IMURIS
Journal of Computational and Information Research

Vol. 01, No. 1, June, 2026
Pages 39-48

Classifying Family Economic Status Using the K-Nearest Neighbor Algorithm in Popalia Village

Istrikah[1], Rabiah Adawiyah[2], Yuwanda Purnamasari Pasrun[3*]

[1,2,3]Department of Information System, Universitas Sembilan Belas November
Jl. Pemuda, Kolaka, 93517, Sulawesi Tenggara, Indonesia

Email Correponding Author: *yuwandapurnamasari@gmail.com*

Article Info

Article history:

Received: February 18, 2026
Revised: April 15, 2026
Accepted: June 20, 2026

Keywords:

Classification,
Confusion matrix,
Family economic status,
K-Nearest Neighbor,
Village information system.

ABSTRACT

Access to accurate family economic data is essential for the equitable distribution of village social assistance. At the Popalia Village Office, Tanggetada Sub-district, Kolaka Regency, identification of eligible recipients previously relied on manual, page-by-page verification of Statistics Indonesia (BPS) census documents, a process that was slow and often produced recipients that did not match the intended criteria. This study develops a web-based classification system using the K-Nearest Neighbor (KNN) algorithm to categorize 160 household heads into "Mampu" (financially capable) and "Tidak Mampu" (financially incapable) classes based on twelve socio-economic criteria, including occupation, monthly income, education, number of dependents, and asset ownership. The system was built following the Waterfall development model using PHP and MySQL with a use-case-driven UML design. Model performance was evaluated using Euclidean-distance-based KNN with 10-fold cross validation and confusion matrix analysis. The system achieved an average classification accuracy of 99.38% (minimum 93.75%, maximum 100%), a precision of 98.21%, a recall of 100%, and an F1-score of 99.10%. Black-box testing further confirmed that all functional modules operated as intended. These findings indicate that KNN is an accurate and practical method for supporting village-level social assistance targeting.

: https://doi.org/10.30598/timurisv1i1pp39-48

1. INTRODUCTION

Popalia Village Office, located in Tanggetada Sub-district, Kolaka Regency, Southeast Sulawesi, is responsible for public administrative services for approximately 160 household heads (Kepala Keluarga, KK). As with most rural administrations in Indonesia, the village office is regularly tasked with distributing social assistance programs such as subsidized rice (raskin), direct cash transfers (BLT), and other forms of aid, which are typically disbursed every few months. Determining which households are genuinely eligible for such programs requires an accurate and consistent measure of family economic status — a stratification that places each household into a defined socio-economic position within the community [1].

In practice, the distribution process at Popalia Village Office still depends on manual inspection of census documents produced by the Central Statistics Agency (Badan Pusat Statistik, BPS). Officers examine each document individually to identify households that qualify for assistance. This procedure is time-consuming and prone to error, frequently resulting in recipients who do not fully satisfy the intended criteria. The criteria considered relevant for assistance eligibility in this study consist of twelve indicators: occupation, monthly income, last education, number of household members, number of school-aged children, building area, floor type, wall type, electric power capacity, cooking-fuel type, vehicle ownership, and asset ownership. Given the number of criteria and households involved, a systematic classification approach is required to accelerate and standardize the decision process.

Classification is a data mining technique that predicts the class label of new, unclassified data based on patterns learned from historical data [2]. Among the many classification techniques available, the K-Nearest Neighbor (KNN) algorithm is attractive for problems of this scale because it is simple to implement, fast to train, tolerant of noisy training data, and generally effective on comparatively small or medium-sized datasets [3], [4]. KNN classifies a new instance according to the majority class among its k nearest neighbors in the training data, where distance is most commonly measured using the Euclidean metric.

Several prior studies have examined classification methods for similar socio-economic and pattern-recognition problems, summarized in Table 1. Annur [5] applied Naive Bayes to classify impoverished communities and obtained 73% accuracy; Maharani [6] used a Backpropagation neural network with momentum and adaptive learning rate and reported 96% accuracy; Syahid et al. [7] used KNN with HSV color features for ornamental plant classification and achieved 92% accuracy; Permatasari [8] applied KNN to classify family economic status in Pacewetan Village, Nganjuk Regency, East Java, and obtained an average accuracy of 98.8% using 5-fold cross validation; and Raysyah et al. [9] combined KNN with Principal Component Analysis (PCA) for coffee-fruit ripeness classification, reaching 97.77% accuracy at k = 3.

Table 1. Summary of Related Studies

No.	Author (Year)	Method	Result
1.	Annur (2019) [5]	Naive Bayes	Classification of low-income communities; confusion matrix accuracy of 73%.
2.	Maharani (2020) [6]	Backpropagation + adaptive learning rate	Momentum and adaptive learning rate accelerated convergence; accuracy of 96%.
3.	Syahid et al. (2020) [7]	KNN + HSV features	Classified ornamental Philodendron leaves; accuracy of 92%.
4.	Permatasari (2019) [8]	KNN, 5-fold CV	Classified family economic status in Nganjuk Regency; average accuracy of 98.8%.
5.	Raysyah et al. (2021) [9]	KNN + PCA	Coffee-fruit ripeness classification by color; accuracy of 97.77% at k = 3.

Compared to the closest precedent Permatasari's [8] KNN-based classification of family economic status in Nganjuk the present study contributes in three ways: (1) it applies the method to a different regional context (Popalia Village, Southeast Sulawesi) with a locally relevant criteria set; (2) it validates the model using a full 10-fold cross validation together with confusion-matrix-derived precision, recall and F1-score, rather than accuracy alone; and (3) it embeds the classifier inside an operational, role-based web information system (administrator and village head accounts) so that classification results can be produced, reviewed, and reported without manual document handling. The remainder of this article is organized as follows: Section 2 describes the research method, Section 3 presents the results and discussion, and Section 4 concludes the study.

2. METHOD

2.1 Research Location, Time, and Data Collection

The research was conducted at the Popalia Village Office, Tanggetada Sub-district, Kolaka Regency, Southeast Sulawesi Province, Indonesia, from February to May 2023. Data were collected through three techniques: (1) direct observation of household records at the village office; (2) interviews with the village head and finance/planning staff regarding the socio-economic condition of residents; and (3) a literature study of books, journal articles, and online references relevant to data mining and the KNN algorithm. The primary data consisted of 160 household records obtained from the village office's BPS census documents, while secondary data included supporting literature used to build the theoretical foundation of the study. All data used in this research are quantitative in nature.

2.2 Research Workflow

The overall research workflow is illustrated in Figure 1. Data collected from the village office were pre-processed and organized into a training dataset with a class label ("Mampu" or "Tidak Mampu") verified against the twelve socio-economic criteria described in Section 1. The KNN model was then evaluated using 10-fold cross validation, and the resulting confusion matrix was used to compute accuracy, precision, recall and F1-score. In parallel, a web-based information system implementing the classifier was developed following the Waterfall model requirements analysis, design, coding, and testing which is well suited to projects whose specifications are unlikely to change substantially during development [10].

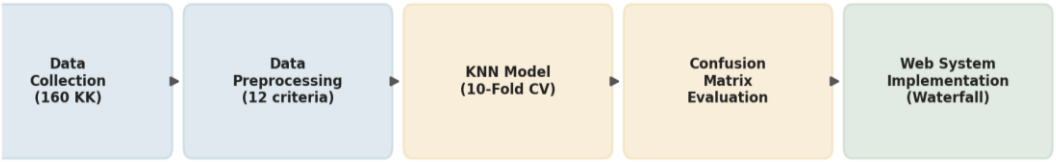

Figure 1. Research Workflow

2.3 Research Workflow

KNN classifies a query instance according to the majority class of its k nearest neighbors in the training set, where proximity is defined by a distance metric — most commonly the Euclidean distance [11], [12]. The Euclidean distance between a training instance X and a testing instance Y, each described by n attributes, is computed as shown in Equation (1).

$$d(X,Y) = \sqrt{\sum(X_i - Y_i)^2} \quad (1)$$

where $d(X,Y)$ is the distance between the two instances, n is the number of attributes (n = 12 in this study), and Xi, Yi are the i-th attribute values of the training and testing instances, respectively. The classification procedure follows five steps [13]: (1) determine the parameter k, the number of nearest neighbors to be considered; (2) compute the Euclidean distance between the query instance and every training instance using Equation (1); (3)

sort the training instances in ascending order of distance; (4) select the class labels of the k nearest instances; and (5) assign the query instance to the majority class among those k neighbors. To illustrate the distance computation, consider two hypothetical instances described by four normalized attribute values, X = [8, 2, 4, 3] and Y = [1, 6, 7, 5]. Applying Equation (1):

$$d(X,Y) = \sqrt{[(8-1)^2 + (2-6)^2 + (4-7)^2 + (3-5)^2]} = \sqrt{78} \approx 9.17$$

A smaller distance value indicates greater similarity between the two instances; consequently, instances with the smallest distances to the query point are the most influential in determining its predicted class. A smaller distance value indicates greater similarity between the two instances; consequently, instances with the smallest distances to the query point are the most influential in determining its predicted class.

2.4 Model Evaluation: K-Fold Cross Validation and Confusion Matrix

Model performance was evaluated using k-fold cross validation with k = 10, a configuration widely recommended in the data mining literature because it offers a favorable bias–variance trade-off compared to smaller fold counts or leave-one-out validation [14]. The 160-record dataset was partitioned into ten equally sized folds of sixteen records each; in each of the ten iterations, nine folds (144 records) were used as training data and the remaining fold (16 records) as testing data, and the resulting confusion matrix was recorded. The confusion matrix summarizes four possible outcomes of a binary classification task: True Positive (TP) for households correctly classified as “Mampu”; True Negative (TN) for households correctly classified as “Tidak Mampu”; False Positive (FP) for households actually “Tidak Mampu” but classified as “Mampu”; and False Negative (FN) for households actually “Mampu” but classified as “Tidak Mampu” [15]. From these four values, four evaluation metrics were computed, given in Equations (2) until (5).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

$$Precision = TP / (TP + FP) \quad (3)$$

$$Recall = TP / (TP + FN) \quad (4)$$

$$F1 - score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (5)$$

In addition to model evaluation, the resulting web application was functionally tested using black-box testing, a technique that verifies whether the software behaves as expected purely from its input–output specification, without inspecting internal code logic [16]. Test cases covered authentication, data-training management, testing (uji data), classification, and reporting modules.

2.5 Software and Hardware

The system was developed on a laptop with an Intel® Celeron® N4000 processor (1.10 GHz) and 4 GB RAM, running Windows 10. The software stack consisted of PHP as the server-side scripting language, MySQL as the database management system, XAMPP as the local web server, phpMyAdmin for database administration, and Sublime Text as the code editor, following common conventions for small-scale institutional web applications [17], [18], [19], [20], [21]. The system architecture was modeled using Unified Modeling Language (UML) diagrams — use case, activity, and sequence diagrams — to represent the

interaction between the Administrator and Village Head actors and the system's core functions (login, data training, testing, classification, and reporting) [22], [23], [24].

3. RESULTS AND DISCUSSION

3.1 Ten-Fold Cross Validation Results

Table 2 presents the confusion matrix components (TP, TN, FP, FN) and the resulting accuracy for each of the ten folds obtained after applying the KNN classifier to the 160-record dataset.

Table 2. Confusion Matrix and Accuracy per Fold (10-Fold Cross Validation)

Fold	TP	TN	FP	FN	Accuracy (%)
1	6	10	0	0	100.00
2	5	11	0	0	100.00
3	5	11	0	0	100.00
4	6	10	0	0	100.00
5	5	11	0	0	100.00
6	6	10	0	0	100.00
7	5	10	1	0	93.75
8	6	10	0	0	100.00
9	6	10	0	0	100.00
10	5	11	0	0	100.00
Total / Average	55	104	1	0	99.38

Nine of the ten folds achieved a perfect classification accuracy of 100%, while fold 7 produced one misclassified instance (a false positive), yielding an accuracy of 93.75%. Figure 2 visualizes the accuracy obtained in each fold together with the overall average.

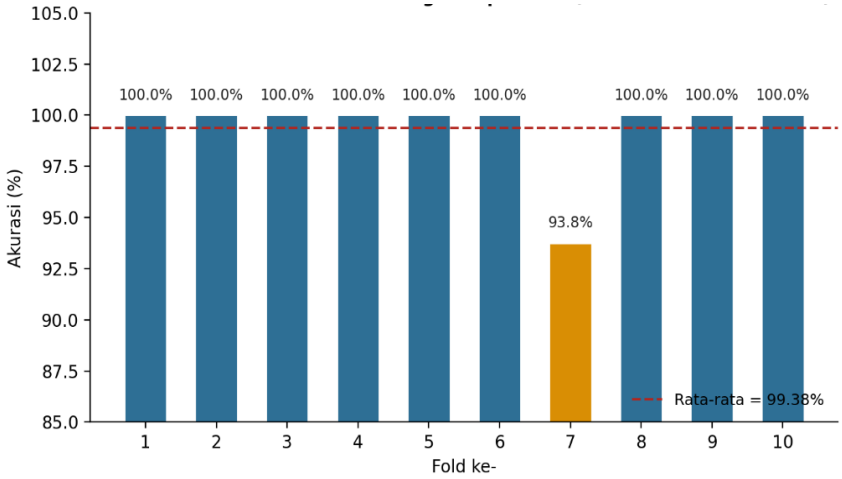

Figure 2. Classification Accuracy per Fold (10-Fold Cross Validation)

The average classification accuracy across the ten folds is 99.38%, with a minimum of 93.75% and a maximum of 100%. Aggregating the confusion matrix values across all folds gives TP = 55, TN = 104, FP = 1, and FN = 0, out of 160 total household records. Figure 3 presents this aggregated confusion matrix.

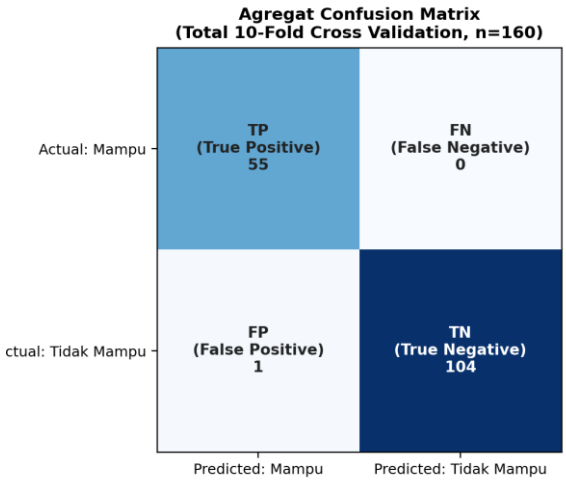

Figure 3. Aggregated Confusion Matrix (10-Fold Cross Validation, n = 160)

3.2 Accuracy, Precision, Recall, and F1-Score

Applying Equation (2) to the aggregated confusion matrix values yields:

$$Accuracy = (55 + 104)/(55 + 104 + 1 + 0) = 159/160 = 0.9938 = 99.38$$

This pooled value is consistent with the average of the ten individual fold accuracies reported in Table 2, confirming the stability of the classifier across folds. Beyond overall accuracy, Equations (3), (4) ,(5) were applied to further characterize the classifier's behavior with respect to the "Mampu" (positive) class:

$$Precision = 55/(55 + 1) = 0.9821 = 98.21$$

$$Recall = 55/(55 + 0) = 1.0000 = 100.00$$

$$F1 - score = (2 \times 0.9821 \times 1.0000)/(0.9821 + 1.0000) = 0.9910 = 99.10$$

A recall of 100% indicates that the classifier did not fail to identify any household that was genuinely "Mampu" (no false negatives), while a precision of 98.21% shows that only one household was incorrectly classified as "Mampu" when it was in fact "Tidak Mampu." In the context of social-assistance targeting, this error pattern is comparatively low-risk: a missed "truly capable" classification (false negative) would have unnecessarily excluded an eligible household from receiving no assistance, whereas the single observed error (a false positive) means one non-eligible household could be mistakenly granted "capable" status an outcome that can be corrected through the village head's manual review step already built into the system.

3.3 Distribution of Classification Results

Of the 160 households evaluated, 55 households (34.4%) were classified as "Mampu" and 105 households (65.6%) were classified as "Tidak Mampu," as shown in Figure 4. This distribution suggests that roughly two-thirds of Popalia Village households fall below the threshold used to define financial capability, underscoring the continuing relevance of village-level social assistance programs and the importance of an accurate, low-error classification tool for directing that assistance to the appropriate households.

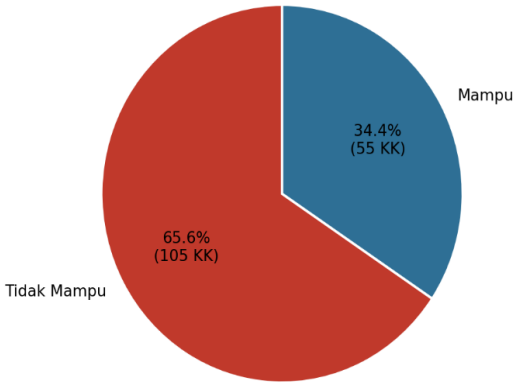

Figure 4. Distribution of Family Economic Status Classification Results

3.4 Functional (Black-Box) Testing

The web-based classification system was tested functionally using black-box test cases covering the authentication module, as summarized in Table 3; equivalent test scenarios were also executed for the data-training, testing, classification, and reporting modules, all of which produced outcomes consistent with expectations.

Table 3. Black-Box Testing Results for the User Authentication Module

No.	Test Case	Expected Result	Status
1.	Username and/or password left empty	System remains on the login form and displays a validation message	Valid
2.	Only username or only password entered	System displays a message requesting the missing field	Valid
3.	Incorrect username entered	System displays an “invalid username or password” message	Valid
4.	Incorrect password entered	System displays an “invalid username or password” message	Valid
5.	Both username and password incorrect	System displays an “invalid username or password” message	Valid

3.5 Comparison with Related Work

Table 4 compares the accuracy achieved in this study with that of the related studies introduced in Table 4.

Table 4. Accuracy Comparison with Related Studies

Study	Method	Accuracy
Annur [5]	Naive Bayes	73.00%
Maharani [6]	Backpropagation (adaptive LR)	96.00%
Syahid et al. [7]	KNN + HSV features	92.00%
Permatasari [8]	KNN, 5-fold CV	98.80%
Raysyah et al. [9]	KNN + PCA	97.77%
This study	KNN, 10-fold CV	99.38%

The KNN classifier developed in this study achieved the highest accuracy among the compared methods, and slightly exceeded the closest precedent Permatasari’s [8] KNN-based economic status classification in Nganjuk (98.8%) despite being evaluated with a more stringent 10-fold rather than 5-fold cross validation. This result is consistent with the

broader literature indicating that KNN performs particularly well on classification tasks with clearly separable, low-dimensional numerical criteria such as income brackets, dependents count, and housing indicators [3], [12]. Nonetheless, two limitations should be acknowledged. First, the dataset is comparatively small (n = 160) and moderately imbalanced (55 "Mampu" versus 105 "Tidak Mampu" households); the single misclassification observed in fold 7 illustrates that performance on minority-adjacent boundary cases can still vary between folds. Second, because KNN is a lazy-learning, complete-storage algorithm, prediction time scales with dataset size [13]; while this is not a practical concern at the current scale of 160 records, it may require optimization (e.g., k-d trees or data reduction) should the system later be scaled to cover multiple villages within the sub-district.

4. CONCLUSION

This study developed and evaluated a web-based classification system that applies the K-Nearest Neighbor algorithm to determine family economic status for 160 households in Popalia Village, Tanggetada Sub-district, Kolaka Regency. Using Euclidean-distance-based KNN validated through 10-fold cross validation, the system achieved an average classification accuracy of 99.38% (minimum 93.75%, maximum 100%), a precision of 98.21%, a recall of 100%, and an F1-score of 99.10%, with an aggregated confusion matrix of TP = 55, TN = 104, FP = 1, and FN = 0. Black-box testing confirmed that all functional modules — authentication, data training, testing, classification, and reporting — operated according to specification. These results confirm that KNN can substantially reduce the time and inconsistency associated with manual, document-by-document verification previously used at the village office, and can support more equitable targeting of social assistance programs. Future work may extend this study by comparing KNN against other classifiers (e.g., Support Vector Machine or ensemble methods), incorporating attribute weighting to address the class imbalance observed between "Mampu" and "Tidak Mampu" households, and expanding the dataset to cover additional villages within Tanggetada Sub-district to test the generalizability of the model.

Acknowledgments

The authors thank the Head of Popalia Village and the staff of the Finance and Planning Divisions of the Popalia Village Office for their assistance in providing the data required for this study, and the Faculty of Information Technology, Universitas Sembilanbelas November Kolaka, for its academic support throughout the research process.

REFERENCES

[1] R. Pratama and A. Sari, "Socio-economic stratification models for village-level social assistance targeting," J. Soc. Policy Stud., vol. 6, no. 2, pp. 88–97, 2020.
[2] S. Hendrian, "Algoritma klasifikasi data mining untuk memprediksi siswa dalam memperoleh bantuan dana pendidikan," Faktor Exacta, vol. 11, no. 3, pp. 266–274, 2018.
[3] A. Maulana, "Konsep dasar data mining," J. Konsep Data Mining, vol. 1, pp. 1–16, 2019.
[4] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan data mining untuk evaluasi kinerja akademik mahasiswa menggunakan algoritma Naive Bayes classifier," Eeccis, vol. 7, no. 1, pp. 59–64, 2019.
[5] H. Annur, "Klasifikasi masyarakat miskin menggunakan metode Naive Bayes," ILKOM J. Ilm., vol. 10, no. 2, pp. 160–165, 2019, doi: 10.33096/ilkom.v10i2.303.160-165.
[6] W. Maharani, "Klasifikasi data menggunakan JST Backpropagation momentum dengan adaptive learning rate," 2020.
[7] D. Syahid, J. Jumadi, and D. Nursantika, "Sistem klasifikasi jenis tanaman hias daun Philodendron menggunakan metode K-Nearest Neighbor (KNN) berdasarkan nilai Hue, Saturation, Value (HSV)," J. Online Inform., vol. 1, no. 1, p. 20, 2020, doi: 10.15575/join.v1i1.63.
[8] D. I. Permatasari, "Klasifikasi status ekonomi keluarga dengan menggunakan algoritma K-Nearest Neighbor di Desa Pacewetan Kecamatan Pace Kabupaten Nganjuk," Simki-Techsin, vol. 1, no. 1, pp. 1–7, 2019.

[9] S. R. Raysyah, V. Arinal, and D. I. Mulyana, "Klasifikasi tingkat kematangan buah kopi berdasarkan deteksi warna menggunakan metode KNN dan PCA," J. Sist. Inf. (JSiI), vol. 8, no. 2, pp. 88–95, 2021, doi: 10.30656/jsii.v8i2.3638.
[10] Swastika, "Model pengembangan perangkat lunak waterfall untuk sistem informasi berbasis web," Bab II Landasan Teori, 2019.
[11] A. Johar, D. Yanosma, and K. Anggriani, "Implementasi metode K-Nearest Neighbor (KNN) dan Simple Additive Weighting (SAW) dalam pengambilan keputusan seleksi penerimaan anggota Paskibraka," Pseudocode, vol. 8, no. 1, pp. 98–112, 2021.
[12] R. Julita, "Algoritma K-Nearest Neighbour dalam mengklasifikasi penilaian peserta didik," JSAI (J. Sci. Appl. Inform.), vol. 3, no. 3, pp. 149–155, 2020, doi: 10.36085/jsai.v3i3.1163.
[13] Julita, "Karakteristik algoritma K-Nearest Neighbor sebagai metode lazy learning," JSAI (J. Sci. Appl. Inform.), vol. 3, no. 3, pp. 149–155, 2020.
[14] N. Suyal and A. Kumar, "A review of k-fold cross-validation strategies for model evaluation," Int. J. Data Sci. Anal., vol. 5, no. 3, pp. 210–219, 2019.
[15] G. Hasugian and A. Shidiq, "Analisis akurasi model klasifikasi menggunakan confusion matrix," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2019.
[16] G. W. Setiawan, "Pengujian perangkat lunak menggunakan metode black box: studi kasus Exelsa," J. Inform., vol. 3, pp. 286–293, 2021.
[17] L. Erawan, "Dasar-dasar PHP," Udinus Tech. Rep., pp. 1–47, 2018.
[18] C. Shah, "MySQL," in A Hands-On Introduction to Data Science, Cambridge, U.K.: Cambridge Univ. Press, 2020, pp. 187–206, doi: 10.1017/9781108560412.008.
[19] Ismai, "Analisa dan perancangan sistem informasi administrasi kursus bahasa Inggris pada Intensive English Course di Ciledug Tangerang," J. Ipsikom, vol. 8, no. 1, 2020.
[20] A. Suprianto, "Sublime text sebagai aplikasi editor kode dan teks lintas platform," J. Teknol. Inf., vol. 7, no. 1, pp. 48–58, 2018.
[21] R. Fauzi, "Pendahuluan JavaScript," J. Pemrograman Web, 2019.
[22] T. M. Prihandoyo, "Unified Modeling Language (UML) model untuk pengembangan sistem informasi akademik berbasis web," J. Inform., vol. 3, no. 1, pp. 126–129, 2018.
[23] T. A. Kurniawan, "Pemodelan use case (UML): evaluasi terhadap beberapa kesalahan dalam praktik use case (UML) modeling," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 1, pp. 77–86, 2018, doi: 10.25126/jtiik.201851610.
[24] Nurul Azwanti, "Sistem informasi penjualan tas berbasis web dengan pemodelan UML," J. Sist. Inf., vol. 4, no. 1, pp. 1–14, 2017.