GAUSSIAN MIXTURE MODEL FOR IMMIGRATION RESIDENCE PERMIT IN KEDIRI IMMIGRATION OFFICE

Priati Assiroj, Besse Hartati, Isidorus Anung, Nurul Maharani, Rasona Sunara, Galuh Boy, Masdar Bakhiar, Atsil Syah, Tiara Okta

Politeknik Imigrasi
Email: priati.assiroj@poltekim.ac.id

Abstrak

Data mining melibatkan pemanfaatan teknik pengenalan pola, konsep matematika, dan algoritme pembelajaran mesin untuk mengungkap tren dan hubungan yang berharga di dalam kumpulan data. Analisis klaster adalah metode fundamental dalam penggalian data, yang membantu dalam mengidentifikasi klaster-klaster yang memiliki kemiripan. Penggunaan algoritme pengelompokan telah menarik perhatian di berbagai domain analisis data, dan teknik pengelompokan berbasis model telah mendapatkan daya tarik dalam mengelola data yang kompleks. Pendekatan pengelompokan berbasis model yang diusulkan menggabungkan berbagai variabel non-kontinyu menggunakan Generalized Linear Latent Variable Model (GLLVM) dan Gaussian Mixture Model (GMM) untuk menganalisis data campuran. Pendekatan ini telah diperluas untuk mempertimbangkan variabel laten dengan distribusi non-Gaussian, yang menghasilkan pembentukan cluster secara alami. Penelitian ini menyoroti pentingnya memanfaatkan teknik data mining untuk mengungkapkan wawasan dan pola dalam set data yang kompleks, yang berkontribusi pada peningkatan manajemen data dan pengambilan keputusan yang tepat. Dataset yang digunakan adalah data izin tinggal, menghasilkan 2 klaster yang sesuai dengan DBI, memberikan informasi bahwa Klaster 1 didominasi oleh warga negara asing (WNA) yang disponsori oleh perusahaan untuk tinggal atau berkunjung ke wilayah Kediri dan sekitarnya. Di sisi lain, Cluster 2 dicirikan oleh WNA yang disponsori oleh perorangan atau tanpa sponsor, dengan tujuan di luar wilayah Kabupaten Kediri.

Abstract

Data mining involves the utilization of pattern recognition techniques, mathematical concepts, and machine learning algorithms to uncover valuable trends and relationships within datasets. Cluster analysis is a fundamental method in data mining, aiding in identifying clusters with similarities. The use of clustering algorithms has garnered attention across various data analysis domains, and model-based clustering techniques have gained traction in managing complex data. The proposed model-based clustering approach combines various non-continuous variables using the Generalized Linear Latent Variable Model (GLLVM) and Gaussian Mixture Model (GMM) to analyze mixed data. This approach has been extended to consider latent variables with non-Gaussian distributions, resulting in the natural formation of clusters. This research highlights the importance of utilizing data mining techniques to reveal insights and patterns within complex datasets, contributing to enhanced data management and informed decision-making. The dataset used is residency permit data, resulting in 2 clusters that are in line with the DBI, providing information that Cluster 1 is predominantly composed of foreign nationals (WNAs) sponsored by corporations for residency or visits to the Kediri area and its surroundings. On the other hand, Cluster 2 is characterized by WNAs sponsored by individuals or without sponsors, with their destinations lying outside the Kediri district.

1. Introduction

Since ancient times, human movement from one location to another, known as migration, has been a common phenomenon. The presence of push and pull factors is the

rationale behind this phenomenon. The main driving force is often the search for employment (known as the push-pull theory) [1].

The Directorate General of Immigration is a governmental institution responsible for carrying out tasks and functions that encompass not only aspects of individuals entering and leaving Indonesian territory, as well as law enforcement, but also providing services to the public. This involves services for Indonesian citizens through passport issuance, and for foreigners through immigration residency permits and related facilities. In the era of the fourth industrial revolution, the government sector must embrace technology effectively to enhance the efficiency and effectiveness of public services and support good governance.

Excellence in service achievements can be attained through adjustments by government institutions, where innovation is the key to providing public services to the community. For efficient service delivery, the government needs to manage transparent and effective governance, which also acknowledges the authority possessed by these institutions [2].

The utilization of technology in public services conducted by the Directorate General of Immigration is not limited to Indonesian citizens (WNI) but also involves foreign nationals (WNA) since both have their respective responsibilities and functions. Immigration Residency Permits are issued to foreign nationals to stay in Indonesia. These permits are issued by Immigration Officials as well as officials from the Ministry of Foreign Affairs (Republic of Indonesia, 2011). Therefore, it can be understood that foreign nationals arriving and residing in Indonesia are required to possess valid and current residency permits.

The increasing cross-border mobility and the need to manage data related to the entry of foreign nationals (FNs) into Indonesia have become increasingly important issues. Every year, thousands of FNs enter Indonesian territory for various purposes, including business visits, tourism, education, or employment. Data processing plays a crucial role in the advancement of information technology. In almost all sectors of work, there is a collection of informational data. This data holds the potential to support analyses in various job tasks [3]. FN entry data, including residency permit data, constitute valuable assets in the management of national security and the formulation of effective policies.

In this context, the utilization of Data Mining techniques emerges as a potential solution for analyzing residency permit data at the Class II Non TPI Kediri Immigration Office. Data Mining techniques enable the identification of hidden patterns and relationships within large and complex datasets, which are difficult to accomplish manually. By applying these techniques, residency permit data can be processed more efficiently and effectively, resulting in deeper and more meaningful insights for various stakeholders, including immigration agencies, security institutions, and other relevant parties.

Therefore, the purpose of this study is to investigate the potential utilization of Data Mining techniques in the analysis of residency permit data at the Class II Non TPI Kediri Immigration Office. By uncovering hidden insights within the data, this research aims to contribute to improving the management of residency permit data, supporting better decision-making, and detecting potential risks and threats to national security. As a result, this study holds high relevance in addressing the challenges of managing FN entry data and national security in the digital era.

Data mining is defined as a process of identifying new relationships, patterns, and meaningful trends by utilizing pattern recognition techniques such as statistical and

mathematical methods [4][5]. The term data mining is used to refer to the discovery of knowledge within a database. Data mining is a process that involves the application of analytical methods [6], mathematical concepts, artificial intelligence, and machine learning to extract and identify valuable information and detailed insights from various large-scale database sources [7]. A book [8] on the fundamental concepts and methods of data mining assists readers in understanding how to apply data mining tools and techniques in the analysis of diverse data.

Data Mining techniques provide a powerful approach to extract valuable information from large and complex datasets. It involves the use of computational algorithms to identify patterns, relationships, and hidden insights within the data. One common method in Data Mining is Clustering. Clustering is used to group data into clusters with similarities. A comprehensive guide to machine learning concepts and practices in the context of data mining is provided in the book [9].

As an efficient technique for unsupervised data clustering, various clustering algorithms have been proposed and widely applied in various applications [10][11][12]. Clustering algorithms have found broad utilization across diverse data analysis fields [13][10]. Clustering methods play a crucial role in the data mining domain and hold significance in various types of data analysis. With the diversity of data types, various variations of clustering techniques have been developed concurrently to process different types of data. Each clustering technique has its strengths and weaknesses in segregating data into groups. The capability and implementation of clustering techniques in the scope of EDM (Exploratory Data Mining) have been detailed in a study [14]. Clustered data challenges can be addressed through a model-based approach, which involves applying specialized models to clusters and striving to optimize the fit between the data and the model. Each cluster (component) can be mathematically represented through parametric distributions, such as Gaussian (continuous) or Poisson (discrete). The entire dataset is then represented by a mixture of these distributions. The single distribution used to characterize a specific cluster is often referred to as a component distribution [15]. In the field of statistics, there is significant interest in clustering mixed and continuous data.

This approach can be categorized into four main categories, as explained by Ahmad and Khan [16]: (i) (i) partitional clustering seeks to minimize the distance between observations and center groups through iterative optimization, as in K-modes or K-prototypes [17]; (ii) (ii) hierarchical algorithms perform hierarchical clustering and merge them to form the final clustering [18][19]; (iii) (iii) model-based clustering [20][21], as the name suggests, relies on probability distributions; (iv) finally, Neural Networks-based algorithms [22] design clusters through connected neurons that learn complex patterns in the data.

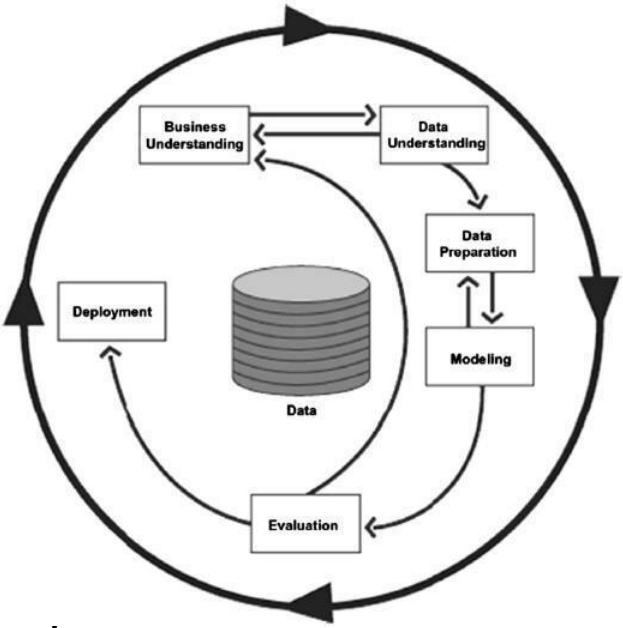
In the context of model-based clustering, we propose a model for clustering mixed data, where various non-continuous variables are combined through the Generalized Linear Latent Variable Model (GLLVM) [23][24]. GLLVM assumes the presence of a linking function between the non-continuous observation space (comprising non-continuous variables) and a latent continuous data space, involving Gaussian latent variables. More recently, Cagnone and Viroli [25] have extended this approach by considering latent variables that no longer follow Gaussian distributions but certain mixtures of Gaussians [21], s resulting in observations naturally forming clusters. In this research, the Gaussian Mixture Model (GMM) will be utilized.

In 1999, the first version of CRISP-DM (CRoss-Industry Standard Process for Data Mining) was introduced, better known as the Cross-Industry Standard Process for Data

Mining [26]. This straightforward methodology was created to outline and guide the general steps in data mining projects. Shortly thereafter, CRISP-DM became the "de facto standard for developing data mining and knowledge discovery projects" [27]. CRISP-DM (CRoss-Industry Standard Process for Data Mining) originated in the 1990s and has been around for approximately two decades. According to various surveys and user feedback, this method is still considered a widely used standard in developing data mining and knowledge discovery projects [28]. In 1979, Davies and Bouldin introduced a scheme used to evaluate a data mining test, which later became known as the Davies Bouldin Index method [29].

2. Method

The research procedure serves as a guide for conducting the study, ensuring that the formulation of the problem and research objectives can be addressed. In this step, the approach employed adopts the Cross Industry Standard for Data Mining (CRISP-DM) methodology. CRISP-DM is a framework that applies a commonly used data development process model, utilized by experts to address various challenges. The research process follows the six stages of CRISP-DM: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. [30]. Here is the intended CRISP-DM method:



1. Result and Discussion

1. Business Understanding

The processing of data concerning Residence Permits in the Class II Non TPI Kediri Immigration Office is essential due to its extensive jurisdiction. Therefore, data analysis is required to identify knowledge patterns that can be used to formulate policies and make decisions related to Residence Permit applications. The clustering process is expected to yield detailed and structured information, which will serve as the foundation for

determining Residence Permit policies and monitoring the presence of foreigners. Additionally, the creation of a vulnerability map for foreigners is also anticipated.

The aim of this step is to enable officials and employees to comprehend the characteristics of Residence Permit applicants easily and swiftly. The initial approach taken involves gathering data regarding the issuance of Residence Permits from the Class II Non TPI Kediri Immigration Office, available within the Immigration Management Information System (SIMKIM). The data utilized in this research encompasses Residence Permit issuances from 2019 to June 2022. The variables employed include sponsorship and destination (address of residence in Indonesia).

2. Data Understanding

a. Data collection

In this stage, the examination of the database and data structure within the Residence Permit Section of the Class II Non TPI Kediri Immigration Office is conducted. The data source is obtained from the Immigration Management Information System (SIMKIM) of the Class II Non TPI Kediri Immigration Office in the format of PDF documents, encompassing a total of 2,096 records.

b. Data Explanation

The Immigration Residence Permit dataset within the Immigration Management Information System (SIMKIM) comprises several variables, including:

Table 1. Residence permit variables

No	Variable	Data Type
1	Residence permit type	Nominal
2	Registration number	Nominal
3	Date of completion	Nominal
4	Valid until	Nominal
5	Name	Nominal
6	Place of birth	Nominal
7	Nationality	Nominal
8	Date of birth	Nominal
9	Gender	Nominal
10	File number	Nominal
11	Passport number	Nominal
12	Passport expiration	Nominal
13	Sponsorship	Nominal
14	Destination	Nominal

c. Data Analysis

The dataset utilized consists of Immigration Residence Permit issuance data from the years 2019 to 2022, obtained in PDF format from which it will be converted into Excel format based on the relevant variables.

d. **Penilaian Kualitas Data**

Evaluation and assessment of data quality are crucial to detect any missing values, commonly referred to as "missing value," within the variables of the Immigration Residence Permit issuance dataset. This step involves scrutinizing to ensure that each record holds complete data and each column is populated with accurate data, free from any inconsistencies. If empty data is identified, the next step is to fill in these values based on an examination of other datasets, as some missing data may be due to system errors. The selection of the sponsor and destination variables is also taken into consideration during this step.

3. Data Preparation

Figure 2. Data sample

	Origin	Sponsor	Gender	Resident Permit	Destination
2050	1	2	1	1	2.0
1776	2	2	1	1	2.0
973	1	1	1	2	1.0
243	1	1	1	2	1.0
1508	1	1	2	2	2.0
642	1	1	1	1	1.0
1739	3	1	2	1	2.0
987	4	2	1	2	1.0
43	1	1	2	2	1.0
1065	2	2	1	2	4.0

Origin is a colum consists of 1 for foreigner from Asia, 2 for foregner from Europe, 3 for foregner from America, 4 for foregner from Australia, 5 for foregner from Africa, and 6 for foregner from Oceania.

Sponsor consists of 1 for corporate sponsorship, 2 for personal sponsorship, and 3 for non-sponsorship. Gender consists of 1 for male, and 2 for female. Resident Permit consists of 1 for restricted residency, 2 for temporary visit, and 3 for permanent residency. Destination is a place where foreigner will stay or live, consists of 1 for Kediri Regency, 2 for Kediri City, 3 for Jombang Regency, 4 for Nganjuk Regency, and 5 for Others. Others mean outside Kediri and its around.

Figure 3. Columns selection for research

```
X = data[['Sponsor', 'Destination']].copy()
X.head(5)
```

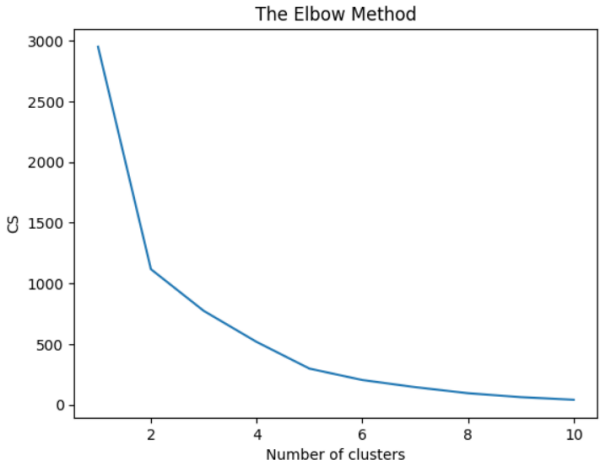
	Sponsor	Destination
0	2	2.0
1	1	4.0
2	1	2.0
3	1	2.0
4	1	2.0

Figure 4. Convert data type into integer

```
X["Destination"] = X["Destination"].astype("int")
X.head(10)
```

	Sponsor	Destination
0	2	2
1	1	4
2	1	2
3	1	2
4	1	2
5	1	2
6	1	2
7	1	1
8	1	3
9	1	1

Figure 5. Elbow method



The elbow method is used to determine an approximately suitable number of clusters. From the plot above, we can observe a change at k=2. Therefore, k=2 can be considered as an appropriate number of clusters to group this data. However, we have observed that we achieved a low classification accuracy of 1% with k=2. We rewrite the necessary code with k=2 again for convenience.

4. Modeling

A Gaussian Mixture Model (GMM) is a statistical model utilized to represent a dataset as a combination of multiple Gaussian distributions. Each Gaussian component in this combination denotes a distinct cluster within the dataset, and the model assumes that the observed data points are generated from these underlying Gaussian distributions with specific probabilities. GMMs are commonly used for tasks involving clustering and density estimation. The primary objective is to identify the inherent clusters within the data and estimate the probability distribution of the data points. The GMM's parameters encompass the mean, covariance, and mixing coefficients of each Gaussian component. Typically, the model is trained using iterative techniques such as the Expectation-Maximization (EM) algorithm, which seeks to determine the optimal parameters that best capture the observed data within the framework of the Gaussian mixture assumption.

Figure 6. Gaussian model result

Cluster 1 is predominantly composed of foreign nationals (WNAs) who corporations sponsor for either residency or visits to the Kediri area and its surroundings. On the other hand, Cluster 2 is characterized by WNAs who are sponsored by individuals or have no sponsors, with their destinations lying outside the Kediri district.

5. Evaluation

Figure 7. Cluster evaluation

```
For n_clusters=2, the silhouette score is 0.5455144939744229
For n_clusters=3, the silhouette score is 0.654899499860855
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans
  warnings.warn(
For n_clusters=4, the silhouette score is 0.6955035313923872
For n_clusters=5, the silhouette score is 0.8240965379856272
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans
  warnings.warn(
For n_clusters=6, the silhouette score is 0.855545655966024
For n_clusters=7, the silhouette score is 0.9198883532384114
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans
  warnings.warn(
For n_clusters=8, the silhouette score is 0.9341621965234231
```

According to Figure 7, the Davies Boulding Index value for Cluster 2 is the smallest, with a value of 0.545, and it is closest to 0 compared to the other clusters.

6. Deployment

Generate a report and publish an article.

4. Conclusion

The choice of 2 clusters is accurate and in line with the DBI, thus providing information that Cluster 1 is predominantly composed of foreign nationals (WNAs) who are sponsored by corporations for either residency or visits to the Kediri area and its surroundings. On the other hand, Cluster 2 is characterized by WNAs who are sponsored by individuals or have no sponsors, with their destinations lying outside the Kediri district.

References:

[1] Lee ES. A Theory of Migration Author (s): Everett S . Lee Published by : Springer on behalf of the Population Association of America Stable URL :

http://www.jstor.org/stable/2060063 Accessed : 05-06-2016 17 : 53 UTC Your use of the JSTOR archive indicates your. Demography 1966;3:47–57.

[2] Lockwood M. Good governance for terrestrial protected areas: A framework, principles and performance outcomes. J Environ Manage 2010;91:754–66. https://doi.org/10.1016/j.jenvman.2009.10.005.

[3] Lia Hananto A, Assiroj P, Priyatna B, Nurhayati, Fauzi A, Yuniar Rahman A, et al. Analysis of Drug Data Mining with Clustering Technique Using K-Means Algorithm. J. Phys. Conf. Ser., vol. 1908, IOP Publishing Ltd; 2021. https://doi.org/10.1088/1742-6596/1908/1/012024.

[4] Sato Y, Izui K, Yamada T, Nishiwaki S. Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization. Expert Syst Appl 2019;119:247–61. https://doi.org/10.1016/j.eswa.2018.10.047.

[5] Gola J, Britz D, Staudt T, Winter M, Schneider AS, Ludovici M, et al. Advanced microstructure classification by data mining methods. Comput Mater Sci 2018;148:324–35. https://doi.org/10.1016/j.commatsci.2018.03.004.

[6] Setiawidayat S, Yuniar Rahman A. New method for obtaining Peak Value R and the duration of each cycle of Electrocardiogram. IEEE; 2018.

[7] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics Med Unlocked 2018;10:100–7. https://doi.org/10.1016/j.imu.2017.12.006.

[8] Larose DT, Larose CD. Discovering knowledge in data : an introduction to data mining. n.d.

[9] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. 2016. https://doi.org/10.1016/c2009-0-19715-5.

[10] Wang Y, Lin X, Wu L, Zhang W, Zhang Q. Exploiting correlation consensus: Towards subspace clustering for multi-modal data. MM 2014 - Proc. 2014 ACM Conf. Multimed., Association for Computing Machinery; 2014, p. 981–4. https://doi.org/10.1145/2647868.2654999.

[11] Liu H, Shao M, Li S, Fu Y. Infinite ensemble for image clustering. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17- August-2016, Association for Computing Machinery; 2016, p. 1745–54. https://doi.org/10.1145/2939672.2939813.

[12] Gao Y, Miao X, Chen G, Zheng B, Cai D, Cui H. On efficiently finding reverse k-nearest neighbors over uncertain graphs. VLDB J 2017;26:467–92. https://doi.org/10.1007/s00778-017-0460-y.

[13] Liu X, Wan C, Chen L. Returning clustered results for keyword search on XML documents. IEEE Trans Knowl Data Eng 2011;23:1811–25. https://doi.org/10.1109/TKDE.2011.183.

[14] Dutt A, Ismail MA, Herawan T. A Systematic Review on Educational Data Mining. IEEE Access 2017;5:15991–6005. https://doi.org/10.1109/ACCESS.2017.2654247.

[15] Sumitra S. Mixture Models and EM Algorithm. n.d.

[16] Ahmad A, Khan SS. Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access 2019;7:31883–902. https://doi.org/10.1109/ACCESS.2019.2903568.

[17] Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min Knowl Discov 2 1998;1:284–90. https://doi.org/10.3923/ajbmb.2011.284.290.

[18] Philip G, Ottaway BS. Mixed Data Cluster Analysis: an Illustration Using Cypriot Hooked-Tang Weapons. Archaeometry 1983;25:119–33. https://doi.org/10.1111/j.1475-4754.1983.tb00671.x.

[19] Chiu T, Fang DP, Chen J, Wang Y, Jeris C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. Proc Seventh ACM SIGKDD Int Conf Knowl Discov Data Min 2001:263–8. https://doi.org/10.1145/502512.502549.

[20] Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2008. https://doi.org/10.1007/s10115-007-0114-2.

[21] Taylor P, Fraley C, Raftery AE. Journal of the American Statistical Association and Density Estimation Model-Based Clustering , Discriminant Analysis , and Density Estimation 2012:37–41.

[22] Kohonen T. The Self-Organizing Map. Proc IEEE 1990;78:1464–80. https://doi.org/10.1109/5.58325.

[23] Moustaki I. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. Br J Math Stat Psychol 2003;56:337–57. https://doi.org/10.1348/000711003770480075.

[24] Moustaki I. Generalized latent trait models 2000;65:391–411.

[25] Cagnone S, Viroli C. A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. AStA Adv Stat Anal 2014;98:1–20. https://doi.org/10.1007/s10182-012-0206-5.

[26] Chapman P. The CRISP-DM User Guide. Cris User Guid 1999:14.

[27] Marbán O, Segovia J, Menasalvas E, Fernández-Baizán C. Toward data mining engineering: A software engineering approach. Inf Syst 2009;34:87–107. https://doi.org/10.1016/j.is.2008.04.003.

[28] Martinez-Plumed F, Contreras-Ochando L, Ferri C, Hernandez-Orallo J, Kull M, Lachiche N, et al. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans Knowl Data Eng 2021;33:3048–61. https://doi.org/10.1109/TKDE.2019.2962680.

[29] Muningsih E, Maryani I, Handayani VR. Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa. EVOLUSI J Sains Dan Manaj 2021;9.

[30] Mariscal G, Marbán Ó, Fernández C. A survey of data mining and knowledge discovery process models and methodologies. Knowl Eng Rev 2010;25:137–66. https://doi.org/10.1017/S0269888910000032.

[31] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0 Step-by-step data mining guide. DaimlerChrysler; 1999.

[32] Kamayanti A. Paradigma Nusantara Methodology Variety : Re-embedding Nusantara Values into Research Tools. Int J Relig Cult Stud 2021;3:123–32.

[33] Setiawan AR. Paradigma Nusantara for the Advancement of Accounting and Other Social Sciences. Int J Relig Cult Stud 2022;4:93–104. https://doi.org/10.34199/ijracs.2022.04.09.

[34] Persson. “Government Value Paradigms-Bureaucracy, New Public Management, and E-Government.” Commun Assoc Inf Syst 2010;27.