Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 2, 2024, pp. 207 – 215

K-MEANS ALGORITHM IN CLUSTERING SALES DATA FOR CALCULATING ESTIMATED HOUSE PRICES

Gatot Tri Pranoto [1*]
[1] *Information System Study Program, Faculty of Science, Engineering and Design, Trilogi University*
email: gatot.pranoto@trilogi.ac.id [1]

Abstract - *Determination of the value of the guarantee to the Bank in the process of applying for Home Ownership Credit (KPR) submitted by prospective customers still refers to the provisions of the Financial Services Authority, where the assessment must follow the existing rules and be carried out by public appraisals or commonly called the Office of Public Appraisal Services (KJPP). Currently the analyst credit officer cannot validate the results of the assessment report from KJPP, so if an error occurs either intentionally or not by KJPP or appraisal parties continue to process according to the given value. In the event of default of payment by the customer due to the lower collateral value of the loan provided, the bank violates Bank Indonesia Regulation number 18/16/PBI/2016 concerning loan to value ratio. This study aims to apply the K-Means algorithm in grouping home sales so that it can be used for the calculation of the estimated value of house prices, and develop a prototype of the house price estimation information system. Data retrieval using crawling or scrapping techniques on the website makes it easier to fulfill data on a dataset. The result of this study is the data of home sales for kebon Jeruk area spread across the internet can be grouped into 3 clusters with the value of David Bouldin Index in duri Kepa sub area, which is 0.096, in South Kedoya sub area of 0.087, in North Kedoya sub area of 0.071, and Kelapa Dua sub area of 0.117. By combining clusterization results using K-Means methodology with land price calculation formula obtained land price estimation in sub area.*

Keywords: K-Means, KPR, Data Scraping, KJPP, MAPPI.

1. INTRODUCTION

The growth of the financial industry is currently very rapid, especially in the world of financial industry such as banking. financial industry such as banking. This increase in growth is also accompanied by the growth of Non-Performing Loan (NPL) in the household sub-sector for residential ownership or commonly called Home Ownership Loans (KPR).

The effect of NPLs can result in reduced bank income, the reduction arises because of the additional costs that arise reduction arises because of the additional costs that arise due to non-performing loans (Sari, 2012). (Sari, 2012). The impact of non-performing loans will certainly also be problematic for bank customers, because the house that is used as collateral for the bank will be auctioned by the bank to cover the bank's losses. bank to cover the bank's losses.

There are many factors that cause NPLs, due to the country's economic conditions economic conditions, the customer is experiencing financial difficulties or a process error in the bank that causes errors in the process errors in the bank that cause errors in providing loan or collateral values.

One of the mistakes in the bank is in the credit analysis process where problems are found by credit analysts in validating the loan value or collateral. problems by credit analysts in validating the value of collateral provided by KJPP (Public Appraisal Services Office).

(Public Appraisal Services Office), because credit analysts do not have data to confirm the validity of the collateral value provided by KJPP.

ensure the validity of the collateral value provided by the KJPP so that it has the potential for errors in granting loan limits that can cause losses to the bank, and these losses will have an impact on the bank. bank, and this loss will have an impact on the risk of the Bank's reputation and can violate the loan to value ratio regulation in Bank Indonesia Regulation Number 18/16/PBI/2016.

In addition to NPL problems, there is also competition between banks where all banks are competing to improve the ability of their information systems to provide speed in the mortgage application process, especially applications that can help sales to be able to estimate the value of collateral for prospective customers.

The k-means algorithm has a better performance than k-medoids, either in terms of average within centroid distance value and time complexity (Mediana et al., 2018), so the authors used the k-means algorithm, 2018), so the authors use the K-Means algorithm for grouping which group is the most dominant by eliminating anomalous data to be used as a reference or fair value on land values in an area in a city.

2. RESEARCH METHOD

In this section of the research methodology, systematic and directed steps will be described that will be used as a reference as a research framework for determining the similarity of problems in determining appraisal price estimates, using K-means so that it can be known which method produces the best cluster results and gets the closest price estimates, which method produces the best cluster results and gets the closest price estimate.

2.1 Knowledge Discovery In Databases (KDD)

KDD is a method for obtaining knowledge from existing databases. In the database there are tables that are interconnected. The results of the knowledge obtained from the process can be used for the knowledge base for decision-making purposes. The KDD process is largely data selection, pre-processing / cleaning, transformation, data mining, evaluation, and knowledge (Riadi et al. 2020).

1) Data Selection
2) Pre-Processing
3) Data Transformation
4) Data Selection

2.2 Population and Sample Selection Method

Sampling data is a technique used to systematically select a smaller number of data representatives from a predetermined population to serve as a source of data for observation or experimentation according to objectives, and in accordance with research science and statistics, sampling procedures must be carried out several important factors (Sharma, 2017).

2.3 Data Collecting Method

Data collection using data collection techniques is quantitative in nature to test hypothesis that has been determined with techniques such as the following:

1) Data Crawling
2) Data Cleaning

2.4 K-Means Clustering

The K-Means algorithm is an iterative clustering algorithm that partitions a data set into a predetermined number of K clusters (Parlina et al., 2018). The method will divide the data into several groups where the groups have the same properties or characteristics. Steps in the clustering stages:

1) Specify k as the number of clusters to be formed.
2) Specify cluster center.
3) Calculate the distance of each data to the cluster center using the Euclidean equation.

$$d_{ik} = \sqrt{\sum_{j}^{m} (Cij - Ckj)^2}$$

4) Group the data into the cluster with the shortest distance using the equation.

$$Min \sum_{K=1}^{K} d_{ik} = \sqrt{\sum_{j}^{m} (Cij - Ckj)^2}$$

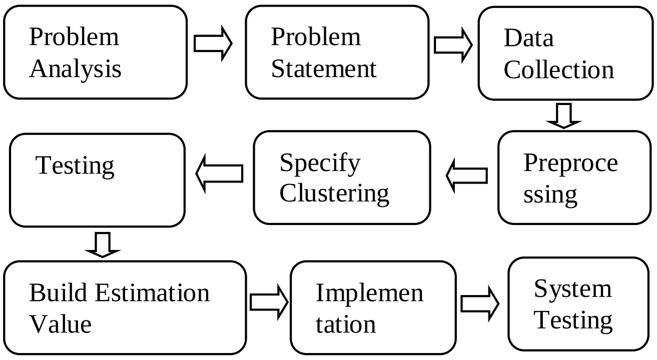
5) Calculate the new cluster center using the equation.

$$C_{kj} = \frac{\sum_{i=1}^{p} xij}{p}$$

6) Repeat steps 2 through 4 until there is no more data moving to other clusters.

2.5 Resarch Step

The steps that will be carried out in this research in general are as shown in the figure below.



3. RESULTS AND DISCUSSION

The K-Means algorithm is an iterative clustering algorithm that partitions a data set into a predetermined number of K clusters (Parlina et al., 2018). The method will divide the data into several groups where the groups have the same properties or characteristics. The purpose of the clustering is to minimize the diversity in a group and maximize the type in the group (Agustin et al., 2015).

3.1 Data Acquisition

For initial data acquisition, what is done is crawling data from e-commerce websites according to the needs of the research. The web-scrapping process takes data from home sales websites such as rumah123.com and rumah.trovit.co.id. In this case, the tools used in web scrapping are python programming language and jupyter notebook. The data is taken from one of the websites for buying and selling houses, namely www.rumah123.com and www.trovit.rumah.co.id, which is one of the e-commerce websites that provides property sales in Indonesia.

3.2 Data Processing

The next stage is processing the data that has been retrieved through crawling. Attribute data that is selected to be included in the research category are title, land area, building area, price, area and also sub area. In this case, the author takes the Kebon Jeruk area data as a calculation simulation and the sub areas obtained in the Kebon Jeruk area are Duri Kepa, Kelapa Dua, South Kedoya, North Kedoya.

3.3 Land Price Calculation

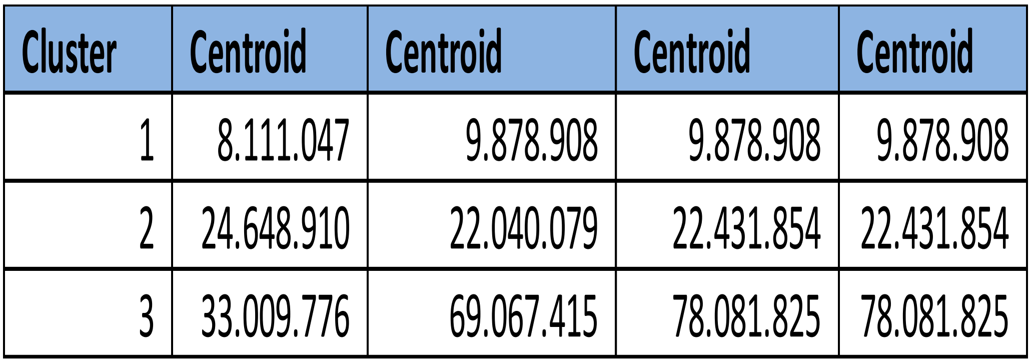
After data transformation, the next step is to calculate the land price with the following formula:

Building Price = Building Area × Building Price Per M^2
Land Price = (House Price - Building Price) / Land Area

3.4 Centroid Data

In the application of the K-means algorithm, the midpoint or centroid value of the data is generated. The process of finding the midpoint value is done by taking the largest value (maximum) for high-level clusters (C1), the average value (average) for medium-level clusters (C2) and the smallest value (minimum) for low-level clusters (C3).

TABLE 1. CENTROID DATA

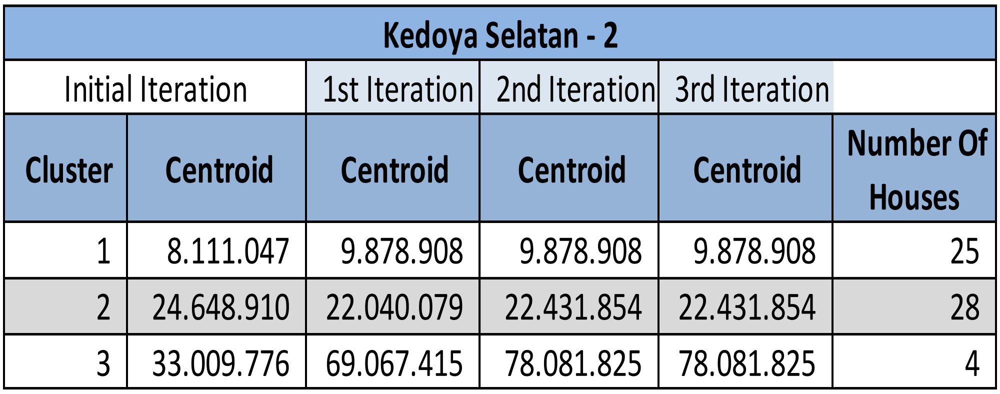


Cluster	Centroid	Centroid	Centroid	Centroid
1	8.111.047	9.878.908	9.878.908	9.878.908
2	24.648.910	22.040.079	22.431.854	22.431.854
3	33.009.776	69.067.415	78.081.825	78.081.825

3.5 Data Clustering

The clustering process using the initial centroid value contained in table 1, will obtain the clustering results in literation 1 which can be seen in the tables below:

TABLE 2. DATA CLUSTERING



Kedoya Selatan - 2					
Initial Iteration		1st Iteration	2nd Iteration	3rd Iteration	
Cluster	**Centroid**	**Centroid**	**Centroid**	**Centroid**	**Number Of Houses**
1	8.111.047	9.878.908	9.878.908	9.878.908	25
2	24.648.910	22.040.079	22.431.854	22.431.854	28
3	33.009.776	69.067.415	78.081.825	78.081.825	4

3.6 Estimatation Calculations

Continuing the grouping process, the raw data that has been formed is used as a data source for calculating land prices. Meanwhile, the building price refers to the calculation of building prices based on the results of the MAPPI matrix table.

Building Price = Building Area × Building Price Per M^2
Land Price = Land Price in the Area × Land Area
Estimated House Price = Land Price + Building Price

3.7 Data Visualitation

The first visualization is to illustrate the distribution of land price data because of grouping and calculating land price estimates combined with geotagging information for longitude and latitude positions in each sub-area, so as to get the distribution of land price estimation results with the results as shown in Figure 1 below:

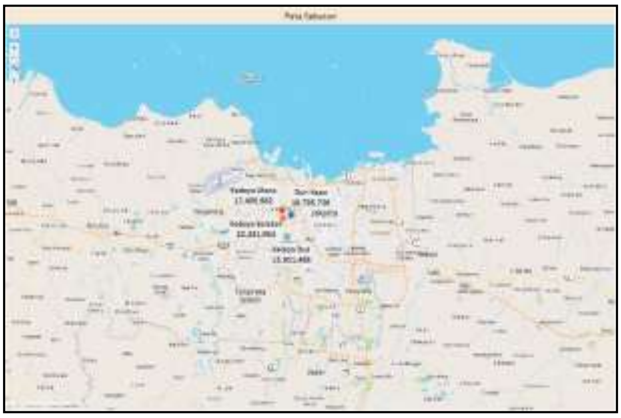


Fig. 1. Data Distribution

In addition to using visualization of the distribution of estimated land prices based on geotagging, the following below is the result of clustering or grouping land prices based on each group that can provide an overview of the estimated land prices in each cluster in a sub-area according to Figure 2 below:

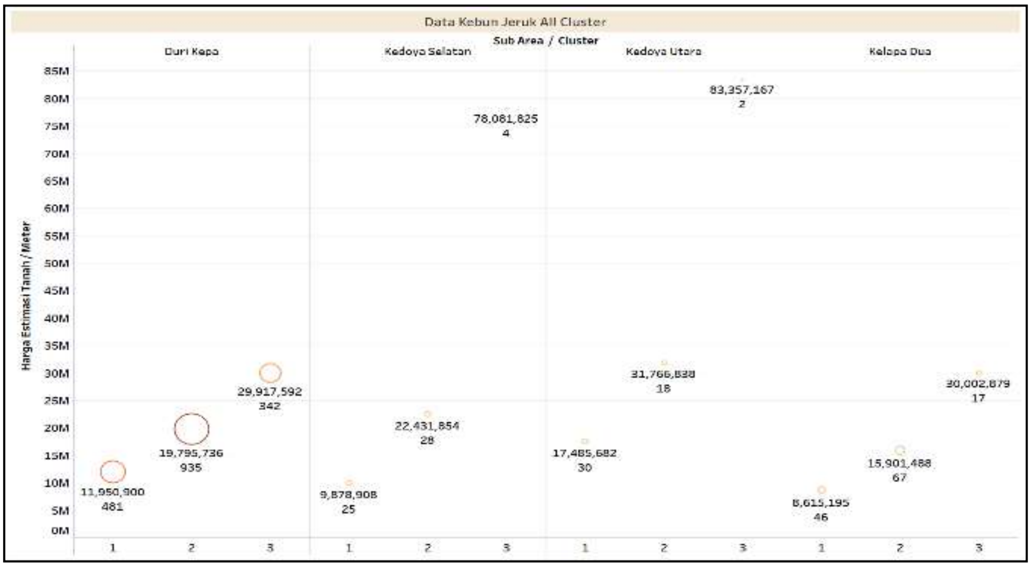
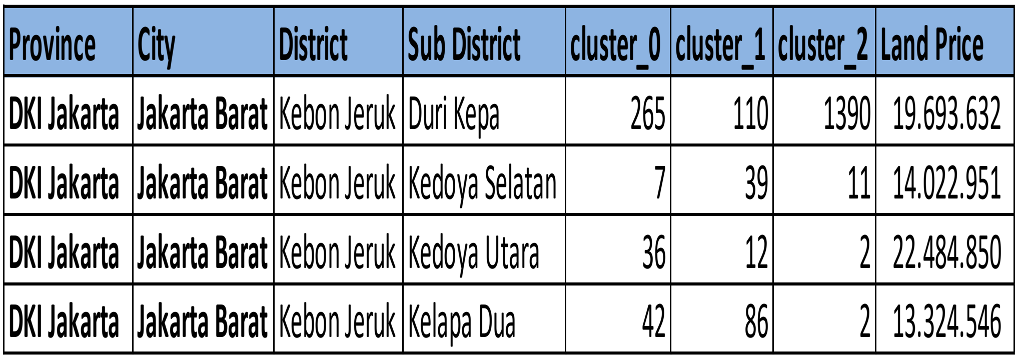


Fig. 2. Sub District Distribution

3.8 Data Result

The results of the implementation of the K-Mean's methodology combined with the land price calculation process, where the dominant clusters will be averaged and produce data as in table 3 belowthe distribution of land price estimation results with the results as shown in figure as below:

TABLE 3. DATA RESULT



Province	City	District	Sub District	cluster_0	cluster_1	cluster_2	Land Price
DKI Jakarta	Jakarta Barat	Kebon Jeruk	Duri Kepa	265	110	1390	19.693.632
DKI Jakarta	Jakarta Barat	Kebon Jeruk	Kedoya Selatan	7	39	11	14.022.951
DKI Jakarta	Jakarta Barat	Kebon Jeruk	Kedoya Utara	36	12	2	22.484.850
DKI Jakarta	Jakarta Barat	Kebon Jeruk	Kelapa Dua	42	86	2	13.324.546

3.9 Integrated System Design

In general, the system requirements are to provide solutions faced by the Bank in the process of receiving credit applications by the marketing team and the credit analysis process by the credit analyst team. The following is a description of the system needed, according to Figure 3 below:

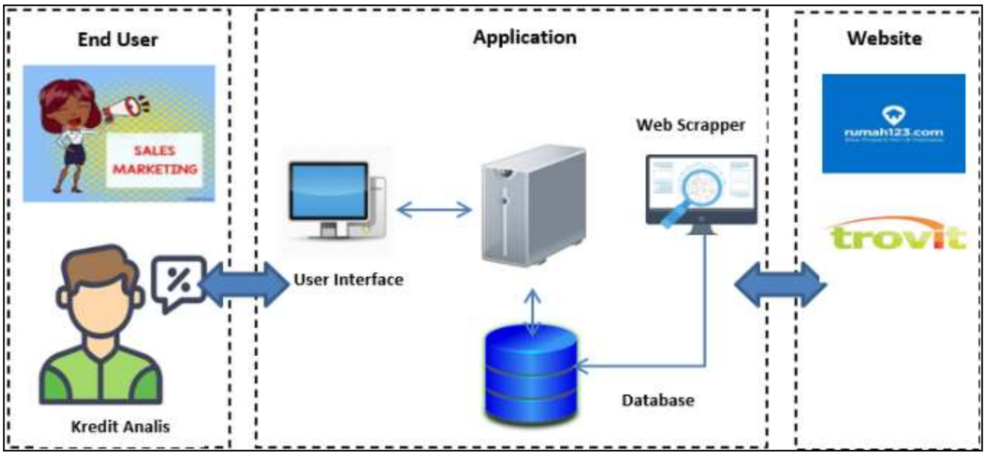


Fig. 3. Sub District Distribution

Figure 3 is a picture of the design of a house price estimation application that uses grouped data that will be used by sales and credit analysis.

3.10 Prototype Implementation

At this stage of prototype implementation, the prototype is built with PHP programming language as an interface display and MYSQL as a database storage of processing results, here is the appearance of the prototype:

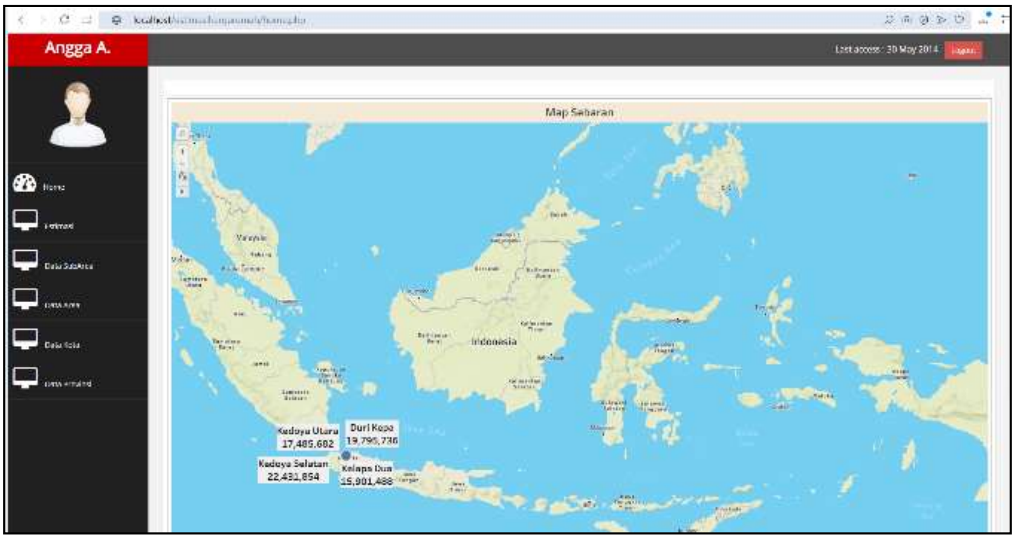


Fig. 4. Sub District Distribution

4. CONCLUSION

The results of the research that has been conducted by the author are as follows:

1. We can take home sales data scattered on the internet using web scrapping to be used as a data source.
2. Tests conducted in this study resulted in 3 clusters with a large David Bouldin Index value in the Duri Kepa sub area of 0.096, in the South Kedoya sub area of 0.087, in the North Kedoya sub area of 0.071, and the Kelapa Dua sub area of 0.117.

3. By combining the clustering results using the K-Means methodology with the land price calculation formula, the estimated land prices in the sub-area.

REFERENCES

[1] Ahmad Syaripul, N., & Mukharil Bachtiar, A. (2016). Visualisasi Data Interaktif Data Terbuka Pemerintah Provinsi Dki Jakarta: Topik Ekonomi Dan Keuangan Daerah. Jurnal Sistem Informasi, 12, 15–29

[2] Agustin, F. E. M. (2015). Implementasi Algoritma K-Means Untuk Menentukan Kelompok Pengayaan Materi Mata Pelajaran Ujian Nasional (Studi Kasus: Smp Negeri 101 Jakarta). Jurnal Teknik Informatika, 8(1), 73–78. https://doi.org/10.15408/jti.v8i1.1938.

[3] Bhatia, P. (2019). Introduction to Data Mining. Data Mining and Data Warehousing, 17–27. https://doi.org/10.1017/9781108635592.003.

[4] Dogan, O., Aycon, E., & Bulut, Z. A. (2018). Customer Segmentation By Using Rfm Model and Klasterisasi Methods : a Case. International Journal of Contemporary Economics and Administrative Sciences, 8(July), 1–19.

[5] Edy Irwansyah, S.T., M. S. (2017). KLASTERISASI. Https://Socs.Binus.Ac.Id. https://socs.binus.ac.id/2017/03/09/Klasterisasi/Klasterisasi merupakan metode segmentasi data, computer vision dan image processing.

[6] Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat-Obatan. Jurnal Nasional Teknologi Dan Sistem Informasi, 5(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24.

[7] Kamila, I., Khairunnisa, U., & Mustakim. (2019). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau. Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi, 5(1), 119–125. http://garuda.ristekdikti.go.id/documents/detail/1051741.

[8] Mardi, Y. (2017). Data Mining: Klasifikasi Menggunakan Algoritma C4.5. Jurnal Edik Informatika, 2(2), 213–219.

[9] Mediana, Madyatmadja, E. D., & Miranda, E. (2018). Application of K-Means and K-Medoids Klasterisasi Pada Data Internet Banking Di Bank Xyz. 349–356.

[10] Metisen, B. M., & Sari, H. L. (2015). Analisis Klasterisasi menggunakan metode K-Means dalam pengelompokkan penjualan produk pada Swalayan Fadhila. Jurnal Media Infotama, 11(2), 110–118.

[11] Muhammad, A. F. (2015). Pengelompokkan Proses Seleksi Pemain Menggunakan Algoritma K-Means (Study Kasus: Tim Hockey Kabupaten Kendal). Jurusan Teknik Informatika FIK UDINUS, 1–5.

[12] Nurul rohmawati, sofi defiyanti, mohamad jajuli. (2015). Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. Jitter 2015, I(2), 62–68.

[13] Sibuea, M. L., & Safta, A. (2017). Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring. Jurteksi, 4(1), 85–92. https://doi.org/10.33330/jurteksi.v4i1.28.

[14] S.C.M. de S Sirisuriya. (2015). A Comparative Study on Web Scraping. 8th International Research Conference KDU, November, 135–140. http://www.kdu.ac.lk/proceedings/irc2015/2015/com-020.pdf.

[15] Steinbach, M., Tan, P., & Boriah, S. (2006). The Application of Klasterisasi to Earth Science Data: Progress and Challenges. The Application of Klasterisasi to Earth Science Data: Progress and Challenges Michael, 1–6.

[16] Wardhani, A. K. (2016). Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan. Jurnal Transformatika, 14(1), 30–37.

[17] Windarto, A. P. (2017). Penerapan Datamining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Klasterisasi Method. Techno.Com, 16(4), 348–357. https://doi.org/10.33633/tc.v16i4.1447.

[18] Sari, T. M., Syam, D., & Ulum, I. (2012). Pengaruh Non-Performing Loan Sebagai Dampak Krisis Keuangan Global terhadap Profitabilitas Perusahaan Perbankan. Jurnal Akuntansi & Investasi, 13(2), 83–98.

[19] Sharma, G. (2017). Pros and cons of different sampling techniques. International Journal of Applied Research, 3(7), 749–752. www.allresearchjournal.com.

[20] Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. International Journal of Engineering Research and Applications Www.Ijera.Com.

[21] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. IEEE Access. https://doi.org/10.1109/ACCESS.2014.2362522

[22] Yadav, J., & Sharma, M. (2013). A Review of K-mean Algorithms. International Journal of Engineering Trends and Technology.