STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, May 2026, pp 4164–4171.
Published online in International Academic Press (www.IAPress.org)

Finding Category Value Using Mean Shift Clustering to Optimize Naïve Bayes Classification

Berlian Rahmy Lidiawaty[1], Arip Ramadan[1], Tita Ayu Rospricilia[1], Najma Attaqiya Alya[2,3], Dwi Rantini[2,4,*], Alhassan Sesay[5]

[1]*Information System Study Program, School of Industrial and System Engineering, Telkom University, Surabaya Campus, Jl. Ketintang No.156, Surabaya 60231, East Java, Indonesia*
[2]*Data Science Technology Study Program, Department of Engineering, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, 60115, Indonesia*
[3]*Institute of Statistics and Data Science, Faculty of Science, National Tsing Hua University, Hsinchu, Taiwan*
[4]*Research Group of Data-Driven Decision Support System, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, 60115, Indonesia*
[5]*Faculty of Transformative Education, the United Methodist University, Sierra Leone*

Abstract The Naïve Bayes classifier is a simple classification method that can make predictions quickly and accurately by considering the independent variables separately from the class. However, in the Naïve Bayes classifier, each independent variable must be divided into several categories, while some of the data remain continuous and uncategorized. Therefore, this study proposes a measurable and precise model to categorize these independent variables effectively. The main objective is to develop a categorization model for independent variables using the Mean Shift clustering algorithm to optimize the performance of the Naïve Bayes classifier. To implement the proposed model, experiments were conducted on two types of datasets. The first dataset contains 191 records with 4 attributes and 6 classes, while the second dataset consists of 2,000 records with 7 attributes and 2 classes. In both datasets, several attributes were initially uncategorized and were categorized using the Mean Shift clustering method. The Mean Shift approach successfully grouped the uncategorized attributes into meaningful categories. In the first dataset, the accuracy of the proposed categorical Naïve Bayes classifier reached 80.1%, representing an improvement of 5.74%. Furthermore, in the second dataset, the accuracy increased to 84.25%, marking a 3% enhancement. The results of this research are expected to contribute to the field of education, especially in the subfield of machine learning.

Keywords Naïve Bayes, Mean Shift Clustering, Classification, Optimize, Education

DOI: 10.19139/soic-2310-5070-3161

1. Introduction

Classification is one of the key aspects in data processing and machine learning which has wide applications in various fields, including data analysis, pattern recognition, and decision making. In the context of classification, choosing the right classification method is very important because it can have a significant impact on the final results. Since classification techniques were developed, there are several classification methods that are commonly used in data mining and as a basis for predicting class determination, including Miscellaneous, decision trees, rule-based, Lazy and Function, Bayes, and so on.

One of the simple yet robust method of classification is Naïve Bayes algorithm, which is an efficient and effective classification method in solving problems in machine learning [1]. The use of Naïve Bayes is solving

*Correspondence to: Dwi Rantini (Email: dwi.rantini@ftmm.unair.ac.id). Data Science Technology Study Program, Department of Engineering, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, 60115, Indonesia

ISSN 2310-5070 (online) ISSN 2311-004X (print)

image classification [2], banking [3], and Husejinovich cases [2, 4, 5, 6]. However, with the advantages of this method, the level of NB accuracy still tends to be low compared to other classification methods [7].

Apart from the classification methods mentioned above, there are several classification methods including Logistic Regression (LR) with an accuracy of 90.57%, Light Gradient Boosting Machine (LGBM) with an accuracy of 90.55%, Stochastic Gradient Descent (SGD) with an accuracy of 90.6%, Random Forest (RF) with an accuracy of 89.84%, AdaBoost (ADB) with an accuracy of 89.30%, and SVM with an accuracy of 67.13%. If the accuracy values are compared, Naïve Bayes is not the classification method that has the lowest accuracy value, but it is quite low [8].

There is something that makes the accuracy value low, namely treating each label individually, causing the accuracy value to drop when there is zero frequency in one of the categories [9, 10]. Currently, many researchers are trying to carry out smoothing to overcome this problem [11]. One of the causes of zero frequency occurrences is the inappropriate assignment of categories to a particular attribute [12]. Therefore, the most important things about Naïve Bayes classification requires data that has been categorized and assumptions regarding the criteria that constitute its attributes [13]. In carrying out Naïve Bayes classification, categories are needed for each variable or predictor and look for data patterns to increase accuracy [1]. As an example, when describing the age attribute, the range of its values can vary from the youngest to the oldest. The challenge lies in how to appropriately group the data based on the available information. Should it be categorized into three groups: adolescent, young, and elderly? Or should it be grouped into a different number of categories?

The main objectives of this research is determining the categories of one or more attributes, so the data can be used in the Naïve Bayes classification method. One approach to data categorization is the use of clustering methods [14]. Moreover, clustering the data can solve the problem of imbalanced data in classification [15]. In substantiating this theory, this research utilized two datasets as experimental instruments to ascertain the optimization of attribute clustering using a clustering method prior to processing it compared to direct processing with the Naïve Bayes classifier available in a library. The cluster method that proposed in this research is Mean Shift because it does not require assumptions about the number of clusters [16]. This research used two datasets. The first dataset comprises customer data with 191 records, 4 attributes, and 6 classes [17]. Conversely, the second dataset still pertains to customer data but with a larger quantity, specifically 2,000 records, encompassing 6 attributes and 2 classes.

2. Literature Review

In this chapter, the focus will be on elucidating how previous studies have optimized Naïve Bayes classification methods and how relevant research has concentrated on determining attribute categories. Additionally, this chapter also expounds upon various methods and equations utilized in this study.

2.1. Improving the accuracy of Naïve Bayes

Multilabel clustering with the Naïve Bayes approach for each variable has been carried out by Kim *et al.* (2020). The proposed Multilabel Naïve Bayes classifier considered the existence of relationships between variables. In their research, it has been proven that the proposed method performs better than conventional methods. To increase the accuracy of the Naïve Bayes method, there is a technique that needs to be considered, namely categorizing the data according to its data type [18]. Apart from this, to further enhance the accuracy of the Naïve Bayes method, it can be combined with other techniques. This approach has been implemented by Setyaningsih and Listiowarni (2021), who combined the Naïve Bayes method with the Chi-Square feature selection and Laplace smoothing. This research does not focus on data types because all data in this study are numerical. Instead, it focuses on how to divide the range of data for each variable and determine appropriate categories.

2.2. Determining Categories for Classification

(Miao et al. 2023b) have conducted research using the K-Means clustering method to determine the categories of human factors that cause accidents in coal mining. The K-Means method requires prior information on the

number of clusters for further analysis [19]. In the K-Means clustering approach, determining the optimal number of clusters (K) often requires specific techniques, methods, or algorithms [19, 20].

The K-Means algorithm divides a set of N samples X into K disjoint clusters C, each represented by the mean μ_j of the samples within that cluster. These means are referred to as the cluster *centroids*; note that they are generally not points from X itself, although they exist in the same feature space. The objective of the K-Means algorithm is to determine centroids that minimize the inertia, or within-cluster sum-of-squares criterion, as expressed in Equation 1.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \|x_i - \mu_j\|^2 \tag{1}$$

Inertia can be interpreted as a measure of how internally coherent the clusters are.

This research utilizes a range of non-parametric data, which will be categorized into multiple classes using the Naïve Bayes method. Therefore, one of the techniques employed in this study is the Mean Shift clustering method, which does not require assumptions about the number of clusters in the dataset. It operates by performing density estimation and iteratively locating the local maxima of the kernel function. Moreover, the dataset used in this study comprises 191 records, which still lies within the optimal range for implementing the Mean Shift clustering method [16]. However, the model developed in this research can potentially reduce the computational complexity of the Mean Shift process, allowing it to be applied to larger datasets. To further understand the Mean Shift clustering approach, readers can refer to [21].

$$M_A = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

2.3. Gaussian Naïve Bayes as The Comparison and Categorical Naïve Bayes

Gaussian Naive Bayes (GNB) is a prominent classification algorithm within the Naive Bayes family, specifically designed for modeling and classifying data with numeric features assumed to follow a Gaussian (Normal) distribution. The term "Naive" conveys the algorithm's fundamental assumption that each feature in the dataset is considered independent of others, meaning the value of one feature is assumed not to depend on the values of other features.

The likelihood of a feature x_i given class y is expressed as:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3}$$

Key aspects of Gaussian Naive Bayes include the assumption of a Gaussian distribution for each observed class, positing that feature values within each class follow a normal distribution. Despite potential deviations from this idealized scenario in real-world datasets, the algorithm maintains the naive assumption of feature independence. Gaussian Naive Bayes estimates parameters for the Gaussian distribution, including the mean μ_y and variance σ_y^2, for each feature within each class. Using Bayes' Theorem, GNB calculates class probabilities based on feature values, considering posterior probabilities, prior probabilities, and likelihoods. After training, the GNB model can classify new data instances by computing probabilities for each class and selecting the class with the highest probability. While GNB relies on relatively simplistic assumptions and may exhibit reduced accuracy if these assumptions are violated, it is often effective and efficient when the assumptions align with the characteristics of the dataset.

Categorical Naïve Bayes is employed for categorical data. For every feature i in the training set X, CategoricalNB estimates a categorical distribution for i conditioned on the class y. The probability is computed as:

$$P(x_i = t \mid y = c; \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i} \tag{4}$$

where $J = \{1, \ldots, m\}$ denotes the index set of samples, N_{tic} is the number of samples in class c with feature i taking value t, N_c is the total number of samples in class c, α is the smoothing parameter, and n_i is the number of possible categories for feature i.

3. Methods

This chapter will delineate the methods or steps undertaken in this research. In broad strokes, the methods employed in this study are outlined in Figure 1. The process begins with understanding the data structure, implementing the Mean Shift clustering, and calculating accuracy. Further details on each of these stages will be expounded upon in their respective sections.

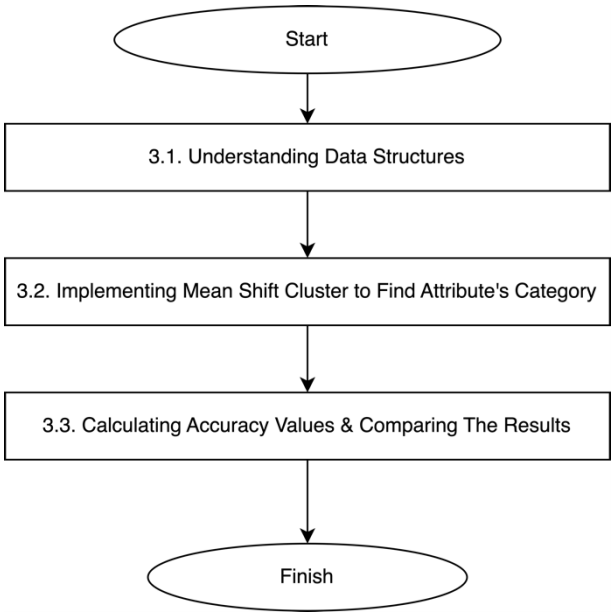

Figure 1. The methods of the research

The explanation of Figure 1 is presented in the following description.

3.1. Understanding data structures

In the process of understanding data structures, you need to look at the data in each dataset. The main goal of understanding data structures is to determine which attributes to break down and how to determine a complete classifier. In general, the data structure and data samples from each dataset can be seen in Table 1.

This study employs two datasets to evaluate the effectiveness of Mean Shift–based categorization in improving Naïve Bayes classification performance. Dataset 1 is a transactional customer dataset derived from the Recency–Frequency–Monetary (RFM) model. It consists of 191 records, four attributes, and six class labels. The attributes include Recency, Frequency, and Monetary, which are continuous numerical variables representing customer purchasing behavior, and a target Cluster variable representing customer segments. All predictive attributes in this dataset are initially continuous and uncategorized, making the dataset suitable for evaluating the proposed categorization framework. Dataset 2 is a larger benchmark dataset consisting of 2,000 records, seven attributes, and two class labels. This dataset contains a mixture of categorical and continuous attributes. While some attributes are already categorical, several numerical attributes remain continuous and require discretization before being processed by the Naïve Bayes classifier. This dataset is used to assess the scalability and generalizability of the proposed Mean Shift–based categorization approach on data with a higher volume and mixed attribute

Table 1. Structure and Characteristics of the Datasets

Indicators	**Dataset 1**	**Dataset 2**
Number of records	191	2,000
Number of attributes	3	7
Uncategorized attributes	Recency, Frequency, Monetary	Age, Income
Categorical attributes	–	Sex, Marital Status, Education, Occupation, Settlement Size
Class label	Cluster (6 classes: 1–6)	Label (2 classes: 0,1)
Attribute type	Numerical	Mixed (Numerical and Categorical)
Sample values	Recency Frequency Monetary Cluster 1522 1 32773758 2 941 1 2739000 1 1940 1 9868150 6 1885 1 9420250 6 787 4 4746000 6	Sex Marital status Age Education Income Occupation Settlement size label 0 0 67 2 124670 1 2 0 1 1 22 1 150773 1 2 1 0 0 49 1 89210 0 0 0 0 0 45 1 171565 1 1 1 0 0 53 1 149031 1 1 1

types. Both datasets were selected to reflect common real-world classification scenarios in which Naïve Bayes is applicable but constrained by its requirement for categorical input features. By using datasets with different sizes, attribute compositions, and class distributions, this study demonstrates the robustness of the proposed Mean Shift clustering method for transforming continuous attributes into meaningful categorical representations.

3.2. Implementing Mean Shift Cluster to Find Attribute's Category

In this stage, it is discerned that in the first dataset, there are three attributes that remain in an uncategorical form. Therefore, these three attributes require the application of the Mean Shift method. Meanwhile, in the second dataset, two attributes need to be categorized first. For each of these uncategorized attributes, a search for the optimal cluster count is conducted, and categorization is performed using Equation 2. The aim of this step is to determine the number of categories required for each attribute and to assign each value of an attribute to its corresponding category. For the implementation, the following steps were performed for each continuous attribute:

- Data Preparation: The attribute values were reshaped into a one-dimensional array compatible with the Mean Shift algorithm.
- Kernel Selection and Bandwidth: A Gaussian kernel was employed to estimate the density of data points. The bandwidth, which controls the radius of the kernel and therefore the level of smoothing in the density estimation, was set to None to allow the algorithm to automatically estimate the optimal bandwidth from the data. This ensures that the identified clusters represent natural groupings in the attribute values.
- Mean Shift Clustering: Using the prepared data and kernel, the algorithm iteratively shifted points toward areas of higher density. Each point was eventually assigned a cluster label corresponding to the mode it converged to.
- Mapping to Categorical Labels: After clustering, each unique cluster label was considered a category. Data points within the same cluster were assigned the same category label (e.g., 0, 1, 2, . . .). This transformed the original continuous attribute into a categorical attribute suitable for subsequent analysis using classification methods.
- Determining Number of Categories: The number of categories for each attribute is equal to the number of clusters identified by Mean Shift, which depends on the data distribution and the automatically estimated bandwidth.
- The resulting categorical representation ensures that each attribute is grouped in a way that reflects the inherent structure of the data, enabling more meaningful downstream analyses such as Naïve Bayes classification.

3.3. Calculating Accuracy Values & Comparing the Results

In calculating accuracy, Equation 5 is employed. In this equation, the accuracy value A is derived by dividing the number of true predicted values (TP) by the total number of data points (n). A true predicted (TP) value is assigned to a record when the label of a testing dataset record is correctly predicted by the developed system. The allocation of data between the training and testing sets in the model development process follows an 80% training data and 20% testing data distribution.

$$A = \frac{TP}{n} \tag{5}$$

In the subsequent accuracy comparison process, initially, data with uncategorized attributes is subjected to classification using Gaussian Naïve Bayes Equation 3. Furthermore, for data that has already been categorized, it is retested using both Gaussian Naïve Bayes and Categorical Naïve Bayes classifiers. Subsequently, a more in-depth analysis is conducted concerning these accuracy outcomes.

4. Result and Discussion

This chapter will elucidate the findings from experiments conducted regarding the process of categorizing attributes using Mean Shift, the developed mathematical model, and the accuracy results compared with other Naïve Bayes methods

4.1. Data Transformation

This study utilizes two datasets, where each dataset is examined to determine whether there are attributes that remain uncategorized. In the first dataset, all three attributes, or in fact all attributes, are yet to be categorized. Consequently, all attributes in Dataset 1 are clustered using the Mean Shift algorithm. Meanwhile, in the second dataset, several attributes have been categorized, representing specific categories. However, two attributes, namely *Age* and *Income*, are found to remain uncategorized. For each attribute that remains uncategorized, clustering is performed using Mean Shift, depending on the range of its data, as illustrated in Table 1.

From Table 1, it can be observed that several uncategorized attributes have varying ranges of values. In the first dataset, the *Recency* attribute, with values ranging from 29 to 3,005 using the Mean Shift clustering algorithm, is categorized into 2 clusters. The *Frequency* attribute, with a range from 1 to 25, is categorized into 9 clusters, while the *Monetary* attribute, with values ranging from 211,250 to 96,126,500 (in Indonesian Rupiahs), is categorized into 7 clusters. Meanwhile, in Dataset 2, it is necessary to categorize the *Age* and *Income* attributes. The *Age* attribute, with values ranging from 18 years old to 76 years old, is categorized into 2 clusters. Similarly, the *Income* attribute, with a value range between 35,832 and 309,362 (in USD), is also categorized into 2 clusters.

Table 2. Dataset Structure and Samples

Indicators	**1st Dataset**	**2nd Dataset**
Attributes	3 (all uncategorized)	7 (2 uncategorized)
Class	Column name: Cluster 2 (1,2,3,4,5,6)	Column name: Label 2 (0,1)
Samples	Recency Frequency Monetary Cluster 1522 1 32773758 2 941 1 2739000 1 1940 1 9868150 6 1885 1 9420250 6 787 4 4746000 6	Sex Marital status Age Education Income Occupation Settlement size label 0 0 67 2 124670 1 2 0 1 1 22 1 150773 1 2 1 0 0 49 1 89210 0 0 0 0 0 45 1 171565 1 1 1 0 0 53 1 149031 1 1 1

4.2. Accuracy Comparison

When the attributes have been categorized, using the Gaussian Naïve Bayes classifier is actually ineffective, and the accuracy value decreases, as shown in Table 3. Meanwhile, by using the categorical Naïve Bayes classifier, the accuracy value increases. The increase in accuracy is also higher for data with smaller dimensions but with more labels (first dataset).

Table 3. Comparison of Accuracy Values

Dataset	**Gaussian Classifier**		**Categorical Classifier**	**Improvement**
	Uncategorical Data	**Categorical Data**	**Categorical Data**	
First	94.87%	74.36%	80.1%	5.74%
Second	93.75%	81.25%	84.25%	3%

In the second dataset, when the number of data classes is lower as it seen in Table 3, even with an increasing dimensionality, the classification results tend to be more stable. Consequently, the improvement is not as pronounced as observed during experimentation with the first dataset. This indicates that, although the Mean Shift–based categorization successfully transforms continuous attributes into discrete categories, its impact on classification performance depends on the characteristics of the dataset, particularly the number of classes and the initial separability of the data. When the original continuous features already exhibit good class discrimination, the benefit of categorization becomes more limited. Therefore, in future research endeavors, additional strategies are required, focusing on optimizing the number of categories for a given attribute. This is necessary because a limitation in this study is that the proposed model has not yet surpassed the accuracy achieved by using the Gaussian Naïve Bayes classifier without prior data categorization.

5. Conclusion and Suggestions

The primary objective of this research is to determine the number of categories for uncategorized data attributes to optimize the Naïve Bayes classification method. The determination of the number of categories employs the Mean Shift method. Consequently, several attributes are categorized based on the number of clusters obtained through the Mean Shift method. In the first dataset, an accuracy of 80.1% is achieved, while the second dataset attains the highest accuracy of 84.25%. Both results exhibit improvements compared to the classification of categorized data using Gaussian Naïve Bayes. However, further observations are necessary in this study to enhance the accuracy values to be on par with the accuracy of uncategorized data using Gaussian Naïve Bayes.

REFERENCES

1. Ko, Youngjoong. 2017. “How to Use Negative Class Information for Naive Bayes Classification.” *Information Processing & Management* 53 (6): 1255–68.
2. Ramesh Kumar, P, and A Vijaya. 2022a. “Naïve Bayes Machine Learning Model for Image Classification to Assess the Level of Deformation of Thin Components.” *Materials Today: Proceedings* 68: 2265–74. https://doi.org/10.1016/j.matpr.2022.08.489
3. Husejinovic, Admel. 2020a. “Credit Card Fraud Detection Using Naive Bayesian and C4.5 Decision Tree Classifiers.” Husejinovic, A.(2020): 1–5.
4. Phoenix, Peter, Richard Sudaryono, and Derwin Suhartono. 2021. “Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier.” *Procedia Computer Science* 179: 498–506.
5. Husejinovic, Admel. 2020b. “Credit Card Fraud Detection Using Naive Bayesian and C4.5 Decision Tree Classifiers.” Husejinovic, A.(2020): 1–5.
6. Salmi, Nafizatus, and Zuherman Rustam. 2019. “Naïve Bayes Classifier Models for Predicting the Colon Cancer.” In *IOP Conference Series: Materials Science and Engineering*, 546:052068. IOP Publishing.
7. Panigrahi, Ranjit, Samarjeet Borah, Akash Kumar Bhoi, Muhammad Fazal Ijaz, Moumita Pramanik, Rutvij H Jhaveri, and Chiranji Lal Chowdhary. 2021. “Performance Assessment of Supervised Classifiers for Designing Intrusion Detection Systems: A Comprehensive Review and Recommendations for Future Research.” *Mathematics* 9 (6): 690.
8. Muneer, Amgad, and Suliman Mohamed Fati. 2020. “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter.” *Future Internet* 12 (11): 187.

9. Kim, Hae-Cheon, Jin-Hyeong Park, Dae-Won Kim, and Jaesung Lee. 2020. "Multilabel Naïve Bayes Classification Considering Label Dependence." *Pattern Recognition Letters* 136: 279–85.
10. Sarkar, A M Jehad, Young-Koo Lee, and Sungyoung Lee. 2010. "A Smoothed Naive Bayes-Based Classifier for Activity Recognition." *IETE Technical Review* 27 (2): 107–19.
11. Upadhyay, Kamlesh, Prabhjot Kaur, SVAV Prasad, and Lingayas Vidyapeeth. 2021. "State of the Art on Data Level Methods to Address Class Imbalance Problem in Binary Classification." *GIS Science Journal* 8 (3): 875–903.
12. Vujović, Ž. 2021. "Classification Model Evaluation Metrics." *International Journal of Advanced Computer Science and Applications* 12 (6): 599–606.
13. Rish, Irina. 2001. "An Empirical Study of the Naive Bayes Classifier." In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3:41–46.
14. Miao, Dejun, Wenhao Wang, Yueying Lv, Lu Liu, Kaixin Yao, and Xiuhua Sui. 2023a. "Research on the Classification and Control of Human Factor Characteristics of Coal Mine Accidents Based on K-Means Clustering Analysis." *International Journal of Industrial Ergonomics* 97: 103481. https://doi.org/10.1016/j.ergon.2023.103481
15. Farshidvard, A, F Hooshmand, and S A MirHassani. 2023. "A Novel Two-Phase Clustering-Based under-Sampling Method for Imbalanced Classification Problems." *Expert Systems with Applications* 213: 119003. https://doi.org/10.1016/j.eswa.2022.119003
16. Ren, Yazhou, Carlotta Domeniconi, Guoji Zhang, and Guoxian Yu. 2014. "A Weighted Adaptive Mean Shift Clustering Algorithm." In *Proceedings of the 2014 SIAM International Conference on Data Mining*, 794–802. SIAM.
17. Ayu, Tita, Syurfah Ayu Ithriah, and Amalia Anjani. 2020. "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada CV Tita Jaya." *Jurnal Informatika Dan Sistem Informasi* 1 (3): 699–708.
18. Vishwakarma, Monika, and Nishtha Kesswani. 2023. "A New Two-Phase Intrusion Detection System with Naïve Bayes Machine Learning for Data Classification and Elliptic Envelop Method for Anomaly Detection." *Decision Analytics Journal* 7: 100233.
19. Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek. 2003. "The Global K-Means Clustering Algorithm." *Pattern Recognition* 36 (2): 451–61.
20. Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. "The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation." *Electronics* 9 (8): 1295.
21. Comaniciu, Dorin, and Peter Meer. 2002. "Mean Shift: A Robust Approach toward Feature Space Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (5): 603–19.
22. Miao, Dejun, Wenhao Wang, Yueying Lv, Lu Liu, Kaixin Yao, and Xiuhua Sui. 2023b. "Research on the Classification and Control of Human Factor Characteristics of Coal Mine Accidents Based on K-Means Clustering Analysis." *International Journal of Industrial Ergonomics* 97: 103481. https://doi.org/10.1016/j.ergon.2023.103481.
23. Setyaningsih, E. R., and I. Listiowarni. 2021. "Categorization of Exam Questions Based on Bloom Taxonomy Using Naïve Bayes and Laplace Smoothing." *In 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 330–333. https://doi.org/10.1109/EIConCIT50028.2021.9431862.