 ISSN 2723-6471

A Comparative Study on Data Collection Methods: Investigating Optimal Datasets for Data Mining Analysis

Hendra Jatnika[1,*], Ari Waluyo[2], Abdul Azis[3]

[1] *Informatics Engineering, Faculty of Telematics Energy, Institut Teknologi PLN, Jakarta 11750, Indonesia*

[2]*Electronics Engineering, Piksi Ganesha Indonesia Polytechnic, Kebumen 54311, Indonesia*

[3] *Information System, AmikomPurwokerto University, Purwokerto 53127, Indonesia*

(Received: November 23, 2023; Revised: December 8, 2023; Accepted: January 14, 2024; Available online: January 29, 2024)

Abstract

This study is dedicated to evaluating the efficiency of diverse data collection methods in obtaining optimal data for computational data mining. The investigation meticulously compares the questionnaire and web mining methodologies within the framework of SVM and NBC algorithms to discern the flexibility inherent in each data type. The outcomes of this comprehensive analysis demonstrate that questionnaires showcase remarkable flexibility, exhibiting accuracy rates surpassing 80% in both algorithms, along with AUC values exceeding 0.9 when contrasted with data acquired through web mining techniques. These results underscore the paramount importance of the dataset collection method in the realm of computational data mining. The study contributes compelling evidence that advocates for the superiority of the questionnaire data collection method over web mining in the specific context of computational data mining. The questionnaire method not only outperforms in terms of flexibility but also achieves high accuracy, making it a more reliable choice for acquiring data in this domain. Beyond its practical implications, the research highlights a critical aspect of methodology in data collection by emphasizing the necessity of exploring and assessing methods that may have been overlooked in previous research endeavors. This underscores the continuous evolution of research methodologies and the need for ongoing exploration to enhance the robustness and effectiveness of data collection in computational data mining studies.

Keywords: Comparative Modeling, Support Vector Machine, Naive Bayes Classifier

1. Introduction

Data mining involves extracting information from extensive and intricate datasets. However, the effectiveness of this process hinges on the careful selection of data. Ensuring the relevance and quality of data is crucial to producing meaningful and valid insights. Inadequate or low-quality data can compromise the accuracy and utility of analysis results [1], [2]. Hence, meticulous data selection is imperative to instill trust in the analysis outcomes, enabling informed decision-making.

Sufficient data is essential for robust and representative analysis results. An insufficient amount may yield non-representative outcomes, while an excess can lead to unwarranted complexity and time consumption in the data mining process. Additionally, proper data selection is pivotal in avoiding significant errors or defects, which could render the analysis results inaccurate or futile [3]. By upholding validity and reliability through careful data selection, the data used in the data mining process remains trustworthy.

Aligning the selected data with the analysis objectives is crucial, considering the varying purposes in different applications. The questionnaire method, a means of collecting data for sentiment analysis [4], [5], employs written inquiries distributed to respondents. This method is adept at capturing subjective data, such as perceptions, opinions, or feelings about a specific topic, and can be applied in both online and offline formats.

Questionnaire-derived data offers insights into respondents' perceptions, opinions, or feelings regarding a product, service, company, event, or occurrence. Concurrently, secondary data sourced from pre-existing outlets like journals, books, or articles, proves valuable in sentiment analysis to discern public opinions or sentiments [6], [7]. This secondary

***Corresponding author: Hendra Jatnika (h.jatnika@itpln.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i1.148

 ISSN 2723-6471

data can unveil sentiments about a product, service, company, event, or occurrence. Integrating questionnaire data with secondary data creates a more comprehensive analysis, where the former provides direct insights from respondents, and the latter offers a broader perspective on general sentiments [8], [9]. This amalgamation produces a more holistic understanding of people's opinions or feelings on a given topic.

This study aims to assess the computational effectiveness of sentiment analysis through the examination of data obtained via both primary and secondary data collection approaches. The acquired data will undergo testing using standard sentiment analysis algorithms, namely the Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), ensuring consistent and reliable computational outcomes. The objective is to identify a data collection method that yields the most accurate and dependable results in sentiment analysis.

What sets this research apart is its focus on the data collection process. While traditional data mining studies emphasize achieving optimal accuracy in computation, this research distinguishes itself by concentrating on identifying the most effective data collection method. It involves a comparative analysis of the factors influencing the advantages and disadvantages of each employed data collection method. By doing so, the study aims to provide a more profound understanding of the data collection process and its impact on the ultimate results of data analysis.

2. Method

The classification technique is a fundamental aspect of data mining wherein a classifier is developed to predict categorical labels, such as "safe" or "risky" for financial lending applications, "yes" or "no" for marketing data, or "treatment A," "treatment B," "treatment C" for medical data. These categories can be represented with values based on specific requirements. Classification, akin to data mining, primarily involves the prediction of class labels [10], [11].

Each classification technique employs a learning algorithm to create a model that effectively captures the relationship between a set of attributes and class labels within the input data. Typically, the input for a classification model consists of a set of records (training set), each comprising a set of attributes, including one designated as the class [12], [13], [14]. The model for the class attribute functions based on the values of the other attributes. To assess the model's accuracy, a test set is utilized. Typically, the dataset is divided into training and test sets, with the former employed for model construction and the latter for validation [15].

The Naive Bayes Classifier (NBC) is a classification technique grounded in Bayes' theorem. NBC assumes the independence of each feature used for classification, hence the term "naive" to denote this feature dependency assumption. NBC facilitates binary and multi-class classification by computing the probability of a potential class for each data point [16], [17], [18].

On the other hand, the Support Vector Machine (SVM) is a classification technique designed to identify a line or hyperplane that separates two classes of data. SVM seeks the line with the maximum distance from each data point of distinct classes, referred to as the support vector. SVM can be applied to both binary and multi-class classification by incorporating kernels [19], [20], [21].

The key distinction between NBC and SVM lies in their treatment of features in the data. NBC presupposes feature independence, while SVM searches for a line or hyperplane to separate data classes. NBC is more suitable for data with independent features, whereas SVM is apt for data featuring intricate and non-independent attributes.

NBC exhibits faster processing compared to SVM due to its quicker calculation of class probabilities. Nevertheless, SVM proves more robust in handling non-linear data and boasts a higher accuracy rate than NBC. NBC is known for its simplicity in comprehension and implementation relative to SVM.

Both techniques possess their own merits and drawbacks, necessitating the selection of the appropriate technique based on the nature of the data and the classification objectives. NBC is well-suited for data with independent features, offering a reasonably high accuracy rate, while SVM is more fitting for data with complex and non-independent features, delivering a superior accuracy rate.

2.1. Data Evaluation

In this study, two distinct data collection methods were employed, namely the utilization of questionnaires and web mining facilitated by RapidMiner. Questionnaires involve the distribution of specific inquiries to respondents, and in this context, questions pertaining to campus services were administered to students on the campus. Respondents were encouraged to furnish responses aligned with their experiences regarding the quality of campus services.

Parallelly, the research incorporated web mining methods through RapidMiner, a tool designed for extracting data from the internet. RapidMiner was employed to scour the web for student comments on campus services, thereby augmenting the dataset for this study. The cumulative dataset amassed for analysis comprised 10 thousand entries in the form of student comments.

Both questionnaire and web mining data underwent identical preprocessing stages before being subjected to analysis. This preprocessing phase involved scrutinizing the accuracy and completeness of the data, while simultaneously adjusting the data format to align with the requirements of the subsequent analytical procedures.

Following the preprocessing stage, the data from both questionnaire responses and web mining were employed in the analysis through machine learning algorithms. These algorithms were utilized to assess the accuracy of the data collected through both methods. The outcomes of this analysis aimed to ascertain which data format yielded the most accurate results.

In essence, the overarching objective of this research is to determine the data format that can deliver optimal accuracy in evaluating campus services. By employing questionnaires and web mining with RapidMiner, the study sought to amass a substantial dataset for thorough analysis.

2.2. Data Preparation (Preprocessing)

This study involves several essential stages of data preprocessing aimed at ensuring equivalence between the two types of data employed. Initially, the removal of duplicates stage is implemented to eliminate identical data that may be present in both datasets. This step is crucial to mitigate the risk of including irrelevant or invalid data that could impact the final study outcomes. Subsequently, the Nominal to Text stage is applied to convert nominal data into textual data. Nominal data often poses challenges for effective analysis due to its lack of clear context. Through this conversion, the data becomes more interpretable, facilitating easier analysis.

Following that, the Transform Case stage is utilized to standardize the letter case of textual data to either uppercase or lowercase. This step is imperative to prevent inaccuracies in data comparison arising from variations in case sensitivity. Additionally, the Token Filter (by Length) stage is employed to exclude tokens with lengths below a specified limit. This measure is essential to avert issues related to inaccurate data comparison caused by the presence of excessively short tokens lacking clear meaning.

Moreover, the Stopword Filter stage is incorporated to eliminate words devoid of significant meaning for the purpose of this research. This step is vital to prevent inaccurate data comparison, as such words lack relevance in the analysis. Through the execution of these data preprocessing stages, it is anticipated that the data utilized in this investigation will yield results that are equivalent and valid.

2.3. Comparative Modeling

The process of computational modeling involves analyzing data and forecasting outcomes. Within the RapidMiner application, this modeling can be executed through the utilization of the Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) algorithms. The NBC employs probability methods for data classification, while SVM utilizes machine learning methods. In the RapidMiner application, the computational modeling process with NBC and SVM entails importing data, applying the chosen algorithm, and subsequently testing the data. Comparisons between the results of computational modeling with NBC and SVM enable the identification of the superior algorithm for application usage. This computational modeling approach with NBC and SVM finds applications in diverse areas such as face recognition, sentiment analysis, and object recognition in images. The use of the RapidMiner application facilitates a more streamlined and efficient execution of computational modeling processes.

2.4. Evaluation

The assessment conducted in this study aims to ascertain the utility of the previously developed model. The evaluation employs the 10-fold cross-validation technique, which involves partitioning the data into 10 segments, each serving as the test dataset in turn.

Two algorithms are utilized in this evaluation to derive Accuracy (evaluated through Confusion Matrix) and AUC (Area Under Curve) values. Accuracy gauges the model's ability to correctly identify data, distinguishing between correct and incorrect classifications. Meanwhile, AUC assesses the model's proficiency in discriminating between positive and negative data. The results of this assessment yield a Receiver Operating Characteristic (ROC) graph that illustrates the AUC value. This graphical representation elucidates the relationship between the True Positive Rate and False Positive Rate of the model. A higher AUC value, approaching 1, signifies that the employed model effectively distinguishes between positive and negative data.

3. Result and Discussion

3.1. Accuracy Value of SVM Algorithm

Based on the outcomes obtained from assessing the aforementioned model, it is evident that the SVM algorithm demonstrates a commendable accuracy level. The accuracy values are presented in Tables 1 and 2, depicting the outcomes of data classification through the utilization of the SVM algorithm. These accuracy values signify the model's proficiency in categorizing the data subjected to the testing process.

Table 1. SVM algorithm accuracy value on questionnaire data

Accuracy: 82.39% +/- 1.44% (micro average: 82.39%)			
	true Positive	true Negative	class precision
pred. Positive	511	109	80.77%
pred. Negative	72	466	86.67%
class recall	87.63%	81.45%	

$$Acc\ (Accuracy)\ = \frac{TP+TN}{TP+TN+FP+FN} = \frac{511+466}{511+109+72+466} = \frac{977}{1158} = 0.823 \quad (1)$$

Table 2. SVM algorithm accuracy value on web mining data

Accuracy: 81.22% +/- 1.76% (micro average: 81.22%)			
	true Positive	true Negative	class precision
pred. Positive	501	98	80.12%
pred. Negative	89	531	81.10%
class recall	81.21%	75.54%	

$$Acc\ (Accuracy)\ = \frac{TP+TN}{TP+TN+FP+FN} = \frac{501+98}{501+89+98+531} = \frac{599}{1219} = 0.812 \quad (2)$$

Moreover, the questionnaire data demonstrates an accuracy outcome of 84.45%, while the accuracy of web mining data is slightly lower at 81.22%. This discrepancy suggests that the questionnaire outperforms web mining in terms of accuracy. The disparity can be attributed to the presence of numerous symbols in web mining data, introducing complexity that hampers accuracy results. These symbols contribute to classification errors, thereby diminishing the overall accuracy of the model. In contrast, questionnaires do not incorporate symbols, leading to a higher accuracy value compared to web mining data.

 ISSN 2723-6471

3.2. Manuscript Format

Table 3. Accuracy value of NBC algorithm on questionnaire data

Accuracy: 81.42% +/- 4.22% (micro average: 81.42%)			
	true Positive	true Negative	class precision
pred. Positive	499	111	79.12%
pred. Negative	72	437	82.34%
class recall	82.34%	78.21%	

$$Acc\ (Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} = \frac{499+437}{499+111+71+437} = \frac{936}{1118} = 0.814 \quad (3)$$

Table 4. Accuracy value of NBC algorithm on web mining data

Accuracy: 78.21% +/- 3.46% (micro average: 78.21%)			
	true Positive	true Negative	class precision
pred. Positive	482	157	76.33%
pred. Negative	77	563	78.20%
class recall	80.02%	73.77%	

$$Acc\ (Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} = \frac{482+563}{482+157+77+563} = \frac{1045}{1279} = 0.782 \quad (4)$$

First, the results of model testing using the NBC algorithm can produce accuracy values that can be seen in Tables 3 and 4. Table 3 shows the accuracy results of the questionnaire data obtained by 81.42%. While Table 4 shows the accuracy results of the web mining data obtained by 78.21%.

Second, the accuracy of the questionnaire data is higher than the web mining data. This can be explained because when filling out the questionnaire, it is not allowed to fill in with symbols. This makes the questionnaire data cleaner and does not contain symbols that can confuse the accuracy results. Whereas in web mining data, many symbols are found so that they can confuse accuracy results. Third, the accuracy results obtained from questionnaire data and web mining data can still be improved by doing a better data cleaning process. In addition, other algorithms such as SVM or Random Forest can also be used to improve the accuracy of the model used.

3.3. AUC Value of SVM Algorithm

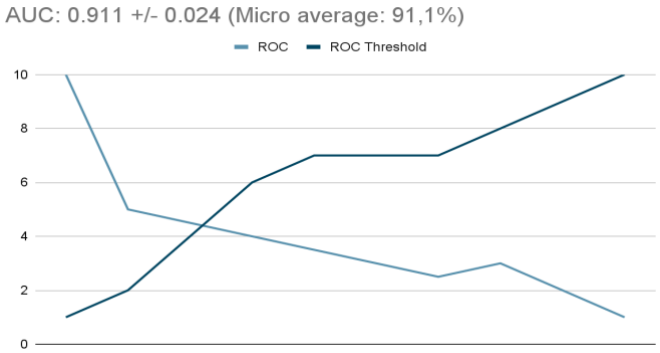

Figure 1. AUC value in SVM algorithm

The outcomes of the conducted model testing reveal that the employed model demonstrates commendable accuracy. Evidently, the Area Under Curve (AUC) value stands at 0.911, signifying the model's ability to effectively discern

 ISSN 2723-6471

between positive and negative data. The proximity of the AUC value to 1 indicates a robust capability in data classification. Moreover, the obtained AUC value attests to the model's high-performance level in accurately categorizing data. The overall accuracy achieved by the model is outstanding, underscoring its proficiency in precisely predicting data classes. Consequently, the model proves adept at making informed decisions and identifying patterns within the data.

In summary, the results of the model testing showcase the model's strong performance in data classification. With an AUC value of 0.911 and excellent accuracy, the model excels in distinguishing between positive and negative data, affirming its accuracy in predicting data classes. This suggests the practical applicability of the model in relevant fields.

3.4. AUC Value of NBC Algorithm

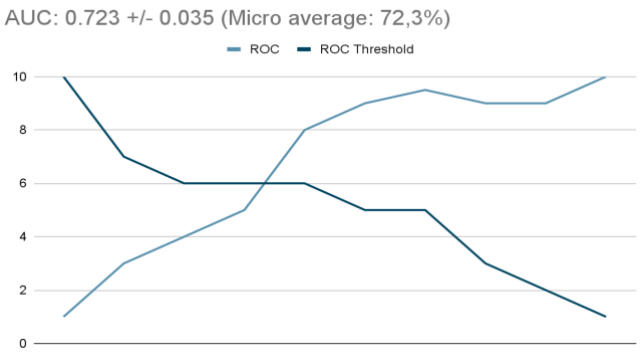

Figure 2. AUC value in NBC algorithm

Initially, the testing of the model aimed to appraise its accuracy in correctly categorizing data, gauging its proficiency. The test results indicated a lackluster accuracy level, pointing to the model's difficulty in appropriately classifying the data. Subsequent evaluation incorporated the utilization of the Area Under Curve (AUC) metric to assess the model's efficacy in binary classification. The obtained AUC value of 0.723 suggested a moderate ability of the model to distinguish between positive and negative classes. Furthermore, the analysis of the ROC graph results revealed the model's inadequacy in accurate data classification, as evidenced by the point (0,0) residing below the reference line, indicating suboptimal accuracy. Consequently, it is imperative to enhance the model to improve both accuracy and AUC performance.

3.5. Performance Comparison

Derived from the outcomes of the aforementioned algorithms analysis, a concise summary is presented in Table 5.

Table 5. Comparison of dataset performance results

	Questionnaire		**Web Mining**	
	SVM	NBC	SVM	NBC
Accuracy	82.39%	81.22%	81.42%	78.21%
AUC	0.911		0.723	

Initially, the utilization of questionnaires provides the researcher with the ability to regulate and structure the data completion process. The researcher retains control over the formulation of questions and can establish specific criteria for collecting the required data, a crucial aspect in research demanding precision and validity.

Furthermore, questionnaires enable researchers to selectively gather data from particular sources, tailoring their approach to respondents meeting predefined criteria such as age, gender, or educational background. This proves especially significant in research necessitating data that accurately represents the targeted population.

 ISSN 2723-6471

In addition, questionnaires afford researchers the flexibility to collect data through diverse channels, including online platforms, phone interactions, or interviews. This flexibility becomes particularly valuable in research scenarios where data accessibility from multiple locations is essential.

Moreover, the structured format of data derived from questionnaires facilitates ease of analysis. The data can be converted into more manageable formats, such as tables or graphs, streamlining the analysis process. This capability is especially crucial in research requiring prompt and efficient data analysis.

Lastly, the data gathered through questionnaires empowers researchers to take informed actions. Based on the obtained data, researchers can implement appropriate measures, such as modifying work patterns or initiating preventive actions. This proves particularly vital in research aiming to utilize data for enhancing work quality or facilitating well-informed decision-making.

4. Conclusion

This research assesses the efficacy of two distinct data collection approaches, specifically employing questionnaires and web mining in the domain of data mining computing. Preceding the model evaluation, the data utilized in this investigation undergoes a preprocessing phase. Subsequent to subjecting the model to NBC and SVM algorithms, the assessment and validation outcomes demonstrate that the questionnaire-based data collection method exhibits notable advantages, showcasing high flexibility and superior accuracy in comparison to the web mining data collection technique. Moreover, the AUC value derived from the questionnaire method is notably high, registering at 0.9, signifying its considerable flexibility. These findings substantiate the superiority of the questionnaire data collection method, particularly within the realm of data mining computing.

5. Declarations

5.1. Author Contributions

Conceptualization: H.J. and A.W.; Methodology: A.W.; Software: H.J.; Validation: H.J. and A.W.; Formal Analysis: H.J. and A.W.; Investigation: A.A.; Resources: A.A.; Data Curation: A.A.; Writing Original Draft Preparation: A.A. and H.J.; Writing Review and Editing: A.A. and H.J.; Visualization: H.J.; All authors, H.J., A.W., A.A., have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. K. J. Groenhof *et al.*, “Data mining information from electronic health records produced high yield and accuracy for current smoking status,” *J. Clin. Epidemiol.*, vol. 118, no. 1, pp. 100–106, 2020.

[2] R. Kusumawati, A. D’arofah, and P. A. Pramana, “Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter’s Classification of Tokopedia Services,” in *Journal of Physics: Conference Series*, 2019, vol.

1320, no. 1, pp. 1-12, 2016.

[3] Y. A. Singgalen, "Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1634–1646, 2022.

[4] F. Tempola, "Implemented PSO-NBC and PSO-SVM to Help Determine Status of Volcanoes," *J. Penelit. Pos dan Inform.*, vol. 9, no. 2, pp. 97–103, 2019.

[5] M. L. Laia and Y. Setyawan, "Perbandingan hasil klasifikasi curah hujan menggunakan metode SVM dan NBC," *J. Stat. Ind. dan Komputasi*, vol. 5, no. 02, pp. 51–61, 2020.

[6] G. Galih and M. Eriyadi, "Perbandingan Model NBC, SVM, dan C4. 5 dalam Mengukur Kinerja Karyawan Berprestasi Pasca Pandemi Covid-19," *J. Inform.*, vol. 9, no. 2, pp. 123–130, 2022.

[7] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, no. 1, pp. 114060-114072, 2021.

[8] J. C. Correa *et al.*, "Evaluation of collaborative consumption of food delivery services through web mining techniques," *J. Retail. Consum. Serv.*, vol. 46, pp. 45–50, 2019.

[9] S. Singh and M. S. Aswal, "Ontology learning procedures based on web mining techniques," 2019.

[10] V. Sathiyamoorthi, "An intelligent system for predicting a user access to a web based E-learning system using web mining," *Int. J. Inf. Technol. Web Eng.*, vol. 15, no. 1, pp. 75–94, 2020.

[11] A. Shalini and M. R. Ambikapathy, "E-Commerce Analysis and Product Price Comparison Using Web Mining," *J. homepage www. ijrpr. com ISSN*, vol. 2582, pp. 7421-7434, 2022.

[12] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, and Y. El Allioui, "A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments," *Educ. Inf. Technol.*, vol. 24, no. 3, pp. 1943–1959, 2019.

[13] P. Sokkhey, S. Navy, L. Tong, and T. Okazaki, "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia," *IEIE Trans. Smart Process. Comput.*, vol. 9, no. 3, pp. 217–229, 2020.

[14] I. Yusuf, D. Purwana, and A. D. Buchdadi, "The Influence of Interpersonal Communication, Universal-Diverse Orientation (UDO), and Self-Efficacy on the Quality of Administrative Services at State University of Jakarta", *Int. J. Appl. Inf. Manag.*, vol. 2, no. 2, pp. 97–105, Dec. 2021.

[15] M. S. Abd and S. F. Behadili, "Recognizing job apathy patterns of Iraqi higher education employees using data mining techniques," *J. Southwest Jiaotong Univ.*, vol. 54, no. 4,pp. 1-8, 2019.

[16] Z. Jin, "Analysis on NSAW Reminder Based on Big Data Technology," *Int. J. Informatics Inf. Syst.*, vol. 5, no. 3, pp. 108–113, Sep. 2022

[17] M. Nie, Z. Xiong, R. Zhong, W. Deng, and G. Yang, "Career choice prediction based on campus big data—mining the potential behavior of college students," *Appl. Sci.*, vol. 10, no. 8, p. 2841, 2020.

[18] D. Maillard, "The Obsolescence of Man in The Digital Society", *Int. J. Appl. Inf. Manag.*, vol. 1, no. 3, pp. 99–124, Jul. 2021.

[19] G. Belostecinic *et al.*, "Teleworking—An Economic and Social Impact during COVID-19 Pandemic: A Data Mining Analysis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 1, pp. 1–36, 2022.

[20] R. Wang, "Research on Network Data Algorithm Based on Association Rules," *Int. J. Informatics Inf. Syst.*, vol. 6, no. 2, pp. 66–73, Mar. 2023

[21] Y. Cui, "Intelligent recommendation system based on mathematical modeling in personalized data mining," *Math. Probl. Eng.*, vol. 2021, no. 1, pp. 1-7, 2021.