Clustering and Association Rule Mining for Retail Prices of Staple Goods and Commodities in Bukittinggi and Padang

Muhammad Fauzan Alkhairi[1], Syafrijon[1], Dony Novalendry[1], Yeka Hendriyani[1]

[1] *Department Electronic Enginering, Universitas Negeri Padang, Padang, Padang, Indonesia*

Abstract **- Price changes are critical indicators of economic dynamics and government policies. This study uses clustering and association rule mining techniques to analyze retail price data of staple goods and commodities in Bukittinggi and Padang from 2017 to 2023. The K-Means algorithm identified significant price patterns, validated by high silhouette scores, revealing distinct clusters that illustrate regional price trends. The Apriori algorithm uncovered strong correlations between commodity prices, such as the relationships between salt and cement and between salt and flour, indicated by high confidence and lift values. These findings provide valuable insights into price behaviours and can inform decision-making in price management and policy formulation. The results demonstrate the potential of data mining techniques to enhance understanding and optimization of regional economic stability, offering a robust approach to analysing complex price data. This information can help local governments make better decisions and manage prices more effectively. Additionally, this study underscores the importance of in-depth data analysis to understand price patterns and develop more effective price management strategies in the future. For further research, it is recommended to explore long-term price trends and external factors affecting prices, as well as to apply other machine learning methods and expand the research to other regions for comparative price dynamics.**

Keywords—**Clustering, K-Means, Association Rule Mining, Apriori, Price Dynamics Bukittinggi-Padang**

I. INTRODUCTION

Price changes reflect economic dynamics and government policies [13]. Price affordability and stable supply are crucial for consumers [15]. For example, a 30-year-old consumer stated that affordability of goods has been their top priority for the past five years. They also expressed great gratitude for being able to afford them [16]. When prices rise, it impacts the amount of loans and leads to higher interest costs.9 Therefore, regulations and policies in a region are essential to maintain the economy, as the economy and technological changes are closely related to the political context [22].

One source of regional data providing statistical information to the public is Statistics Indonesia [17]. The Statistics of Sumatera Barat Province reported the average retail prices of staple goods, cement, and gold in Bukittinggi and Padang from 2017 to 2023 [7][8]. This data is used for clustering based on price movements [2] and identifying association rules between prices [21]. Therefore, clustering and association rule mining are highly necessary [3].

To perform clustering based on price movements, the K-Means algorithm can be used [2]. This K-Means algorithm shows 98.7% classification accuracy in segmentation [23]. With varying cluster numbers, the silhouette score is used to evaluate the optimal cluster [18]. Meanwhile, to identify association rules between prices, the Apriori algorithm can be used [21]. The relevance and strength of the generated association rules are determined using specific support and confidence levels [4].

This research aims to apply clustering and association rule mining techniques to the retail price data of staple goods and commodities in Bukittinggi and Padang. Thus, it is expected that the results of this study can reveal significant patterns and relationships in the retail price data, which can be used to optimise price management and improve community welfare.

II. METHOD

This research uses retail price data of staple goods and commodities from Bukittinggi and Padang reported by BPS Sumatera Barat Province from 2017 to 2023. The data includes the average prices of staple goods such as rice (Kg), chicken meat (Kg), cooking oil (Kg), granulated sugar (Kg), salt (Kg), chicken eggs (10 pieces), infant milk (400 grams package), red chilli (Kg), wheat flour (Kg), and other commodities like cement (50 sacks), and 24-carat gold (gram) [7][8].

Main tools were used for data analysis: Orange 3.36.2 was used to run the K-Means Clustering and Apriori Association algorithms and data visualisation.

DOI: 10.18782/ijeitel-xx
https://doi.org/10.18782/ijeitelxx-xx

Corresponding author: Muhammad Fauzan Alkhairi, *Universitas Negeri Padang*
Email: fauzanalkhairi27@gmail.com,

Received: August 2024
Revised: August 2024
Accepted: October 2024
Published: October 2024

A. Process Mining

Process mining is a method for objectively observing and understanding processes by analysing event data in detail. This method automatically creates models that show how events occur, thus enhancing process transparency and understanding [1]. The Data Preparation stage is crucial and includes several key steps: collecting data from various relevant sources, merging data to form a comprehensive dataset, filling in missing values and addressing inconsistencies, reducing data volume through feature selection, and transforming data into a suitable format for analysis [14]. Process mining allows for dynamic analysis of process models from start to finish, providing information to improve and optimise workflows [1].

B. K-Means Clustering

Clustering in Data Mining is an unsupervised machine learning method that groups objects into clusters based on similarity, so that objects within a cluster are more similar to each other than to those in other clusters [11]. The K-means method is well-known for its ease of implementation, computational speed, and ability to handle large datasets, requiring only one parameter, the number of clusters (K). The algorithm starts with the formation of initial clusters and iterates until there are no significant changes, indicating convergence [18].

$$\underline{x_i} = \frac{1}{m}\sum_{j=1}^{m} \quad = x_{ij} \tag{1}$$

Where:

- x_i is the coordinate of the i-th centroid,
- m is the number of objects in the cluster,
- x_{ij} is the value of the i-th attribute of the j-th object in the cluster.

This process is repeated until the centroids' positions are stable and there are no significant changes, indicating the algorithm's convergence [20]. To find the silhouette coefficient, use equation (2). The silhouette coefficient is a metric for assessing the quality of clustering, determining whether the algorithm has grouped the data well. Higher silhouette coefficients indicate better clusters [12].

$$s(i) = \frac{b(i)-a(i)}{max\{a(i)-b(i)\}} \tag{2}$$

Where:

- $a(i)$ is the coordinate of the i-th centroid,
- $b(i)$ is the average distance of object i to all clusters other than the one containing object i [18].

Calculate the silhouette value comparatively for clusters 2, 3, and so on by repeating the above steps. The cluster with the highest silhouette value is the best according to the evaluation method. Calculate the average silhouette value of all objects to evaluate the entire cluster [18].

C. Apriori Association

Association Rule Mining (ARM) is an important data mining technique for extracting hidden knowledge from databases. As part of Machine Learning (ML), ARM identifies interesting relationships between objects in the data [5]. The initial stage of the Apriori Algorithm forms a set of popular items by setting a minimum support threshold, which affects the number of rules generated. The second stage is the formation of association rules from the identified popular itemsets [6]. Association rules are usually represented in an IF-THEN format, making it easier for researchers to interpret and understand the results. The mined association rules are filtered using three main statistical indicators: support, confidence, and lift ratio [4].

Support measures how often an item or combination of items appears in the dataset [19]. It is expressed in the form of $A \rightarrow B$ as in equation (3). A is known as the antecedent or LHS, and B is known as the consequent or RHS [10].

$$Support(A \rightarrow B) = P(A \cap \quad B) = \frac{\#(A \cap B}{N} \tag{3}$$

Confidence assesses how often item B appears in transactions that also contain item A [19], as reflected in equation (4). The higher the confidence value, the more likely the consequent and antecedent appear together [10].

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support\ (A)} = \frac{P(A \cap B)}{P(A)} = \frac{\#(A \cap B)}{\#A} \tag{4}$$

Lift is the ratio between the confidence of a rule and the expected confidence, depicted in equation (5). Lift indicates the strength of the relationship between the antecedent and the consequent [10].

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{P(B)} = \frac{P(A \cap B)}{P\ (A) \times P(B)} \tag{5}$$

A lift value greater than 1 indicates a strong relationship between A and B, signifying that the presence of A increases the likelihood of B's presence. A lift value less than 1 indicates that A and B are less likely to occur together, thus having no strong relationship or having a negative relationship [19].

III. RESULTS

A. Data Integration

The data we obtained comes from the official reports of average retail prices by the Statistics of Sumatera Barat Province. This data was collected and integrated into a single .csv file to facilitate further processing and analysis. The data integration process involved merging several data sets from various time periods and different types of commodities. This step is crucial to ensure that the data used accurately reflects the retail price conditions in the Bukittinggi and Padang areas.

B. Testing

The testing phase is a crucial step in our analysis to ensure the accuracy and reliability of the results obtained from both clustering and association rule mining techniques. The data we used comes from the Statistics of Sumatera Barat Province, which includes monthly average retail prices from 2017 to 2023 for staple goods, cement, and gold in Bukittinggi and Padang. This data was meticulously integrated into a single .csv file, facilitating seamless processing and analysis.

For the clustering analysis, we first processed the data by editing the domain type of the "Month" column to Time and setting the "City" column as Metas. This setup allowed us to focus on the 11 features representing the average retail prices of various goods, ensuring the analysis was aligned with the research objectives. In the main steps as listed in figure 1.

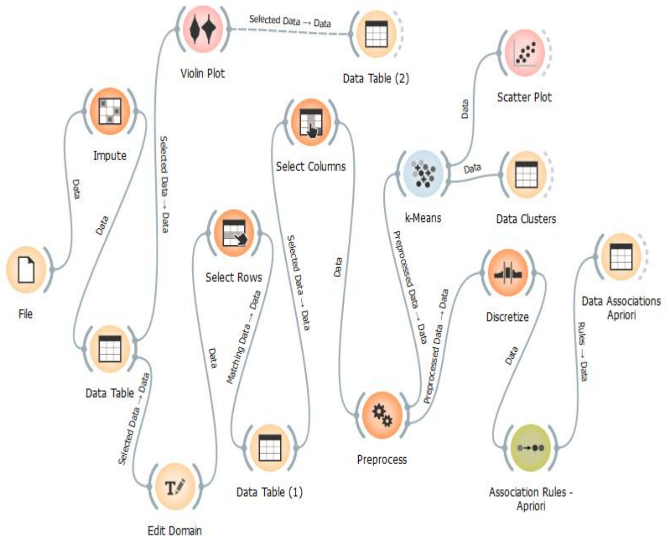

Fig. 1. Workflow in Orange 3.36.2 for data analysis

The K-Means clustering was then performed with a fixed number of clusters set to 3. This decision was guided by the need to identify distinct price patterns and group them accordingly. To evaluate the quality of these clusters, we employed the silhouette score, a robust metric for assessing the coherence and separation of the clusters. The silhouette score ranges from -1 to 1, with higher values indicating well-defined clusters. By calculating the silhouette score for different numbers of clusters, we determined that 3 clusters provided a meaningful and interpretable segmentation of the data.

For the association rule mining, we discretized the features using the equal frequency method with k=4, as shown in Figure 2. This step was essential to transform the continuous price data into categorical intervals, making it suitable for the Apriori algorithm. The discretization process involved dividing the data into four equal-sized intervals, ensuring that each interval contained an approximately equal number of data points. This approach facilitated the identification of meaningful associations between different price levels.

Beras (Kg) (equal freq, k=4): -0.384255, 0.054358, 0.605436
Daging Ayam Ras (Kg) (equal freq, k=4): -0.2969317, 0.1126015, 0.5031950
Minyak Goreng (Kg) (equal freq, k=4): -0.6984114, -0.4833236, -0.3071613
Gula Pasir (Kg) (equal freq, k=4): -0.391304, -0.120730, 0.032609
Garam (Kg) (equal freq, k=4): -0.850638, 0.771277, 0.976915
Telur ayam ras (10 butir) (equal freq, k=4): -0.268260, 0.025240, 0.307795
Susu untuk Balita (kemasan 400 gram) (equal freq, k=4): 0.0320308, 0.3326132, 0.5682886
Cabe Merah (Kg) (equal freq, k=4): -0.5046727, -0.0860839, 0.1589780
Tepung Terigu (Kg) (equal freq, k=4): -0.598149, -0.463225, 0.010229
Emas 24 karat (gram) (equal freq, k=4): -0.709472, -0.408525, 0.550284
Semen(50sak) (equal freq, k=4): -0.809335, -0.525437, -0.301613

Fig. 2. Result of Discretized Data

Next, we applied the Apriori algorithm to generate association rules. We set the minimum support threshold to 20% and the minimum confidence threshold to 50%. The support threshold ensured that only frequently occurring itemsets were considered, while the confidence threshold focused on the strength of the association between items. The generated rules were then filtered based on these criteria to retain only the most significant and relevant associations.

C. Evaluation of Clusters

When evaluating the quality of the best cluster, the average Silhouette score is assessed across all clusters [18]. The scatter plot visualisation in Figure 3 shows the clustering quality across three clusters. The low Silhouette Score in C1 indicates poor cluster quality in both Bukittinggi and Padang. C2 shows better quality compared to C1, with a higher score. The best quality is found in C3, which has a high Silhouette Score in both cities, indicating that the data within this cluster is very similar to each other and well-separated from other clusters.

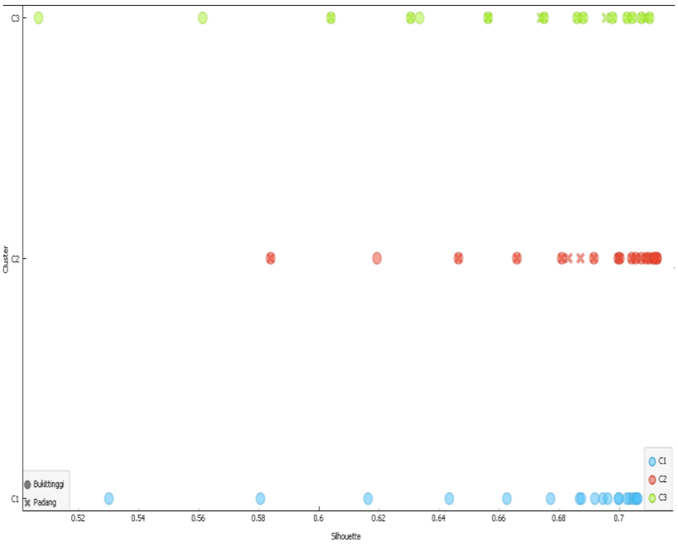

Fig. 3 Silhouette Score of Retail Price Clusters

Based on the clustering results shown in Table 1, there are several notable differences in the average retail prices between Bukittinggi and Padang. In cluster C1, Bukittinggi has more price items in 2017 (11 count) and 2018 (8 count) compared to Padang, which is not included in this cluster. In cluster C2, Bukittinggi shows a significant increase in the number of price items in 2021 (11 count), while Padang experiences fluctuations with the highest number of price items in 2020 (11 count) and a sharp decrease in 2022 (2 count). In cluster C3, Bukittinggi shows an increase in the number of price items in 2019 (8 count) and a decrease in 2021 (2 count). Meanwhile, Padang shows a significant increase in the number of price items in 2019 (12 count) and a decrease in 2022 (2 count). These differences highlight the varying retail price patterns between the two cities, reflecting market dynamics and economic factors that may influence commodity prices in each region. Further analysis is required to understand the causes of these fluctuations and how local policies may impact retail prices in Bukittinggi and Padang.

TABLE I
CLUSTERS OF RETAIL PRICE IN BUKITTINGGI AND PADANG

Clusters	City	2017	2018	2019	2020	2021	2022
C1	Bukittinggi	11	8				
C2	Bukittinggi				6	11	
	Padang				5	11	2
C3	Bukittinggi				2		
	Padang				2		

From the results of this analysis, it can be concluded that there is a significant difference in the number of items in the average retail price of basic commodities, cement and gold between Bukittinggi and Padang. Clustering produces three significant clusters with fairly high validity, based on Silhouette Score analysis. These results provide insight into price trends and variations between cities, which can be used for decision making in pricing and distribution policies of basic commodities.

D. Evaluation of Associations

By applying a minimum support of 20% and a minimum confidence of 50% in discovering association rules, we can gain an understanding of how often item B appears in conjunction with item A. This can be seen in Fig. 5, which shows that a lift value greater than 1 indicates a strong relationship between A and B, suggesting that the presence of A increases the likelihood of B's presence (Samuel et al., 2022). The details of these association rules can be found in Table 2.

TABLE II
PRODUCTS PRICE ASSOCIATION RULES

Antecedent	Consequent	Support	Confidence	Lift
Garam (Kg) =0.771277-0.976915	Semen(50sak) =<-0809335	0.241	0.87	3.437
Garam (Kg) =0.850638-0.771277	Semen(50sak) =-0809335--0525437	0.241	0.952	3.04
Garam (Kg) =<-0.850638	Tepung Terigu (Kg) =>0.010229	0.217	0.947	3.744
Semen (50sak) =0.809335--0.525437	Garam (Kg) =-0850638-0.771277	0.241	0.769	3.04
Semen (50sak) =<-0.809335	Garam (Kg) =0.771277-0.976915	0.241	0.952	3.437
Tepung Terigu (Kg) =>0.010229	Garam (Kg) =<-0.850638	0.217	0.857	3.744

From the analysis in table 2, it was found that if the price of salt falls within the range of 0.771277 to 0.976915, then it is highly likely that the price of cement per 50 sacks will be less than or equal to -0.809335. This is supported by a confidence of 87%, a lift of 3.437, and a support of 24.1%, meaning that approximately 24.1% of all transactions meet this rule. This rule indicates a strong relationship between the given salt price range and a decrease in the price of cement.

Furthermore, if the price of salt is within the range of -0.850638 to 0.771277, it is very likely that the price of cement per 50 sacks will be within the range of -0.809335 to -0.525437. With a confidence of 95.2%, a lift of 3.040, and a support of 24.1%, this rule applies to about 24.1% of all transactions. This shows a very strong correlation between this range of salt prices and lower cement prices.

Additionally, if the price of salt is less than or equal to -0.850638, it is very likely that the price of flour will be greater than or equal to 0.010229. A confidence of 94.7%, a lift of 3.744, and a support of 21.7% indicate that this rule applies to approximately 21.7% of all transactions. This rule suggests that a lower salt price is correlated with a higher flour price.

If the price of cement per 50 sacks is within the range of -0.809335 to -0.525437, it is likely that the price of salt will be within the range of -0.850638 to 0.771277. With a confidence of 76.9%, a lift of 3.040, and a support of 24.1%, this rule applies to about 24.1% of all transactions. This indicates a significant relationship between the prices of cement and salt within this range.

Moreover, if the price of cement per 50 sacks is less than or equal to -0.809335, it is highly likely that the price of salt will be within the range of 0.771277 to 0.976915. With a confidence of 95.2%, a lift of 3.437, and a support of 24.1%, this rule applies to about 24.1% of all transactions. This demonstrates a very strong correlation between lower cement prices and the given range of salt prices.

Finally, if the price of flour is greater than or equal to 0.010229, it is likely that the price of salt will be less than or equal to -0.850638. With a confidence of 85.7%, a lift of 3.744, and a support of 21.7%, this rule applies to about 21.7% of all transactions. This shows that higher flour prices are correlated with lower salt prices. Through the application of minimum support and confidence in association analysis, we can identify significant patterns between the prices of staple goods, cement, and flour in Bukittinggi and Padang.

IV. CONCLUSIONS

This research demonstrates the efficacy of applying clustering and association rule mining techniques to retail price data of staple goods and commodities in Bukittinggi and Padang. The use of K-Means clustering revealed significant patterns in the price movements, allowing us to identify distinct clusters that provide insights into regional price trends. The silhouette score analysis confirmed the validity of these clusters, highlighting the differences in retail price behaviours between the two cities.

Moreover, the application of the Apriori algorithm for association rule mining uncovered meaningful relationships between the prices of various commodities. High confidence and lift values in the generated rules indicate strong correlations, such as the relationship between salt and cement prices, as well as between salt and flour prices. These findings can guide decision-making processes in price management and policy formulation.

Overall, the study's results contribute to a better understanding of regional price dynamics and their underlying factors. By leveraging statistical techniques and robust data analysis, this research provides valuable insights for optimizing pricing strategies and improving economic stability in Bukittinggi and Padang. Future research can build upon these findings by exploring additional variables and employing more advanced machine learning models to further enhance the analysis of retail price data.

ACKNOWLEDGMENT

We express our sincere gratitude to the Informatics Study Program, Faculty of Engineering, Universitas Negeri Padang, for their support and facilities, which were essential for this research. We also thank Badan Pusat Statistik (BPS) Sumatera Barat for providing the crucial public data. Special thanks to Mr. Agariadne Dwinggo Samala for helping one of the authors create an ORCID account.

REFERENCES

[1] Y. Pan and L. Zhang, “A BIM-data mining integrated digital twin framework for advanced project management,” Automation in Construction, vol. 124, p. 103564, Apr. 2021, doi: 10.1016/j.autcon.2021.103564.

[2] D. Das, P. Kayal, and M. Maiti, “A K-means clustering model for analyzing the Bitcoin extreme value returns,” Decision Analytics Journal, vol. 6, p. 100152, Mar. 2023, doi: 10.1016/j.dajour.2022.100152.

[3] L. Rosliadewi, “Analysis of Transaction Data for Modeling the Pattern of Goods Purchase Supporting Goods Location,” Journal of Applied Data Sciences, vol. 1, no. 2, pp. 65–75, Dec. 2020, doi: 10.47738/jads.v1i2.54.

[4] M. H. Santoso, “Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom,” Brilliance: Research of Artificial Intelligence, vol. 1, no. 2, pp. 54–66, Dec. 2021, doi: 10.47709/brilliance.v1i2.1228.

[5] N. Wahidi and R. Ismailova, “Association rule mining algorithm implementation for e-commerce in the retail sector,” Journal of Applied Research in Technology & Engineering, vol. 5, no. 2, pp. 63–68, Apr. 2024, doi: 10.4995/jarte.2024.20753.

[6] A. Roskladka, N. Roskladka, V. Kulazhenko, A. Desiatko, and A. Shtanova, “Association Rule System for Effective Risk Management of a Cinema Chain,” in Cybersecurity Providing in Information and Telecommunication Systems 2024, 2024, pp. 216–227.

[7] Statistics of Sumatera Barat Province, “Average Retail Price of Essential Commodities, Cement, and Gold in Bukittinggi. Statistics of Sumatera Barat Province,” Badan Pusat Statistik Provinsi Sumatera Barat. https://sumbar.bps.go.id/indicator/102/323/1/harga-eceran-rata-rata-bahan-pokok-semen-dan-emas-di-kota-bukittinggi.html

[8] Statistics of Sumatera Barat Province, “Average Retail Price of Essential Commodities, Cement, and Gold in Padang,” Badan Pusat Statistik Provinsi Sumatera Barat. https://sumbar.bps.go.id/indicator/102/322/1/harga-eceran-rata-rata-bahan-pokok-semen-dan-emas-di-padang.html

[9] Brian Uhler, “California Housing Affordability Tracker (2nd Quarter 2024) [EconTax Blog].” https://lao.ca.gov/LAOEconTax/Article/Detail/793 (accessed Aug. 07, 2024).

[10] A. B. Z. Chai, B. T. Lau, M. K. T. Tee, and C. McCarthy, “Comparative analysis of regional variations in road traffic accident patterns with association rule mining,” EAI Endorsed Transactions on Pervasive Health and Technology, vol. 9, Nov. 2023, doi: 10.4108/eetpht.9.3173.

[11] R. Rahim, U. H. Salsabila, A. A. Dani, E. M. S.S. Ciptaningsih, and M. M, “Data Mapping using Combining Clustering Methods and C.45 Classification,” International Journal of Electronics and Communication Engineering, vol. 10, no. 5, pp. 96–104, May 2023, doi: 10.14445/23488549/ijece-v10i5p109.

[12] V. H. Antonius and D. Fitrianah, “Enhancing Customer Segmentation Insights by using RFM + Discount Proportion Model with Clustering Algorithms,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, 2024, doi: 10.14569/ijacsa.2024.0150390.

[13] Sahabat Pegadaian, “Indeks Harga: Pengertian, Jenis, Tujuan, dan Cara Hitungnya,” PT Pegadaian, Jan. 31, 2024. [Online]. Available: https://sahabat.pegadaian.co.id/artikel/keuangan/indeks-harga-adalah

[14] A. Dogan and D. Birant, “Machine learning and data mining in manufacturing,” Expert Systems with Applications, vol. 166, p. 114060, Mar. 2021, doi: 10.1016/j.eswa.2020.114060.

[15] Greenpeace Indonesia, “Masyarakat Sipil Menggugat Menteri Perdagangan dan Jokowi atas Kegagalan Negara Menjamin Pasokan dan Stabilisasi Harga Minyak Goreng ,” Greenpeace Indonesia, Jun. 04, 2022. https://www.greenpeace.org/indonesia/siaran-pers/46189/masyarakat-sipil-menggugat-menteri-perdagangan-dan-jokowi-atas-kegagalan-negara-menjamin-pasokan-dan-stabilisasi-harga-minyak-goreng/ (accessed Aug. 07, 2024).

[16] N. Sherman, “Trying to buy a house is ‘playing a game you can’t win,’” BBC News, May 26, 2024. https://www.bbc.com/news/articles/cmj66r4lvzzo

[17] Pejabat Pengelola Informasi dan Dokumentasi Badan Pusat Statistik, “Profil BPS,” Portal PPID BPS RI. https://ppid.bps.go.id/app/konten/0000/Profil-BPS.html

[18] P. Anitha and M. M. Patil, “RFM model for customer purchase behavior using K-Means algorithm,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.

[19] S. Samuel, A. Sani, A. Budiyantara, M. Ivone, and F. Frieyadie, “SALES LEVEL ANALYSIS USING THE ASSOCIATION METHOD WITH THE APRIORI ALGORITHM,” Jurnal Riset Informatika, vol. 4, no. 4, pp. 331–340, Sep. 2022, doi: 10.34288/jri.v4i4.422.

[20] S. A. Evdokimova, “Segmentation of store customers to increase sales using ABC-XYZ-analysis and clustering methods,” Journal of Physics: Conference Series, vol. 2032, no. 1, p. 012117, Oct. 2021, doi: 10.1088/1742-6596/2032/1/012117.

[21] J. B. Hernández C., A. García-Medina, and M. A. Porro V., “Study of the Behavior of Cryptocurrencies in Turbulent Times Using Association Rules,” Mathematics, vol. 9, no. 14, p. 1620, Jul. 2021, doi: 10.3390/math9141620.

[22] P. Żuk and P. Żuk, “The role of trust, information and legal stability in the development of renewable energy: the analysis of non-economic factors affecting entrepreneurs’ investments in green energy in Poland,” Environment, Development and Sustainability, vol. 26, no. 7, pp. 18499–18534, May 2023, doi: 10.1007/s10668-023-03400-z.

[23] L. Torkzadeh et al., “Value-Based Market Segmentation of West Asian Medical Tourism Market: A Case of Iran,” Iranian Journal of Public Health, May 2024, doi: 10.18502/ijph.v53i4.15570.