuracy and predictive power, understanding the rationale behind their decisions remains a challenge. Addressing this gap requires further exploration into techniques for enhancing model interpretability without compromising performance, thus bridging the divide between theoretical understanding and practical application in classification tasks.

2.2. Random Forest

The evolution of the "Random Forest" (RF) method, pioneered by Tin Kam Ho and further developed by Leo Breiman, represents a significant milestone in data science [30]. RF leverages Bagging and random feature selection to create an ensemble of decision trees, offering robustness and resilience to noise and overfitting [31], [32]. While the ensemble approach enhances predictive performance, critical analysis reveals potential trade-offs between model complexity and interpretability. The intricate nature of RF models poses challenges in understanding and explaining their decision-making processes, raising concerns about their applicability in domains requiring transparent and interpretable models.

Criteria such as Information Gain, Gini Index, and Gain Ratio guide the construction of decision trees within the RF framework, facilitating informed splitting decisions [33]. But this is not the extent of its magic; RF also wields a trio of criteria: the enigmatic Information Gain, the insightful Gain Ratio, and the formidable Gini Index [26]. These criteria stand as the gatekeepers, guiding the formation of the decision tree's root node and branching rules, much like a trio of guardians overseeing the growth of a majestic tree. The Information Gain criterion, in particular, ascends to its zenith when it identifies the attribute that imparts the most knowledge, serving as the foundation upon which the tree's wisdom is rooted. The calculation of Information Gain as in (1).

$$Information\ Gain = Entropy_{parent} - Entropy_{children} \quad (1)$$

Meanwhile, the Gini Index is obtained by calculating the squared probability of each class in the dataset. The equation of Gini Index as in (2).

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \quad (2)$$

Then, the gain ratio is obtained from the ratio of the information gain and split information values by selecting the lowest subset of each attribute. Gain Ratio calculation as in (3).

$$Gain\ Ration\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo(S,A)} \quad (3)$$

However, synthesizing existing literature sheds light on the need for deeper exploration into the interpretability of RF models. While these criteria optimize predictive accuracy, their implications for model transparency and explainability warrant further investigation. Addressing these concerns can

enhance the trustworthiness and adoption of RF models across diverse domains, fostering a deeper understanding of their strengths and limitations in real-world applications.

2.3. Cross Validation (K-Fold)

K-fold cross-validation stands as a quintessential evaluation technique, a bedrock upon which the edifice of machine learning model assessment is built [34]. Its overarching aim? To provide an insightful estimation of a model's performance accuracy, an invaluable metric for data scientists and practitioners seeking to understand the true potential of their algorithms [35]. Like a seasoned conductor orchestrating a symphony, K-fold cross-validation extracts the full melodic range of a dataset, allowing the model to train and perform against various slices of the data, thus unveiling its robustness and adaptability.

In the grand tableau of machine learning, the essence of K-fold cross-validation takes form in Figure 1, a visual representation that encapsulates the iterative process of model training and validation. Each fold, like a distinct movement in a symphony, represents a unique segment of the dataset, with the model performing a graceful dance across these segments. The harmonious collaboration of training and testing phases, as depicted in Fig. 1, paints a vivid picture of how K-fold cross-validation encapsulates the essence of model evaluation, forging a path towards a comprehensive understanding of a model's capabilities. Through this cyclic process, K-fold cross-validation bestows upon data scientists a refined appreciation of their models, allowing them to fine-tune their algorithms and strike the right chord in the ever-evolving symphony of machine learning.

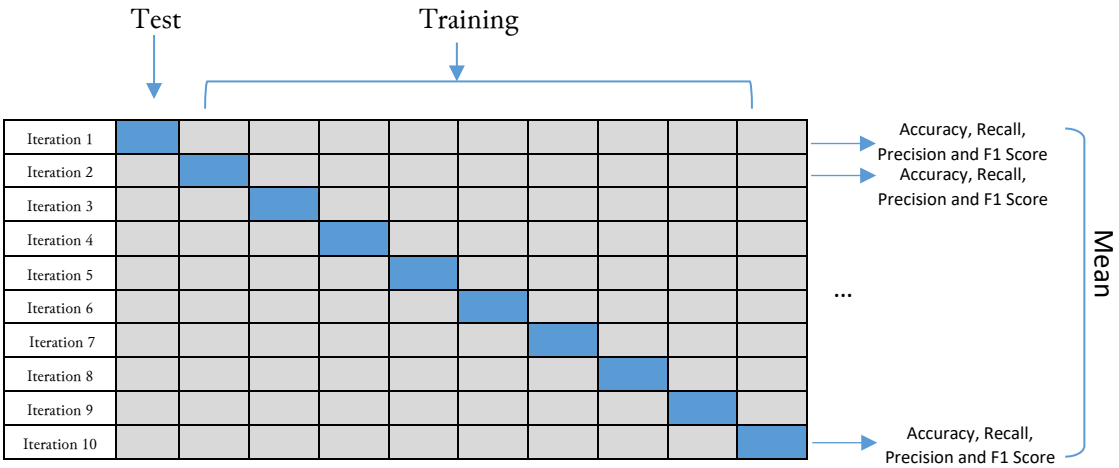

Fig. 1. Cross-validation

However, critical synthesis suggests exploring potential biases in certain data distributions or the impact of hyperparameters on cross-validation results. Identifying such nuances could enrich the understanding of a model's true capabilities.

3. Method

The Cross Industry Standard Process for Data Mining (CRISP-DM) research method is chosen. CRISP-DM is a stage that focuses on research on data mining [36]. The CRISP-DM method has six stages of the research flow listed in Fig. 2.

3.1. Business Understanding

The preliminary stage of business understanding serves as the compass guiding the data-driven journey, with the primary goal of discerning and addressing the needs and objectives from a business perspective [37]. In the context of our study, this phase takes on particular significance as we delve into the realm of SCImago Journal Rank (SJR) and its unique SJR classification system. Our quest for insights begins with the extraction of data from SCImago Journal Rank, explicitly focusing on journals

spanning the year 2020. Within the landscape of scholarly journals, SJR classification wields a distinct influence, acting as a lodestar for academia. In our pursuit, this classification system becomes the focal point, as we navigate through the labyrinthine corridors of data to uncover patterns, trends, and invaluable intelligence. By tapping into the wealth of data within SJR, encompassing the diverse tapestry of journals, we aim to illuminate the landscape of scholarly publishing for the year 2020. Through the lens of business understanding, our mission crystallizes: to extract actionable insights from this data trove, shedding light on the dynamics and nuances that underpin the world of academic journals in the contemporary era.

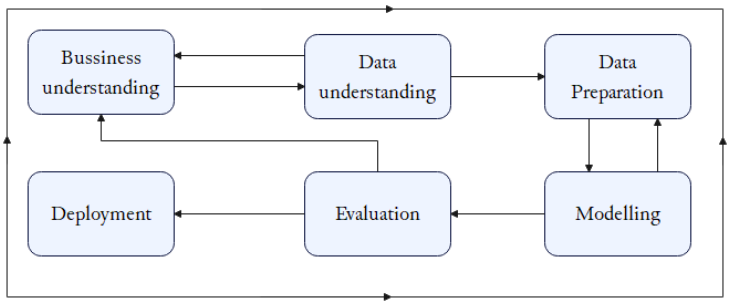

Fig. 2. CRIPS-DM Research Model

3.2. Data Understanding

In the intricate tapestry of our data-driven journey, the data understanding stage emerges as the vital bridge that spans the chasm between raw data and actionable insights [38]. This pivotal phase is dedicated to the meticulous collection and comprehensive assessment of data quality, laying the foundation upon which our analysis shall rest. The crux of this stage is to foster a profound understanding of the dataset in all its intricate detail, unfurling the rich narrative it carries. The dataset in our possession presents itself as a mosaic composed of 20 distinct attributes, each possessing its unique data type, weaving a tapestry of complexity and diversity. This mosaic encompasses a grand total of 26,497 instances, a vibrant and dynamic tapestry that mirrors the rich ecosystem of scholarly journals. Within this expansive landscape, we encounter the quintessential quartile classification, represented by Q1, Q2, Q3, and Q4, each a distinct facet in the scholarly gemstone. Yet, amidst these quartile luminaries, we also find the intriguing presence of NQ (Non-Quartile), a category that beckons us to explore its unique characteristics. To unveil the secrets embedded within this dataset, we turn our gaze to the meticulous documentation encapsulated in Table 1 and Table 2.

Table 1. Dataset Attribute Details

Attributes	Data Type	Description
Rank	**Integer**	**1 – 24978**
SJR	Integer	100 – 62937
SJR Best Quartile	Nominal	Q1, Q2, Q3, Q4, NQ
H Index	Integer	0 – 1226
Total Docs. (2017)	Integer	0 – 21801
Total Docs. (3years)	Integer	1 – 61528
Total Refs	Integer	0 – 1033089
Total Cites (3years)	Integer	0 – 282734
Citable Docs.(3years)	Integer	1 – 61524
Cites/Doc.(2years)	Real	0 – 126340
Ref. / Doc	Real	0 – 859500
Source	Real	12001-19600157914
Title	Nominal	2D Materials, 3 Biotech, etc
Type	Nominal	Journal
Issn	Nominal	12343, 12610, etc
Country	Nominal	Albania, Argentina, etc
Region	Nominal	Africa, Africa/Middle East
Publisher	Nominal	Association pour la Diffusion de la Recherche lilteraire, etc
Coverage	Nominal	2015, 2017, etc
Categories	Nominal	Accounting (Q1), etc

Table 1 and Table 2 serve as treasure maps, guiding us through the labyrinthine dataset, revealing the attributes and their diverse data types. As we traverse this landscape, we shall unearth hidden patterns, discern valuable insights, and ultimately, harness the power of data to shed light on the intricate world of journal classification. In this endeavor, the data understanding stage is the lantern that illuminates our path, ensuring that we tread with clarity and purpose on our journey of discovery. The value of NQ makes the percentage unbalanced in each class label. The smallest class is NQ, with a percentage of 5.73%, and the highest is Q1, with a value of 28.61%

Table 2. Number of SJR 2020

Category	Number of Instances	Percentage
Q1	7580	28,61%
Q2	6411	24,20%
Q3	5829	21,99%
Q4	5158	19,47%
NQ	1519	5,73%
Total	26.497	100%

3.3. Data Preparation

The data preparation stage emerges as the forge where raw data is refined and shaped into a form suitable for analysis, a crucial juncture in our data-driven odyssey [39]. This stage is aptly described as the preprocessing stage, where the art of data refinement takes center stage. Within the realm of preprocessing, we wield four potent methods, each a distinct tool in our arsenal: data cleaning [40], data transformation [41], data reduction [42], and data integration [43]. However, in the context of this research, our focus narrows to the skilled application of two specific methods—data cleaning and data transformation.

3.3.1. Data Cleaning

Data cleaning, the first pillar of our preparation journey, is akin to the meticulous restoration of a priceless artifact. It involves the art of identifying and rectifying inaccuracies, anomalies, and inconsistencies that may tarnish the purity of our dataset. Through this process, we endeavor to ensure that our data gleams with accuracy, laying a solid foundation for our analysis. The data cleaning process removes missing values in the dataset [44]. From the data, we obtained 26497, with details in Table 3. Table 3 explains five classes in the SJR dataset, namely Q1, Q2, Q3, Q4 and NQ. This research categorizes journals with quartile values so that the NQ class label can be removed.

Table 3. Number of SJR 2020 without NQ

Category	Number of Instances	Percentage without NQ
Q1	7580	30.35%
Q2	6411	25.67%
Q3	5829	23.34%
Q4	5158	20.65%
Total	26.497	100% (24.978)

The removed NQ value can increase the percentage of each class label. The largest class is Q1, with a value of 30.35%, and the lowest is Q4, with a value of 20.65%. From the percentage range, it can be assumed that data cleaning on NQ class labels can make the data balanced to eliminate the re-sampling process.

3.3.2. Data Transformation

Data transformation, is the alchemical process through which we transmute raw data into a more refined and insightful form [45]. Like a sculptor shaping a block of marble into a masterpiece, data transformation allows us to extract relevant features, encode categorical variables, and normalize distributions. Through this metamorphosis, we aim to unlock the latent potential within our dataset,

enabling it to reveal its secrets and nuances with greater clarity. Data transformation is used to modify data modifications made using attribute selection and normalization. Of the 20 attributes, only nine attributes are used. These attributes were chosen because theywill be displayed on the old SCImago Journal Rank for normalization using min-max normalization. Min-max normalization was chosen because it matches the Random Forest method and dataset [46]. Selected attributes show as Table 4.

Table 4. Details of Selected Attributes

Attributes	Data Type	Description
SJR Best Quartile	**Nominal**	**Q1, Q2, Q3, Q4**
H Index	Real	0 – 1
Total Docs. (2017)	Real	0 – 1
Total Docs. (3years)	Real	0 – 1
Total Refs	Real	0 – 1
Total Cites (3years)	Real	0 – 1
Citable Docs.(3years)	Real	0 – 1
Cites/Doc.(2years)	Real	0 – 1
Ref. / Doc	Real	0 – 1

The yellow color means the attribute class label. The class attribute is the target class used as the quartile class in each journal. The numeric data type is still used because the Random Forest classification model can process it.

3.4. Modeling

The modeling stage creates a model for this research [47]. The model used is the Random Forest method. Random Forest is a development model of a decision tree with ensemble (bagging) techniques [48]. The modeling framework is shown in Fig. 3.

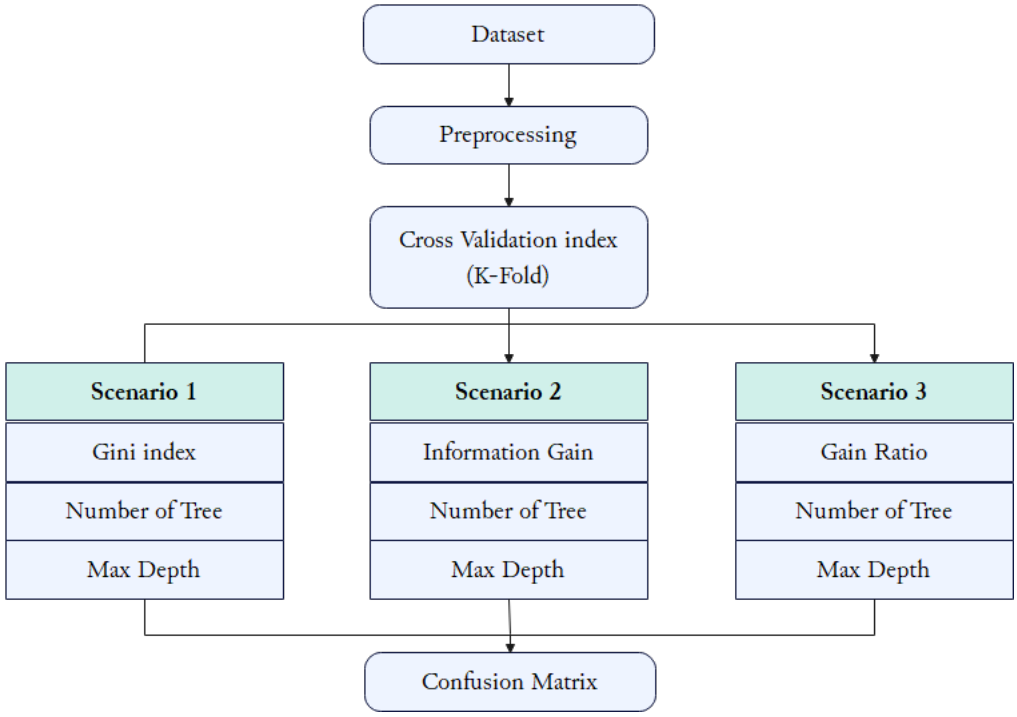

Fig. 3. Research framework

Three scenarios will be applied to the Random Forest algorithm. Each scenario will use criterion parameters, number of trees, and max depth. Scenario 1 contains the criterion gini index, scenario 2 contains information gain, and scenario 3 contains the gain ratio. Pseudocode of the Random Forest scenariois shown in pseudocode 1 (Fig. 4).

```
Start
    Dataset input Preprocessing
    Cross Validation (K-Fold)Random Forest Classifier
    Show Confusion matrix, Accuracy, Precision,Recall, and F1 Score
End
```

Fig. 4. Random forest

The pseudocode conveys the classification process with Random Forest in general. The process will begin with the input dataset and then proceed with preprocessing. Then, proceed with the model with scenario1, scenario 2, and scenario 3, changing the criterion, number of trees, and max depth. Then, the results willbe displayed as a confusion matrix. Details of the type of criterion, Number of trees, and Max_Depth are in Table 5.

Table 5. NoT and Max depth values

Scenario		
Criterion	Criterion	Criterion
Gini_Index	100	5
Information_Gain	150	6
Gain ratio	200	7
	250	8
		9

3.5. Evaluation

The evaluation stage is used to show the validation results of a study [49]. This stage displays the confusion matrix value. Confusion matrix is a tool to analyze the method's performance [50]. The results obtained from CM are statistical values: accuracy, recall,precision, and f1 score. The description of CM can be seen in Table 6.

Table 6. Confusion Matrix

	Pre Positive	Pre Negative
Positive Act	TP (True Positive)	FN (False Negative)
Negative Act	FP (False Positive)	TN (True Negative)

where TP is the amount of true data with the truth value being true, TN is the number of false-valued data whose truth value is true, FP is the amount of data that is true with a truth value that is false, and FN is the number of false-valued data whose truth value is false.

Table 6 produces accuracy, precision-recall, and f1 score values. Accuracy is the percentage of correctness of a model [51]. Recall is a level of sensitivity showing success in retrieving information [52]. Precision is the prediction of the total positive ratio [53]. The F1 score is the ratio of precision and recall [54]. The equation of accuracy, recall, precision, and F1 Score respectively as in (4) to (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{5}$$

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

$$F1\ Score = \frac{2\ x\ (Recall+Precision)}{(Recall+Precision)} \tag{7}$$

3.6. Deployment

The deployment stage is the final stage of the CRISP-DM process. This stage contains a report on the evaluation results of the research [55]. The results are presented as a confusion matrix, and the result values are in the form of accuracy, recall, precision, and f1 score. The evaluation value is obtained on the test results in each criterion: gini index, information gain, and gain ratio. Each criterion has several tree parameters with four combinations and is then supported by the max depth parameter with five combinations.

4. Results and Discussion

This research uses the Random Forest method as the primary model for classification testing in analyzing the performance of Random Forest using the criterion Gini index, information gain, and gain ratio. Each criterion will use a combination of several trees and max depth. The research results are shown in Table 7.

Table 7. Evaluation Results

Criterion	NoT	Max depth	Accuracy (%)	Recall (%)	Precision (%)	F1-Score(%)
Gini index	100	5	57.83	58.57	57.40	57.98
		6	58.40	59.03	57.83	58.42
		7	59.48	60.13	59.19	59.66
		8	60.53	60.98	60.08	60.53
		9	**62.12**	**62.47**	**61.78**	**62.12**
	150	5	57.89	58.61	57.42	58.01
		6	58.42	59.03	57.84	58.43
		7	59.35	59.98	58.97	59.47
		8	60.53	60.97	60.07	60.52
		9	62.08	62.45	61.76	62.10
	200	5	57.86	58.60	57.40	57.99
		6	58.47	59.12	57.97	58.54
		7	59.44	60.08	59.12	59.60
		8	60.63	61.06	60.22	60.64
		9	62.01	62.38	61.71	62.04
	250	5	57.97	58.72	57.57	58.14
		6	58.60	59.62	58.15	58.88
		7	59.54	60.18	59.19	59.68
		8	60.69	61.11	60.29	60.70
		9	61.99	62.37	61.69	62.03
Information gain	100	5	57.67	58.31	56.81	57.55
		6	58.37	58.92	57.52	58.21
		7	58.89	59.38	58.25	58.81
		8	60.23	60.59	59.6	60.09
		9	61.35	61.68	60.7	61.19
	150	5	57.54	58.24	56.86	57.54
		6	58.19	58.78	57.48	58.12
		7	58.87	59.41	58.21	58.80
		8	60.10	60.53	59.49	60.01
		9	61.45	61.80	60.86	61.33
	200	5	57.54	58.26	56.93	57.59
		6	58.23	58.84	57.56	58.19
		7	58.86	59.38	58.21	58.79
		8	60.16	60.59	59.57	60.08
		9	61.54	61.88	61.01	61.44
	250	5	57.58	58.29	56.96	57.62
		6	58.29	58.91	57.64	58.27
		7	58.86	59.38	58.19	58.78
		8	60.11	60.54	59.49	60.01
		9	**61.55**	**61.89**	**60.99**	**61.44**

Criterion	NoT	Max depth	Accuracy (%)	Recall (%)	Precision (%)	F1-Score(%)
Gain ratio	100	5	56.29	56.53	55.43	55.97
		6	56.08	56.71	55.16	55.92
		7	56.40	56.97	55.59	56.27
		8	56.02	56.53	55.55	56.04
		9	56.55	57.11	55.99	56.54
	150	5	56.60	57.16	55.44	56.29
		6	56.17	56.80	55.21	55.99
		7	56.45	57.02	55.63	56.32
		8	55.94	56.45	55.36	55.90
		9	56.19	56.73	55.61	56.16
	200	5	54.26	53.94	55.78	54.84
		6	56.79	57.40	55.76	56.57
		7	56.49	57.04	55.71	56.37
		8	56.00	56.54	55.44	55.98
		9	56.22	56.78	55.71	56.24
	250	5	53.68	53.05	56.27	54.61
		6	**56.93**	**57.34**	**55.82**	**56.57**
		7	56.39	56.93	55.63	56.27
		8	55.91	56.44	55.44	55.94
		9	56.37	56.92	55.92	56.42

Table 7 shows that the best value on the criterion gini index is at several trees 100 withmax depth 9. The values obtained on the criterion gini index are accuracy 62.12%, recall 62.47%, precision 61.78%, and f1 score 62.12%. At the same time, the lowest performance is at several trees 100 max depth 5. When added, the number of trees on the gini index can decrease the accuracy value, as shown in Fig. 5. This is because adding the value of the number of trees increases the value of the gini impurity so that it approaches 1 [56].

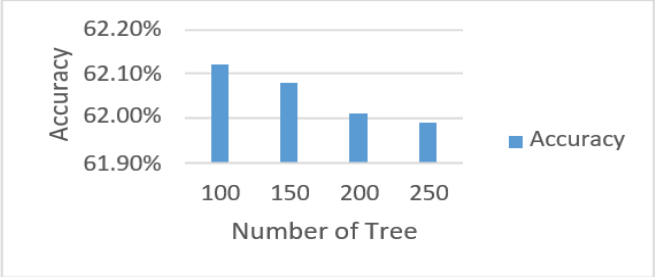

Fig. 5. Accuracy Change of Number of Tree Value

Table 7 conveys that criterion information gain obtained the best value at several trees 250 with max depth 9: accuracy 61.55%, recall 61.89%, precision 60.99%, and F1 Score 61.44%. The lowest performance is several trees 150 with a max depth of 5. Adding several trees in criterion information gain causes an increase in accuracy value. Adding the number of tree valuesmakes the entropy value low, making it better at separating information [57]. The increase in accuracy value can be seen in Fig. 6.

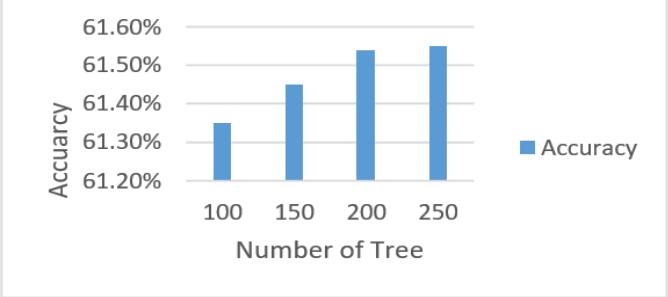

Fig. 6. Accuracy Change of Number of Tree Value

Table 7 conveys the optimal results on the gain ratio criterion at several trees 250 with max depth 6. The performance values obtained are accuracy 56.93%, recall 57.34%, precision 55.82%, and f1 score 56.57%. Unlike the previous criterion, the gini index criterion on each number of trees and max depth obtained varying values. This is due to the uneven ratio value with split information, so the number of trees and max depth obtain different values [58]. Fig. 7 conveys the difference in accuracy values at each number of trees and max depth.

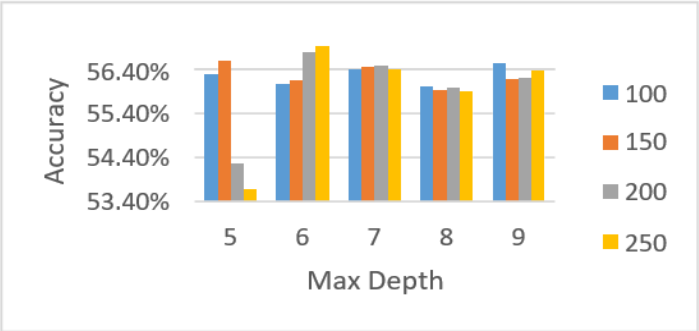

Fig. 7. Accuracy Change of Number of Tree Value

From the Gini index criterion, information gain and gain ratio, the best performance is on the Gini index criterion with an accuracy value of 62.12%. This is because the advantage of the Gini index performance is that it focuses on selecting the subset with the lowest value from the dataset that is partitioned to be purer so that the separation of the target class becomes more efficient [59]. Other research also supports this by conveying the criterion of the Gini index obtaining the highest value [60]. In addition, other research conveyed that from the criterion gini index,information gain, and gain ratio, the gini index obtained the highest value [61]. These results cannot be separated from the weaknesses of information gain, namely that bias can occur if too much data is used [62]. However, it cannot be separated from the weakness of the gain ratio, namely, the more complex thedata results in overfitting and poor generalized data [63]. Overall, this analysis provides valuable insights into the performance of Random Forest using different criteria, emphasizing the strengths and weaknesses of each criterion in the context of journal quartile classification. These findings can inform the selection of an appropriate criterion for specific classification tasks and contribute to a more nuanced understanding of their impact on model performance.

From the results obtained, it is essential to note that the Random Forest method, although powerful and versatile, is not immune to specific challenges. One limitation is the potential complexity of the model, especially when using large numbers of trees and deep trees. While increasing the number of trees and tree depth can improve model performance, it also increases computational complexity and the risk of overfitting, especially with limited training data. Additionally, the Random Forest algorithm may not capture subtle nonlinear relationships or interactions present in the data, which may impact its ability to accurately classify journals based on SJR. Additionally, Random Forest performance may be sensitive to the choice of hyperparameters, such as the number of features considered at each split and the minimum samples required for leaf nodes, which may require careful tuning to optimize model performance. These methodological limitations underscore the importance of rigorously validating model results and considering alternative approaches to ensure robust and reliable classification results.

5. Conclusion

In conclusion, the results of our tests have led to several significant findings. Firstly, it is evident that the Gini index criterion outperformed both the information gain and gain ratio criteria in the context of classifying journal quartiles. This assertion is strongly supported by the extensive analysis presented in the evaluation results, which meticulously compares the performance of different criteria across various combinations of the number of trees and maximum depth. Specifically, the Gini index consistently exhibits higher accuracy, recall, precision, and F1-score values compared to alternative criteria, such as

information gain and gain ratio. This underscores the importance of selecting the proper criterion for decision tree-based algorithms, with the Gini index proving to be a superior choice in this instance. Moreover, the optimal accuracy of 62.12% achieved by the Random Forest method highlights its potential for enhancing the classification of journal quartiles. However, this achievement also suggests the need for further exploration and refinement. For future research, we recommend a more granular approach by separating data into distinct domains rather than amalgamating them. This can potentially yield more precise insights and improve classification outcomes. Additionally, the success of the Gini index criterion prompts us to encourage researchers to explore its use in tandem with appropriate model tuning, as it could further enhance the performance of Random Forest for similar datasets.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] A. P. Wibawa, A. C. Kurniawan, H. A. Rosyid, and A. M. M. Salah, "International Journal Quartile Classification Using the K-Nearest Neighbor Method," in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Oct. 2019, pp. 336–341, doi: 10.1109/ICEEIE47180.2019.8981413.

[2] K. D. S. Mendes, R. C. de C. P. Silveira, and C. M. Galvão, "Use of the bibliographic reference manager in the selection of primary studies in integrative reviews," *Texto Context. - Enferm.*, vol. 28, 2019, doi: 10.1590/1980-265x-tce-2017-0204.

[3] M. Hennink and B. N. Kaiser, "Sample sizes for saturation in qualitative research: A systematic review of empirical tests," *Soc. Sci. Med.*, vol. 292, p. 114523, Jan. 2022, doi: 10.1016/j.socscimed.2021.114523.

[4] I. Ianoş and A.-I. Petrişor, "An Overview of the Dynamics of Relative Research Performance in Central-Eastern Europe Using a Ranking-Based Analysis Derived from SCImago Data," *Publications*, vol. 8, no. 3, p. 36, Jul. 2020, doi: 10.3390/publications8030036.

[5] P. dwi Nurfadila, A. P. Wibawa, I. A. E. Zaeni, and A. Nafalski, "Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal," *Int. J. Artif. Intell. Res.*, vol. 3, no. 2, Jul. 2019, doi: 10.29099/ijair.v3i2.99.

[6] R. P. Adiperkasa, A. P. Wibawa, I. A. E. Zaeni, and T. Widiyaningtyas, "International Reputable Journal Classification Using Inter-correlated Naïve Bayes Classifier," in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, Sep. 2019, pp. 49–52, doi: 10.1109/IC2IE47452.2019.8940887.

[7] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.

[8] J. N.P. and R. Aruna, "Big data analytics in health care by data mining and classification techniques," *ICT Express*, vol. 8, no. 2, pp. 250–257, Jun. 2022, doi: 10.1016/j.icte.2021.07.001.

[9] K. Trang and A. H. Nguyen, "A Comparative Study of Machine Learning-based Approach for Network Traffic Classification," *Knowl. Eng. Data Sci.*, vol. 4, no. 2, p. 128, Jan. 2022, doi: 10.17977/um018v4i22021p128-137.

[10] R. Snell *et al.*, "Methods for Rapid Pore Classification in Metal Additive Manufacturing," *JOM*, vol. 72, no. 1, pp. 101–109, Jan. 2020, doi: 10.1007/s11837-019-03761-9.

[11] N. S. Pangaribuan and F. Marpaung, "Analysis of Corn Agriculture Data to Predict Harvest Results with Data Mining Algorithm C4. 5," *Login J. Teknol.*, vol. 14, no. 2, pp. 235–243, 2020. [Online]. Available at: https://login.seaninstitute.org/index.php/Login/article/view/53.

[12] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan, "Mining in Educational Data: Review and Future Directions," 2020, pp. 92–102, doi: 10.1007/978-3-030-44289-7_9.

[13] M. A. Ledhem, "Data mining techniques for predicting the financial performance of Islamic banking in Indonesia," *J. Model. Manag.*, vol. 17, no. 3, pp. 896–915, Aug. 2022, doi: 10.1108/JM2-10-2020-0286.

[14] O. Almadani and R. Alshammari, "Prediction of Stroke using Data Mining Classification Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, 2018, doi: 10.14569/IJACSA.2018.090163.

[15] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, pp. 234–242, May 2020, doi: 10.1016/j.neucom.2018.11.101.

[16] G. M. Borkar, L. H. Patil, D. Dalgade, and A. Hutke, "A novel clustering approach and adaptive SVM classifier for intrusion detection in WSN: A data mining concept," *Sustain. Comput. Informatics Syst.*, vol. 23, pp. 120–135, Sep. 2019, doi: 10.1016/j.suscom.2019.06.002.

[17] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2018, pp. 301–306, doi: 10.1109/CSCI46756.2018.00065.

[18] S. Mishra, H. K. Tripathy, P. K. Mallick, A. K. Bhoi, and P. Barsocchi, "EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis," *Sensors*, vol. 20, no. 14, p. 4036, Jul. 2020, doi: 10.3390/s20144036.

[19] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, Jun. 2021, doi: 10.26599/BDMA.2020.9020016.

[20] C. Bogdal, R. Schellenberg, O. Höpli, M. Bovens, and M. Lory, "Recognition of gasoline in fire debris using machine learning: Part I, application of random forest, gradient boosting, support vector machine, and naïve bayes," *Forensic Sci. Int.*, vol. 331, p. 111146, Feb. 2022, doi: 10.1016/j.forsciint.2021.111146.

[21] M. Radak, H. Y. Lafta, and H. Fallahi, "Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 12, pp. 10473–10491, Sep. 2023, doi: 10.1007/s00432-023-04956-z.

[22] Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," 2020, p. 080020, doi: 10.1063/1.4977376.

[23] M.-J. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 11, pp. 2149–2167, Nov. 2021, doi: 10.1080/13658816.2021.1887490.

[24] N. Singh and P. Singh, "A novel Bagged Naïve Bayes-Decision Tree approach for multi-class classification problems," *J. Intell. Fuzzy Syst.*, vol. 36, no. 3, pp. 2261–2271, Mar. 2019, doi: 10.3233/JIFS-169937.

[25] B. Ramosaj and M. Pauly, "Consistent estimation of residual variance with random forest Out-Of-Bag errors," *Stat. Probab. Lett.*, vol. 151, pp. 49–57, Aug. 2019, doi: 10.1016/j.spl.2019.03.017.

[26] R. R. Putra and H. W. Dhany, "Determination of accuracy value in id3 algorithm with gini index and gain ratio with minimum size for split, minimum leaf size, and minimum gain," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, p. 012088, Jan. 2020, doi: 10.1088/1757-899X/725/1/012088.

[27] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[28] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.

[29] I. Iddrisu, P. Appiahene, O. Appiah, and I. Fuseini, "Exploring the Impact of Students Demographic Attributes on Performance Prediction through Binary Classification in the KDP Model," *Knowl. Eng. Data Sci.*, vol. 6, no. 1, pp. 24–40, 2023, doi: 10.17977/um018v6i12023p24-40.

[30] V. K. Pandey, K. K. Sharma, H. R. Pourghasemi, and S. K. Bandooni, "Sedimentological characteristics and application of machine learning techniques for landslide susceptibility modelling along the highway corridor

Nahan to Rajgarh (Himachal Pradesh), India," *CATENA*, vol. 182, p. 104150, Nov. 2019, doi: 10.1016/j.catena.2019.104150.

[31] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, p. 104089, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104089.

[32] M. Zounemat-Kermani, D. Stephan, M. Barjenbruch, and R. Hinkelmann, "Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models," *Adv. Eng. Informatics*, vol. 43, p. 101030, Jan. 2020, doi: 10.1016/j.aei.2019.101030.

[33] A. Vrisna, H. Ar, M. Yasser, and S. Nazir, "Optimizing Random Forest Algorithm to Classify Player ' s Memorisation via In-game Data," *Knowl. Eng. Data Sci.*, vol. 6, no. 1, pp. 103–113, 2023, doi: 10.17977/um018v6i12023p103-113.

[34] T.-T. Wong and P.-Y. Yeh, "Reliable Accuracy Estimates from k -Fold Cross Validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.

[35] M. A. Khan *et al.*, "Geopolymer Concrete Compressive Strength via Artificial Neural Network, Adaptive Neuro Fuzzy Interface System, and Gene Expression Programming With K-Fold Cross Validation," *Front. Mater.*, vol. 8, May 2021, doi: 10.3389/fmats.2021.621163.

[36] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[37] M. I. Zulfa, A. Fadli, and Y. Ramadhani, "Classification model for graduation on time study using data mining techniques with SVM algorithm," 2019, p. 020006, doi: 10.1063/1.5097475.

[38] J. F. Pinto da Costa and M. Cabral, "Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works," *Mathematics*, vol. 10, no. 6, p. 993, Mar. 2022, doi: 10.3390/math10060993.

[39] K. Rahayu, L. Novianti, and M. Kusnandar, "Implementation Data Mining With K-Means Algorithm For Clustering Distribution Rabies Case Area In Palembang City," *J. Phys. Conf. Ser.*, vol. 1500, no. 1, p. 012121, Apr. 2020, doi: 10.1088/1742-6596/1500/1/012121.

[40] A. Tawakuli, D. Kaiser, and T. Engel, "Transforming IoT Data Preprocessing," in *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*, Nov. 2022, pp. 1083–1088, doi: 10.1145/3560905.3567762.

[41] C. V. Gonzalez Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, Apr. 2019, pp. 2086–2090, doi: 10.1109/ICDE.2019.00245.

[42] H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," *Comput. Methods Programs Biomed.*, vol. 195, p. 105635, 2020, doi: 10.1016/j.cmpb.2020.105635.

[43] M. D. Luecken *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nat. Methods*, vol. 19, no. 1, pp. 41–50, Jan. 2022, doi: 10.1038/s41592-021-01336-8.

[44] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review," *J. Healthc. Eng.*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/4302425.

[45] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, no. 1, p. 16884, Nov. 2019, doi: 10.1038/s41598-019-52737-x.

[46] U. Ayub and S. A. Moqurrab, "Predicting crop diseases using data mining approaches: Classification," in *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG)*, Apr. 2018, pp. 1–6, doi: 10.1109/ICPESG.2018.8384523.

[47] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, and J. R. Coronado-Hernández, "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test," *Procedia Comput. Sci.*, vol. 198, pp. 512–517, 2022, doi: 10.1016/j.procs.2021.12.278.

[48] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Min. Knowl. Discov.*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1249.

[49] T. Darmawan, "Credit Classification Using CRISP-DM Method On Bank ABC Customers," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2375–2380, Jun. 2020, doi: 10.30534/ijeter/2020/28862020.

[50] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *J. Intell. Comput. Heal. Informatics*, vol. 1, no. 1, p. 17, Mar. 2020, doi: 10.26714/jichi.v1i1.5570.

[51] S. Li, S. Ning, Y. Yezhou, T. Jingjing, Z. Wenxue, and C. Liang, "Application of Data Mining Technology in the Recall of Defective Automobile Products in China ——A Typical Case of the Construction of Digital China," in *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*, Nov. 2019, pp. 541–545, doi: 10.1109/IICSPI48186.2019.9095877.

[52] W. Li, "Big Data Precision Marketing Approach under IoT Cloud Platform Information Mining," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Jan. 2022, doi: 10.1155/2022/4828108.

[53] A. Yulianto, P. Sukarno, and N. A. Suwastika, "Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset," *J. Phys. Conf. Ser.*, vol. 1192, p. 012018, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012018.

[54] J. Yu, A. W. Schumann, Z. Cao, S. M. Sharpe, and N. S. Boyd, "Weed Detection in Perennial Ryegrass With Deep Learning Convolutional Neural Network," *Front. Plant Sci.*, vol. 10, Oct. 2019, doi: 10.3389/fpls.2019.01422.

[55] N. Tanjung, D. Irmayani, and V. Sihombing, "Implementation of C5.0 Algorithm for Prediction of Student Learning Graduation in Computer System Architecture Subjects," *Sinkron*, vol. 7, no. 1, pp. 274–280, Feb. 2022, doi: 10.33395/sinkron.v7i1.11259.

[56] K. Devi and S. Ratnoo, "Predicting student dropouts using random forest," *J. Stat. Manag. Syst.*, vol. 25, no. 7, pp. 1579–1590, Oct. 2022, doi: 10.1080/09720510.2022.2130570.

[57] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," *Procedia Manuf.*, vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.

[58] T. Berhane *et al.*, "Decision-Tree, Rule-Based, and Random Forest Classification of High-Resolution Multispectral Imagery for Wetland Mapping and Inventory," *Remote Sens.*, vol. 10, no. 4, p. 580, Apr. 2018, doi: 10.3390/rs10040580.

[59] Y. Liu and J. L. Gastwirth, "On the capacity of the Gini index to represent income distributions," *METRON*, vol. 78, no. 1, pp. 61–69, Apr. 2020, doi: 10.1007/s40300-020-00164-8.

[60] A. N. Iman and T. Ahmad, "Improving Intrusion Detection System by Estimating Parameters of Random Forest in Boruta," in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, Feb. 2020, pp. 1–6, doi: 10.1109/ICoSTA48221.2020.1570609975.

[61] D. Liu *et al.*, "Optimisation and evaluation of the random forest model in the efficacy prediction of chemoradiotherapy for advanced cervical cancer based on radiomics signature from high-resolution T2 weighted images," *Arch. Gynecol. Obstet.*, vol. 303, no. 3, pp. 811–820, Mar. 2021, doi: 10.1007/s00404-020-05908-5.

[62] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 199, Dec. 2020, doi: 10.1186/s12874-020-01080-1.

[63] V.-H. Nhu *et al.*, "Shallow Landslide Susceptibility Mapping by Random Forest Base Classifier and Its Ensembles in a Semi-Arid Region of Iran," *Forests*, vol. 11, no. 4, p. 421, Apr. 2020, doi: 10.3390/f11040421.