Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Improving the Voter List Clustering Model Fixed (DPT) using the K-Means Algorithm in Girinata Village

Rizki Aldi[1]*, Nana Suarna[2], Irfan Ali[3], Dendy Indriya Efendi[4]

[1,2,3,4]STMIK IKMI CIREBON

rizkialdi972@gmail.com [1], nana.ikmi@gmail.com [2], irfanali0.0@gmail.com [3], dendy.ikmi@gmail.com[4]*

Abstract

Elections are one of the pillars of democracy that require accurate voter data to ensure transparency and fairness. The Permanent Voter List (DPT) is a crucial element in supporting the smooth running of elections, but there are often data validity problems such as duplicate data, voter location errors, or voter data that does not meet the requirements. This research focuses on the application of the K-Means algorithm to increase the accuracy and validity of the DPT at TPS 05, Girinata Village. The problem formulation in this research includes the accuracy level of the DPT, the effectiveness of the K-Means algorithm in identifying inaccuracies, as well as factors that influence the accuracy of voter data. This research aims to analyze the accuracy level of the DPT, evaluate the effectiveness of the K-Means algorithm in grouping data, and identify factors contributing to the validity of the DPT. The analysis results show that the K-Means algorithm succeeded in grouping voter data with good quality, with a Davies-Bouldin Index (DBI) value of 0.389, which indicates clearly defined clusters. The main factors that influence clustering are age, distance to TPS, and location (RT and TPS). This research shows that the K-Means algorithm can be used to detect inaccuracies in voter data, such as data that does not match the TPS location or age that does not meet the requirements as a voter. With these results, the K-Means algorithm makes a significant contribution to validating voter data, thereby supporting a more transparent and accountable election process.

Keywords: *Permanent Voter List, K-Means, Clustering, Davies-Bouldin Index, Data Validation*

1. Introduction

General elections (Pemilu) are one of the main pillars of democracy which function to ensure people's representation in government. In this context, the Permanent Voter List (DPT) is a crucial element that determines the validity of the votes cast by citizens. However, developments in technology and data mining present opportunities to improve DPT accuracy, especially with the K-Means algorithm which is effective in grouping data with similar characteristics. As the complexity of voter data increases, various computational approaches have been tested to overcome problems in managing and validating voter data. Nevertheless, challenges related to data accuracy and the suitability of the algorithms used in grouping voter data remain important issues that need further research[1] [2]. Considering the importance of data accuracy in DPT, this research aims to explore the effectiveness of the K-Means algorithm in improving the accuracy and reliability of DPT at Girinata Village TPS, a representative area for this kind of study.

Even though various efforts have been made to improve the accuracy of the DPT, there are still a number of problems that have not been resolved, such as duplication of voter data, errors in recording, and a lack of comprehensive validation. Weaknesses in this data collection system can lead to multiple voters or even unregistered voters, which in the end can affect the overall election results. In the current literature, it is proven that the use of data mining algorithms, especially K-Means, is able to overcome most of these problems, even though it is not completely optimal [3][4]). Therefore, this research focuses on identifying the main gaps and challenges in implementing the K-Means algorithm in DPT, in order to provide solutions that are more accurate and can be implemented in the context of elections in Indonesia. Various previous studies have been carried out to improve the accuracy of DPT by using information technology approaches and data mining algorithms. For example, [5]) developed the Nakula Sadewa algorithm to overcome voter duplication, but still faces limitations in terms of scale and data complexity. On the other hand, [1]studied the application of the K-Means algorithm in gubernatorial elections, showing promising results in clustering voter data, but still lacking in terms of wider data validation. Other studies such as those conducted by [2] have also explored the application of mobile technology for recapitulating election results, but have not focused on increasing the accuracy of the DPT.

Research according to (Mahendra, 2018)[6] and (Weriza et al., 2019)[7] shows the importance of voter data collection policies and the role of the work culture of data updating officers, even though the technical aspects of data processing have not been the main concern. Although

these studies have made significant contributions, there is a gap in research focused on applying the K-Means algorithm to improve DPT accuracy, especially in smaller TPS and with high data complexity. This research aims to fill this gap by offering a more comprehensive approach and more rigorous validation, which can be widely applied in the context of elections in Indonesia. This approach will not only optimize the use of the K-Means algorithm, but will also consider various variables that influence the accuracy of DPT, such as demographic changes, population migration, and other social factors. This research aims to analyze the effectiveness of the K-Means algorithm in increasing accuracy DPT at Girinata Village TPS. Through this research, it is hoped that a more accurate and efficient voter data grouping model will be obtained, which can be implemented widely in the election system in Indonesia. The significance of this research lies in its contribution in filling knowledge gaps related to voter data management, as well as providing practical solutions that can increase public confidence in election results. The approach used in this research involves applying the K-Means algorithm to analyze and group voter data at TPS 05, Girinata Village. Data will be processed using data mining techniques by utilizing relevant software. This approach will be tested through cross-validation to ensure the accuracy and reliability of the results obtained. In addition, this research will also consider external factors that can influence grouping results, such as demographic changes and population migration.

1.1. K-Means

K-Means Clustering is a data analysis method or Data Mining method that carries out an unsupervised learning modeling process and uses a method that groups data from various partitions.
K-Means is one of the most popular clustering algorithms in unsupervised learning. This algorithm is used to divide a set of data into a number of groups (clusters) based on certain similarities.

K-Means is a partition-based clustering algorithm that aims to divide datasets into a number of groups (clusters) based on feature similarities. In this algorithm, each data will be put into one cluster so that data in the same cluster is more similar to each other compared to data in different clusters.

1.2. Knowledge Discovery Database

Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, potentially useful, and understandable patterns from large datasets. It involves multiple steps, starting with data selection, preprocessing, and transformation, followed by data mining to extract patterns, and finally, interpreting and evaluating the results. KDD serves as the foundation for extracting actionable insights and knowledge from raw data, often leveraging advanced techniques such as machine learning, statistics, and data visualization to uncover hidden relationships and trends. [8]

2. Research Method

2.1. Research Method

The methodology employed in this research is Knowledge Discovery in Databases (KDD). The workflow or sequence of steps utilized throughout the study is depicted in Figure 1 below, outlining the systematic approach taken for data analysis and interpretation. The data analysis technique in this research was carried out using a data mining approach, which aims to explore, find patterns and analyze the data in depth. The data mining process is carried out by following the Knowledge Discovery in Databases (KDD) stages, which include systematic steps to process raw data into meaningful information.

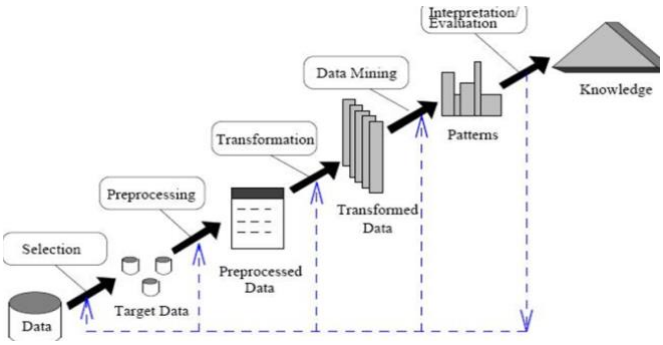

Fig. 1: Research Method

3. Result and Discussion

3.1. Selection

This stage aims to select the data that will be used in the classification process. The dataset used in this study consists of primary data collected through surveys and interviews with the population. In my research, I used raw data from the Girinata Village government in 2024 which has 12 attributes and 506 records. This data was stored in Microsoft Excel (XLSX) file format.

Table 1: Collected Data

no	nama	jenis kelamin	usia	rt	tps	blok	kelurahan	kecamatan	kabupaten	provinsi
1	aan darwati	p	38	003	1	2	girinata	dukupuntang	cirebon	jawa barat
2	abdulloh mauludi mujtaba	l	18	003	1	2	girinata	dukupuntang	cirebon	jawa barat
3	abet	l	71	003	1	2	girinata	dukupuntang	cirebon	jawa barat
4	achmad fadila	l	33	003	1	2	girinata	dukupuntang	cirebon	jawa barat
5	achmad permana hadi	l	61	003	1	2	girinata	dukupuntang	cirebon	jawa barat
6	achmad rohim	l	29	003	1	2	girinata	dukupuntang	cirebon	jawa barat
7	ade suharto	l	51	003	1	2	girinata	dukupuntang	cirebon	jawa barat
8	aef saefudin	l	41	003	1	2	girinata	dukupuntang	cirebon	jawa barat
9	agna yasa nurhabillah	p	17	002	2	1	girinata	dukupuntang	cirebon	jawa barat
10	agus adi saputra	l	32	003	1	2	girinata	dukupuntang	cirebon	jawa barat
11	agus rahmat effendi	l	30	003	1	2	girinata	dukupuntang	cirebon	jawa barat
12	agus surahman	l	32	003	1	2	girinata	dukupuntang	cirebon	jawa barat
13	ahari murti	l	17	003	1	2	girinata	dukupuntang	cirebon	jawa barat
14	ahmad badrus salam	l	22	003	1	2	girinata	dukupuntang	cirebon	jawa barat
506	ajat	l	35	003	1	2	girinata	dukupuntang	cirebon	jawa barat

.

3.2. Preprocessing

After the data collection stage, the next stage will be Data Cleaning or data cleaning so that there is no duplication of data, checking for inconsistent data and correcting errors in the data such as printing errors, so that the data can be processed and entered into the data mining process. After all the required data has gone through the data cleaning stage. Next, the data will be processed into training data and testing data which are ready to be entered into the software used in this research.

no	nama	jenis kelamin	usia	rt	tps	blok	jarak
1	aan darwati	p	38	003	1	2	200
2	abdulloh mauludi mujtaba	l	18	003	1	2	100
3	abet	l	71	003	1	2	100
4	achmad fadila	l	33	003	1	2	100
5	achmad permana hadi	l	61	003	1	2	100
6	achmad rohim	l	29	003	1	2	100
7	ade suharto	l	51	003	1	2	100
8	aef saefudin	l	41	003	1	2	100
9	agna yasa nurhabillah	p	17	002	2	1	100
10	agus adi saputra	l	32	003	1	2	100
11	agus rahmat effendi	l	30	003	1	2	100
12	agus surahman	l	32	003	1	2	100
13	ahari murti	l	17	003	1	2	100
14	ahmad badrus salam	l	22	003	1	2	100
506	ajat	l	35	003	1	2	100

3.3. Transformation

Data transformation is the stage of processing data into a form suitable for processing in data mining. In this research, the data to be processed comes from Microsoft Excel and will be used for processing in Rapidminer software. This data includes several attributes, namely number, name, gender, age, environment, Tps, block and distance.

no	nama	jenis kelamin	usia	rt	tps	blok	jarak
1	aan darwati	p	38	003	1	2	200
2	abdulloh mauludi mujtaba	l	18	003	1	2	100
3	abet	l	71	003	1	2	100
4	achmad fadila	l	33	003	1	2	100
5	achmad permana hadi	l	61	003	1	2	100
6	achmad rohim	l	29	003	1	2	100
7	ade suharto	l	51	003	1	2	100
8	aef saefudin	l	41	003	1	2	100
9	agna yasa nurhabillah	p	17	002	2	1	100

10	agus adi saputra	l	32	003	1	2	100
……	……	……	……	……	……	……	……
……	……	……	……	……	……	……	……
506	ajat	l	35	003	1	2	100

3.4. Data Mining

In this stage, the research focuses on model development using the K-Means clustering method. The implementation involves detailed steps, including determining optimal parameters for the *K-Means*, such as the node selection criteria and the maximum tree depth. Below is the application of the data mining process performance on *Rapidminner* using the K-Means method.

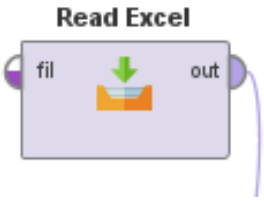

Fig. 2: read excel for data
After the Read Excel operator is executed, information is obtained as in the table

no	uraian	isi
1	record	500
2	special attribute	1
3	regular attribute	7
4	attributes:	
	nama	polynominal, mising 0
	cluster	nomial, mising 0
	jenis kelamin	integer, mising 0
	usia	integer, mising 0
	rt	integer, mising 0
	tps	integer, mising 0
	blok	integer, mising 0
	jarak	integer, mising 0

In the next stage, the Filter Examples operator is an operator used to filter data based on certain criteria. This operator ensures that only data that is relevant and meets predetermined conditions is passed to the next process. In this process, Filter Examples are used to filter data, for example to delete rows with empty values (missing values) or select data based on certain conditions, such as AGE > 18 to filter data on voters who are more than 18 years old. The use of this operator is very important to ensure that the dataset used in the analysis only contains clean and relevant data. The following picture can be seen

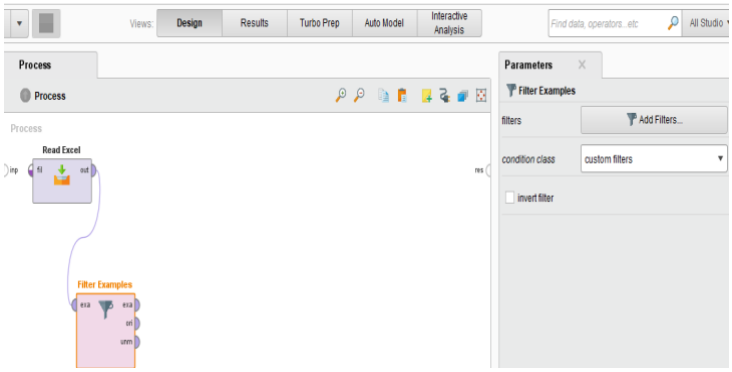

Fig. 3: filter example

The select attributes operator is used to determine the selected attributes in data processing in Rapidminer software. In the dataset, there were initially 10 attributes, then 6 attributes were selected.

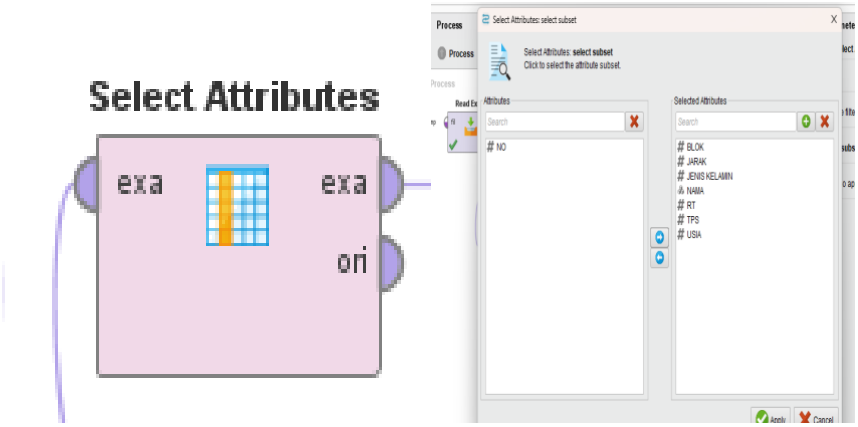

Fig. 4: select atribute

The parameters for the operator set role used are as No Parameter Contents Attributes Name Target role ID

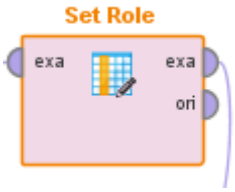

Fig. 5: Set Role

The next stage is determining the number of clusters which must be carried out. The cluster or K value used is determined based on the number of optional clusters and the number of optional members in each cluster. In this process. is the main operator used in modeling the clusterization process.. The parameters used in the K-Means Clutering operator are the values K=5 to k10, the number of Max Optimization steps 1-6 and Measure Type Numerical Measures with the type Euclidean Distance, and Manhattan Distance. The parameters of the K-Means Clustering operator used are, for example

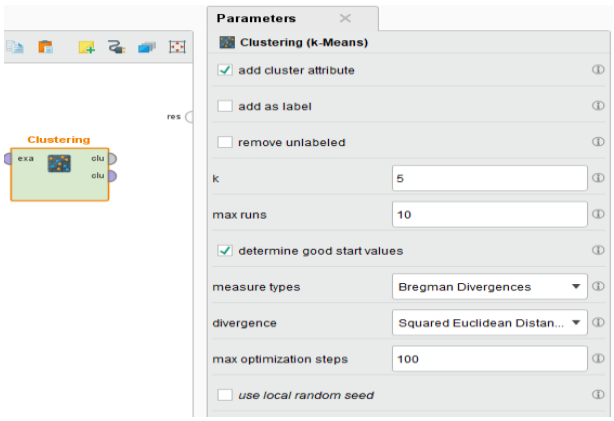

Fig. 6: clustering

Value of Number of Cluster Members

Cluster	Jumlah Anggota
Cluster 0	153 items
Cluster 1	126 items
Cluster 2	50 items
Cluster 3	58 items
Cluster4	113 items

Based on the results of the Davies-Bouldin Index (DBI) analysis, the optimal number of clusters is K=5, with a DBI value of around 0.389. This value is based on the number of clusters studied, indicating that the cluster with K=5 has the best quality. The clusters formed with K=5 have clear separation between clusters and high compression in each cluster, so that the data in each cluster is denser around the centroid, one after another. On the other hand, with a larger number of clusters, such as K=6 (DBI=0.450) or K=7 (DBI=0.470), the DBI increases, indicating that the quality of the clusters is getting better. Increasing the number of clusters does not provide any further benefit, but instead makes the clusters less efficient and harder to read. Thus, the K=5 cluster is considered the optimal configuration because it produces the most optimal and efficient clustering in analyzing data. These results are broken down into average distance values per cluster:
cluster_0: 283,986
cluster_1: 56,776

cluster_2: 260,653
cluster_3: 267,703
cluster_4: 104,715

3.5. Evaluation

Clustering results using the K-Means algorithm show that the average distance of data to the centroid in each cluster (avg. within centroid distance) has significant variations. Clusters with a low average distance, such as Cluster_1 (56,776), indicate that the data in that cluster has high similarity. On the other hand, clusters with a higher average distance, such as Cluster_0 (283,986), indicate that the data is more spread out or that there are outliers in the cluster. The Davies-Bouldin Index (DBI) value of 0.389 indicates that the clustering has very good quality. A DBI value close to 0 indicates well-defined clusters, with significant distances between one cluster and another cluster and data focused around the centroid. Evaluation of the average distance between centroids in each cluster gives an idea that some clusters may have different data distributions. less than optimal. Clusters such as Cluster_3 (267,703) and Cluster_0 (283,986).

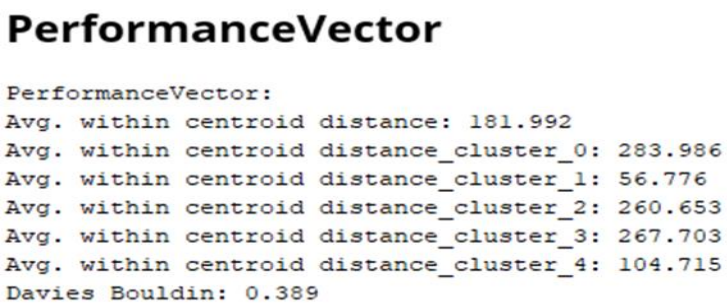

Fig. 7 evaluate model

Can be further evaluated to ascertain whether the cause of this spread is an outlier or a natural pattern in the data. Based on these results, clustering was successful in dividing data with fairly high quality, but there is still opportunity to improve the cluster distribution by performing additional preprocessing. Steps such as data normalization or outlier removal can sharpen cluster division. Clusters with low average distance values, such as Cluster_1, indicate that the data in that cluster is more focused and has stronger relationships between data. This is in accordance with the purpose of clustering, where data with similar characteristics is grouped in one cluster. Significant differences in the average distance between clusters indicate variations in the size or distribution of data within each cluster. This could be an indication that the number of clusters (k) parameter needs to be reviewed to ensure the optimal number of clusters. The evaluation results show that the K-Means algorithm is able to group data well, but some clusters, such as Cluster_0 and Cluster_3, need further research. This is important to ensure that the cluster is not too large or has too diverse data. A low DBI value also indicates that the clusters have good separation. This means that the data between clusters is sufficiently separated so that there is no significant overlap, which is an indication of effective clustering. Use of metrics such as avg.within centroid distance and DBI provide deep insight into the quality of clustering. These values help identify areas that need improvement, such as the distribution of data within certain clusters or the selection of a more optimal number of clusters. Overall, the evaluation results show that the clustering has good quality with clear separation between clusters. However, to improve the results further, additional analyzes can be performed by adjusting the data preprocessing or re-evaluating the number of clusters (k) parameter. This evaluation provides a strong basis to support the interpretation of results and further implementation.

The results of this research show that the K-Means algorithm was successfully applied to group the Permanent Voter List (DPT) data at the Girinata Village TPS with very good quality, as shown by the Davies-Bouldin Index (DBI) value of 0.389. This value shows that the clusters formed have clear separation and good compaction. This is in accordance with research (Huda et al., 2021), which states that the K-Means algorithm is effective in analyzing data with previously unknown structures. Data preprocessing, such as normalization and handling empty values, also makes a significant contribution to optimal clustering results, as suggested by (Nasution et al., 2020). Factors that influence cluster formation in this study include the attributes age, RT, TPS, and distance to polling station. These findings support research by (Puteri et al., 2023) and (Susilowati & Wicaksono, 2024), which highlight the importance of demographic attributes and geographic analysis in voter data. In addition, the clustering results identified variations in the data distribution in several clusters, such as Cluster_0 and Cluster_3, which showed a larger average distance to the centroid. This is in line with the findings of (Patimah et al., 2021),

which states that outliers or unbalanced data distribution can influence clustering results**.** This research also shows that the K-Means algorithm can be used to detect inaccuracies in DPT data, such as data that does not match the TPS location or age groups that do not qualify as voters. This is in line with research (Mulyana et al., 2024), which discusses the effectiveness of K-Means in analyzing numerical data to support data-based decision making. In addition, selecting the right number of clusters is important, as explained by (Fauzi et al., 2022), to ensure optimal data distribution and more representative clustering results. Overall, this research strengthens previous literature and shows that the K-Means algorithm is effective in helping to improve the accuracy of DPT by systematically grouping data. With quality clustering results, this research provides a strong basis for better voter data management, while supporting more transparent and accountable election data validation.

References

[1] S. D. Hilda, A. Voutama, and Y. Umaidah, "Analisis Daftar Pemilih Tetap Pemilihan Gubernur dan Wakil Gubernur menggunakan Algoritma K-Means," *JATISI (Jurnal Tek. …*, vol. 10, no. 3, pp. 398–408, 2023, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/4921%0Ahttps://jurnal.mdp.ac.id/index.php/jatisi/article/download/4921/1600

[2] A. Fauzi, A. Maulana, and A. Setiawan, "Perancangan Aplikasi Rekapitulasi Hasil Pemilu Sementara Berbasis Android Mobile," *CONTEN Comput. Netw. Technol.*, vol. 2, no. 1, pp. 17–26, 2022, doi: 10.31294/conten.v2i1.1179.
[3] K. Mulyana, N. Rahaningsih, and R. Danar Dana, "Analisis Pengelompokkan Dataset Pemilu 2014 Dan 2019 Dpr Ri Di Kota Cirebon Menggunakan Algoritma K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3821–3829, 2024, doi: 10.36040/jati.v7i6.8212.
[4] W. I. Rahayu, A. Anindita, and M. N. Fauzan, "PENENTUAN VALIDASI DATA PEMILIH DAN KLASIFIKASI HASIL PEMILU DPRD KAB.BONE UNTUK MEMPREDIKSI PARTAI PEMENANG MENGGUNAKAN METODE NAIVE BAYES Program Studi D4 Teknik Informatika 123 Politeknik Pos Indonesia 123," *J. Tek. Inform.*, vol. 14, no. 1, pp. 32–39, 2022.
[5] Y. P. Prayogi, H. Wintolo, and Y. Indrianingsih, "Perancangan Dan Penerapan Algoritma Nakula Sadewa Untuk Mengatasi Duplikasi Pemilihan Di Tempat Pemungutan Suara," *Compiler*, vol. 2, no. 2, pp. 1–12, 2013, doi: 10.28989/compiler.v2i2.43.
[6] I. Mahendra, "Implementasi Kebijakan Pendataan Pemilih Dalam Pemilihan Kepala Daerah Kota Malang 2013," *J. Ilm. Ilmu Sos. dan Ilmu Polit.*, vol. 8, no. 1, pp. 28–36, 2018.
[7] W. Weriza, A. Asrinaldi, and E. Arief, "Budaya Kerja Petugas Pemutakhiran Data Pemilih Dalam Pemilukada Di Kota Padang Panjang," *J. Antropol. Isu-Isu Sos. Budaya*, vol. 20, no. 2, p. 213, 2019, doi: 10.25077/jantro.v20.n2.p213-222.2018.
[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.