Jurnal E-Komtek
Vol.8 No. 2(2024) pp. 399-409
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Advanced Credit Scoring with Naive Bayes Algorithm: Improving Accuracy and Reliability in Financial Risk Assessment

Adam Afandi [1], Herman Bedi Agtriadi [2], Luqman [3], Meilia Nur Indah Susanti [4]

[1,2,3,4] Computer Science, Institut Teknologi PLN, Indonesia, 11750

adam2330005@itpln.ac.id

https://doi.org/10.37339/e-komtek.v8i2.2160

Published by Politeknik Piksi Ganesha Indonesia

Artikel Info
Submitted:
04-12-2024
Revised:
23-12-2024
Accepted:
23-12-2024
Online first :
27-12-2024

Abstract

This study develops a credit application recommendation system based on the Naive Bayes method to improve accuracy and reliability in financial risk assessment. Using the CRISP-DM framework, the research process starts with understanding the business needs to implement a web-based system. The Naive Bayes algorithm was chosen because of its ability to handle binary data classification and generate reliable predictions even with limited training data. This study combines feature selection and unbalanced data handling techniques to improve model performance. The evaluation results showed that the system achieved an accuracy of 70.50%, with a precision of 92.16%, a recall of 64.57%, and an F1-score of 75.83%. This system is implemented as a web-based application to help financial institutions make credit decisions quickly and accurately. These findings significantly contribute to developing a data-based classification system for the banking sector, especially in reducing the risk of bad loans and improving decision-making efficiency.

Keywords: *Naive Bayes, Credit Assessment, CRIPS-DM, Recommendation System, Data Mining*

Abstrak

Penelitian ini mengembangkan sistem rekomendasi aplikasi kredit berbasis metode Naive Bayes untuk meningkatkan akurasi dan reliabilitas dalam penilaian risiko keuangan. Dengan menggunakan kerangka kerja CRISP-DM, proses penelitian dimulai dengan memahami kebutuhan bisnis untuk mengimplementasikan sistem berbasis web. Algoritma Naive Bayes dipilih karena kemampuannya menangani klasifikasi data biner dan menghasilkan prediksi yang andal bahkan dengan data pelatihan yang terbatas. Penelitian ini menggabungkan pemilihan fitur dan teknik penanganan data yang tidak seimbang untuk meningkatkan kinerja model. Hasil evaluasi menunjukkan bahwa sistem mencapai akurasi 70,50%, dengan presisi 92,16%, recall 64,57%, dan skor F1 75,83%. Sistem ini diimplementasikan sebagai aplikasi berbasis web untuk membantu lembaga keuangan membuat keputusan kredit dengan cepat dan akurat. Temuan ini berkontribusi secara signifikan untuk mengembangkan sistem klasifikasi berbasis data untuk sektor perbankan, terutama dalam mengurangi risiko kredit macet dan meningkatkan efisiensi pengambilan keputusan.

Kata-kata kunci: *Naive Bayes, Penilaian Kredit, CRIPS-DM, Sistem Rekomendasi, Data Mining*

1. Introduction

The banking sector, including Indonesia, faces increasingly fierce competition in the rapidly developing digital era. One of the main products of banks and financial institutions is credit, and the process of obtaining credit relies heavily on making the right decisions to reduce the risk of bad loans [1]. Therefore, an efficient system to assess customer creditworthiness is needed [2].

Data mining is an advanced analytical technique to extract patterns and important information from large, complex data sets. Data mining is also known as knowledge-discover in database (KDD). The main goal is to acquire knowledge that can be used to help them make better decisions, including credit risk prediction. In data mining, Naive Bayesian algorithms, which are based on statistical theory and probability, are used to classify data into specific categories by calculating the probabilities of each possible class. Naive Bayes' ability to handle binary and multinomial classifications and the ability to produce good results despite having relatively little training data are two of its main advantages [3].

A credit recommendation system using Naive Bayes is hoped to speed up decision-making and improve the accuracy of credit assessments. This system will provide more reliable guidance in determining whether a client is eligible for credit. Previous research has explored various applications of Naive Bayes' algorithm in the context of credit recommendation systems, such as Aida Krichene [4] The study applied Naive Bayes' algorithm to predict the risk of default on a short-term loan in a Tunisian commercial bank. The results show a good classification rate of 63.85%, highlighting the potential of Naive Bayes in credit risk assessment, Okesola et al. [5] The study developed a credit scoring model using Naive Bayes' algorithm, which showed improved accuracy in credit risk evaluation compared to traditional methods, Shaona Hua et al [6] The study developed a credit risk assessment model using federated three-way Naive Bayes, which considers data uncertainty and privacy protection. The results show an increase in accuracy in credit risk evaluation compared to traditional methods,

The purpose of this study is to create a Naive Bayes-based recommendation system that can accurately classify credit applications and assess its performance to ensure its reliability and effectiveness in the banking sector. Hopefully, this system will reduce the risk of bad credit and speed up and improve credit decisions.

2. Method

This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach, which is a systematic framework for the development of data-based models [7][8]. This process consists of six main stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Here is a picture of the research method used:

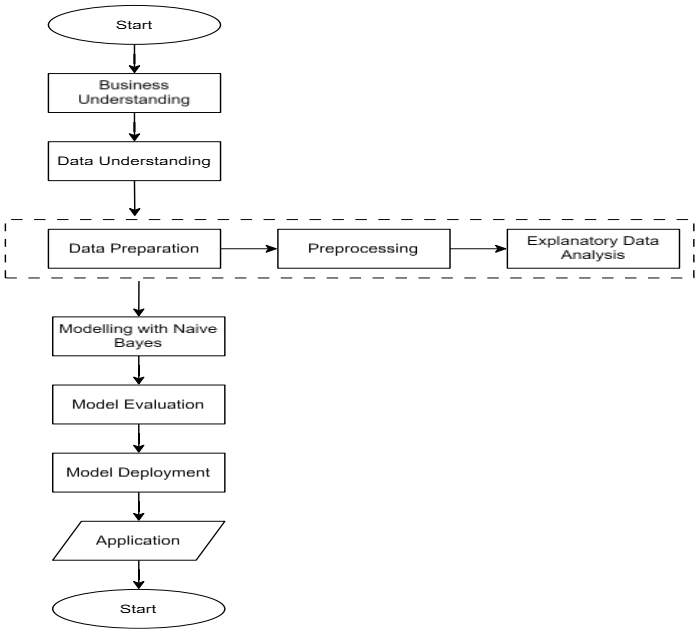

Figure 1. CRIPS-DM Research Method

Figure 1 above illustrates the process of the CRISP-DM (Cross Industry Standard Process for Data Mining) research method, this stage includes:

1. Business Understanding, understanding business objectives, in this case, financial institutions need solutions that can provide reliable decisions about customer credit applications because the credit decision-making process is very important in the banking sector and often faces major problems. This is especially true regarding reducing the risk of bad loans and ensuring proper customer assessments.
2. Data Understanding, Collecting, and Understanding Relevant Data At this stage, the data is collected and understood thoroughly to assess its structure, quality, and characteristics and support the research objectives.
3. Data Preparation involves processing, analyzing, and interpreting data with the Naive Bayes method to develop a predictive model for the credit application recommendation system.
4. Modelling, and building a credit risk assessment model using the Naive Bayes algorithm.

5. Evaluate the model's performance based on accuracy and F1-score metrics.
6. Deployment: Applying the model to a real system to be used in credit scoring in an application.

2.1 Naive Bayes Algorithm

Modeling is an important stage in the use of machine learning algorithms to generate predictive models based on data that has been processed [9]. The Naïve Bayes method is a classification algorithm based on Bayesian probability principles assuming that features are independent [10]. Despite its simplicity, Naïve Bayes has proven very effective in many classification applications, especially when features are considered independent. This algorithm is particularly beneficial for applications such as credit recommendation systems, where fast and accurate classification is essential. The study used Python to model Naïve Bayes through computational simulations on five limited training data. The calculation process of the Gaussian Naive Bayes Classifier is as follows:

Stage 1

Separating training data, test data, and targets: The first step in the Gaussian Naïve Bayes Classifier (GNBC) calculation is to separate the training data (Xtrain) from the target data (Ytrain). Table 1 shows this study's training data.

Table 1. Training Data (X)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16
1.65	9.08	0.04	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	0.054	16.08	1.13	0.007	37.75
2.36	5.6	0.01	0.01	1.35	0.004	1.28	0.62	1.88	0.33	0.13	0.007	0.021	18.6	1.78	0.007
1.01	14.02	0.04	0.58	0.008	4.2	0.53	0.02	0.99	0.05	0.003	0.078	14.16	1.11	0.006	50.28
1.36	11.33	0.04	2.96	0.001	7.2	0.03	1.66	1.08	0.71	0.71	0.016	1.41	1.29	0.004	9.12
0.92	24.33	0.03	0.2	0.2	0.006	2.67	0.69	0.57	0.61	0.13	0.001	0.117	6.74	0.003	16.9

The training data has 16 types of features, which are then assumed to be x1, x2, ..., xn. The target (Y) in this study can be seen in **Table 2**.

Table 2. Data target (Y)

y
1
0
0
1
1

The testing data in this study uses training data (Xtest) = x 1, x x n in the second row to simplify the calculation simulation process.

Stage 2

Calculating the Estimated Prior Probability of each class, here is the equation for calculating the probability of the prior P(Ytrain = yi) for each class.

$$P(Y_{train} = y_i) = \frac{(jumlah\ kelas\ dengan\ sampel\ y_i)}{jumlah\ total\ sampel} \quad (1)$$

The value calculation is as follows:P_0

$$P_0 = (Y_{train} = 0) = \frac{2}{5} = 0{,}4$$

Next, the value calculation P_1 with the results below:

$$P_1 = (Y_{train} = 1) = \frac{3}{5} = 0{,}6$$

The calculation above states that 3 instances of Ytrain = 1 out of a total of 5 instances exist. This shows that 60% of the training data has a target variable Y = 1.

Stage 3

Calculate the mean value and standard deviation of each class. The process of calculating the mean and standard deviation in the class $P(0)$ can be seen below:

$\mu_1 = \big(\mu(x_1), \mu(x_2), \dots\dots, \mu(x_n)\big)$, for the mean class 0 and class 1 $(Y = 0)(Y = 1)$

$$\mu_0 = \frac{(2.36 + 1.01)}{2}, \frac{(5.6 + 14.02)}{2}, \dots\dots, \frac{(0.08 + 0.01)}{2} = (1.68, 9.81, \dots\dots, 0.045$$

$$\mu_1 = \frac{(1.65 + 1.36 + 0.922)}{3}, \frac{(9.08 + 11.33 + 24.33)}{3}, \dots, \frac{(0.02 + 0.05 + 0.02)}{3} = (1.31, 14.91, \dots, 0.03)$$

After obtaining the values of μ0 and μ1, the next step is determining the standard deviation values on labels 0 and 1.

Furthermore, the calculation of the standard deviation uses equation (2) on the condition that the training data is {Xi|yi = 0} in **Table 1**, so that the following results are obtained:

$$\sigma_0 = \left(\sigma(x_1), \sigma(x_2), \dots, \sigma(x_n)\right) \quad (2)$$

Calculate the standard deviation for class 0 $Y(0)$

$$\sigma_0 = \sqrt{\frac{(2.36-1.68)^2+(1.01-1.68)^2}{2}}, \sqrt{\frac{(5.6-9.81)^2+(14.02-9.81)^2}{2}}, \dots, \sqrt{\frac{(0.08-0.045)^2+(0.01-0.045)^2}{2}}$$

$$\sigma_0 = (0.67, 4.21, \dots\dots, 0.03)$$

Calculating standard deviations for grade 1 $Y(1)$

$$\sigma_1 = \sqrt{\frac{(1.65-1.31)^2+(1.36-1.31)^2+(0.922-1.31)^2}{3}}, \sqrt{\frac{(9.08-14.91)^2+(11.33-14.91)^2+(24.33-14.91)^2}{3}},$$

$$\dots, \sqrt{\frac{(0.02-0.03)^2+(0.05-0.03)^2+(0.02-0.03)^2}{3}} = (0.089, 45.18, \dots\dots, 0.0002)$$

Stage 4

Probability Estimation *Using* Gaussian *Estimation*

Calculate probability P(xi = x|Y = yi) for each feature xi and class yi. *Gaussian Naive Bayes* considers the distribution of features xi in class yi is a Gaussian (normal) distribution. Therefore, it is necessary to calculate the *mean* μ and standard deviation of σ of each feature in each class. The Gaussian function formula is generally defined in equation (3).

$$f(\text{Xtest}, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X_{test}-\mu)^2}{2\sigma^2}\right) \quad (3)$$

Meanwhile, to calculate the likelihood function in the Gaussian estimation using equation (4).

$$P(\text{Xtest}|\text{C}) = \prod_{j=1}^{D} f\ (\text{Xtestj}, \mu j, \sigma j) \quad (4)$$

For Y=0

$$\text{P}(\text{Xtest}|0) = (0.35)(0.05)\cdots(6.91)$$

$$\text{P}(\text{Xtest}|0) = 3034927.11.$$

For Y=1

$$\text{P}(\text{Xtest}|0) = (0.002)(0.02)\cdots(0.05)$$

$$\text{P}(\text{Xtest}|0) = 1.279$$

Stage 5

Calculating *the Posterior value*, in general,the posterior value of class C is defined in equation (5).

$$P(\text{C}|\text{Xtest}) = P(C) \cdot P(\text{Xtest}|\text{C}) \quad (5)$$

Based on equation 5, to calculation of *the posterior* for each class is as follows:

For class 0:

$$\text{P}(0|\text{Xtest}) = 0.4 \cdot 3034927.11,$$

$P(0 \mid Xtest) = 1213970.84,$

For 1st grade:

$P(1 \mid Xtest) = 0.6 \cdot 1.279,$

$P(1 \mid Xtest) = 7.675.$

Stage 6

The last factor in predicting the Xtest class is determining it based on the previous posterior results. The requirements for determining the class can be seen in equation (6).

$$\bar{y} = \{0, \text{ for } P(C = 0 \mid X_{test}) > P(C = 1 \mid X_{test}) \quad (6)$$

for $P(C = 0 \mid X_{test}) \leq P(C = 1 \mid X_{test})$ based on the results of the previous calculation, because $P(C = 0 \mid X_{test}) > P(C = 1 \mid X_{test}) \Leftrightarrow 1213970.84 > 7.675$ so the y label for the X_{test} class is 0.

3. Result and Discussion

This section discusses the implementation of the interface, application testing with various conditions, and a discussion.

3.1 Result

This website-based application is designed to support financial institutions or banks in making decisions to accept or reject customers applying for credit, giving to prospective customers in a friendly and reliable way. By using this web-based application, the agency team can easily provide information to provide decisions to clients. In addition, this application also contributes to increasing work productivity from financial institutions and banks to provide fast and precise results for customers.

3.1.1 Input Data Interface

Figure 2 describes the display of some features of the decision-making system interface that are used to stimulate the probability of receiving credit from customers. Here is an explanation of the features.

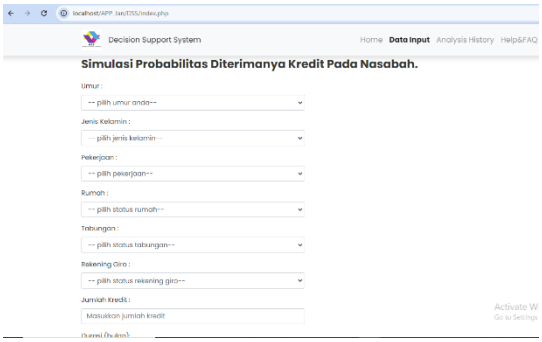

Figure 2. Input Data Interface

Figure 2 shows the main interface of the Naïve Bayes method based on the recommendation system for applying for consumptive credit. This page is designed to simulate the probability of receiving a customer's credit application. The system provides data input forms for age, gender, occupation, home status, savings, checking account, credit amount, and duration of application. The interface has intuitive navigation, such as the Home menu, Data Input, Analysis History, and Help & FAQ, to make it easier for users to explore the system's features. This feature facilitates financial institutions or users entering customer data to obtain credit eligibility predictions quickly and accurately based on probabilistic models. The user-friendly design ensures easy accessibility and efficiency for users in utilizing the system for decision-making.

3.1.2 Prediction Result Interface

Figure 3 shows the prediction results of the Naïve Bayes method based on the recommendation system for applying for consumptive credit.

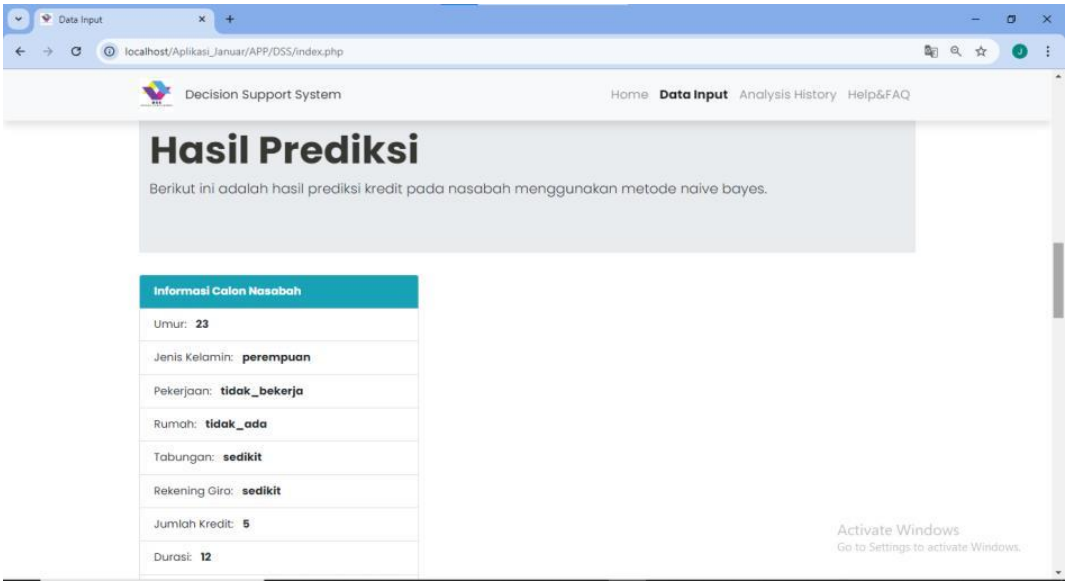

Figure 3. Prediction Result Interface

Figure 3 shows the page display of the prediction results of the credit application recommendation system based on the Naïve Bayes method. This page displays information about potential customers, such as age, gender, occupation, home status, savings, checking account, credit amount, and duration of application in a structured table format. Predictions are made based on the data entered, with a system assessing customer creditworthiness.

This display is designed to provide clear and user-friendly feedback, making it easier for users to verify input data and understand the analysis results. This page supports a data-driven credit decision-making process quickly and accurately.

3.1.3 Decision Result Interface

Figure 4 shows the results of evaluating the prediction system for recommending consumptive credit applications based on the Naïve Bayes method.

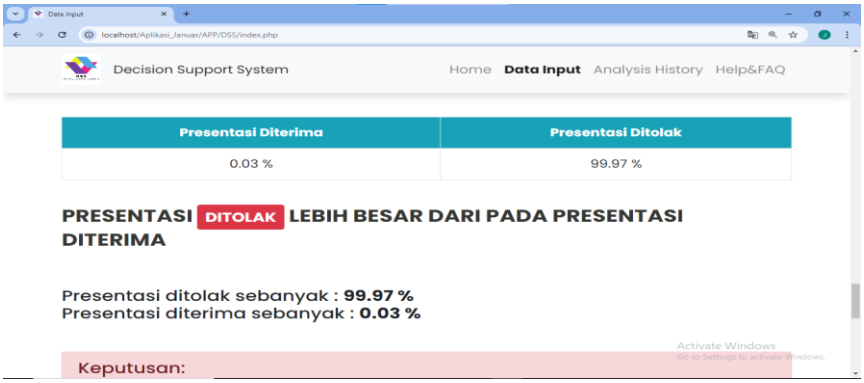

Figure 4. Decision Result interface

Figure 4 shows the results of evaluating credit application predictions with a system based on the Naïve Bayes method. The analysis shows a probability of acceptance of 0.03% and rejection of 99.97%, with the conclusion that the credit application is rejected based on the highest probability. This information is presented visually with bold and informative elements, making it easier for users to understand the system's decisions. This page is designed to help financial institutions make decisions objectively and based on probabilistic data.

3.2 Discussion

The tuning results in this study show that the optimal var_smoothing value is 1e-09, with an average cross-validation score of 70%, indicating that this model has good prediction ability on validation data. The best models achieve 70.50% accuracy on the test set. While this accuracy is good, several factors need to be considered such as data imbalances, where categories such as homeownership and savings types are unbalanced, which can affect the performance of Naïve Bayes' model which assumes feature independence; feature engineering, which can improve performance by replacing the 'unknown' category with a median or mode value; and data handling is not balanced with oversampling techniques such as SMOTE or undersampling.

Additionally, comparisons with other algorithms such as Decision Tree or SVM can provide additional perspective on model performance. Although Naïve Bayes provides adequate results with a cross-validation score of 70% and an accuracy of 70.50%, there is potential for improvement through advanced techniques. The confusion matrix in this study allows the calculation of evaluation metrics with an accuracy of 70.50%, accuracy of 92.61%, recall of 64.57%,

and F1 Score of 75.83%. Quite good accuracy shows the model's performance in predicting credit applications, but there is room for improvement, especially in dealing with data imbalances. High precision indicates accuracy in positive predictions, while lower recall values indicate many false negatives, signaling the need for improvement in detecting credit acceptance cases. The F1 Score of 75.83% shows a balance between precision and recall, but further improvements, especially in recall, are still needed.

The classification report shows that although the Naïve Bayes model effectively predicts accepted credit applications, improvements in detecting rejected ones are still needed. Steps that can be taken include data balancing by oversampling or undersampling, feature engineering, applying ensemble techniques such as Random Forest or Gradient Boosting, and experimenting with other algorithms to improve the accuracy and reliability of the credit application recommendation system.

4. Conclusion

This study succeeded in developing a recommendation system for applying for consumptive credit based on the Naïve Bayes method. Performance evaluation results included accuracy of 70.50%, precision of 92.16%, recall of 64.57%, and F1-score of 75.83%. This system is designed as a web-based application to simplify the process of simulating creditworthiness predictions efficiently and data-based.

This system is expected to help financial institutions, especially banks in Indonesia, reduce the risk of bad loans and improve decision-making efficiency. This research also contributes as a reference for further development in applying data mining methods for classification in similar fields.

References

[1] A. S. Aphale and D. S. R. Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval," *Int. J. Eng. Res. Technol.*, vol. 9, no. 8, pp. 991–995, 2020, [Online]. Available: www.ijert.org.

[2] M. K. Nallakaruppan *et al.*, "Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence," *J. Econ. Financ. Adm. Sci.*, no. Ml, pp. 1–18, 2024.

[3] A. A. Hussin Adam Khatir and M. Bee, "Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination?," *Risks*, vol. 10, no. 9, 2022, doi: 10.3390/risks10090169.

[4] A. Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank," *J. Econ. Financ. Adm. Sci.*, vol. 22, no. 42, pp. 3–24, 2017, doi: 10.1108/JEFAS-02-2017-0039.

[5] O. J. Okesola, K. O. Okokpujie, A. A. Adewale, S. N. John, and O. Omoruyi, "An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach," *Proc. - 2017 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2017*, pp. 228–233, 2018, doi: 10.1109/CSCI.2017.36.

[6] S. Hua *et al.*, "An FTwNB Shield: A Credit Risk Assessment Model for Data Uncertainty and Privacy Protection," *Mathematics*, vol. 12, no. 11, pp. 1–17, 2024, doi: 10.3390/math12111695.

[7] Y. Rosmansyah, B. L. Putro, A. Putri, N. B. Utomo, and Suhardi, "A simple model of smart learning environment," *Interact. Learn. Environ.*, 2022, doi: 10.1080/10494820.2021.2020295.

[8] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput. Sci.*, vol. 6, pp. 1–43, 2020, doi: 10.7717/PEERJ-CS.267.

[9] A. Husejinović, "Credit card fraud detection using naive Bayesian and c4.5 decision tree classifiers," *Period. Eng. Nat. Sci.*, vol. 8, no. 1, pp. 1–5, 2020.

[10] F. Itoo and M. Satwinder, "¨ ve Bayes Comparison and analysis of logistic regression , Naı and KNN machine learning algorithms for credit card fraud detection," *Int. J. Inf. Technol.*, 2020, doi: 10.1007/s41870-020-00430-y.