Client Segmentation of Mobile Payment Parking Data Using Machine Learning

Ilze Andersone, Agris Ņikitenko, Valdis Bergs, Uldis Jansons

Client Segmentation of Mobile Payment Parking Data using Machine Learning

Ilze Andersone[1] , Agris Ņikitenko[1], Valdis Bergs[2] and Uldis Jansons[2]

[1] Riga Technical university, Riga, Latvia
[2] SIA Mobilly , Riga, Latvia

Abstract. This paper addresses the analysis of mobile payment parking data for client segmentation. The transaction data transformation into client-specific attributes is performed from the company data set to achieve the goal. Two clustering algorithms – K-Means and DBScan – are compared for multiple data subsets. For the clustering result interpretation, decision tree representation is used. As a result, the most appropriate combination of the clustering algorithm, its parameters and attribute combination is determined.

Keywords: Client segmentation, Clustering, Mobile payments, Parking

1 Introduction

The client segmentation from the historical data is a relevant research object for many companies, as it can give multiple benefits – help to choose an appropriate marketing strategy [1, 2, 3], assist in the creation of personalized offers to the existing clients [4], as well as help to understand the clients better and/or react to their potential needs in a timely manner [5]. Since the company offers multiple services, client clustering can also benefit from identifying their most essential services and how they relate to their customer base. In our case study, the company offers several services through mobile payments, emphasizing client parking data.

A lot of research has been done on the parking data, but it mostly focuses on the assistance in assisting in finding parking spots where geographical information is available [6, 7, 8, 9, 10]. Sometimes specific topics are addressed – such as autonomous vehicle parking [11] or smart parking [12]. Some papers address individual parking behaviour [13, 14], but it is mainly done only to simulate parking space availability.

The goal of the client segmentation in our research is to identify the primary client groups from the historical data of their mobile payment information so that the most appropriate marketing strategies and loyalty programs can be developed without overwhelming the clients with irrelevant offers. E.g., clients who mostly park inexpensive parking places would have different needs from those that only ever use public transport (buses or trains).

Though some client groups can be identified by proposing hypotheses and then statistically confirming or rejecting them, often the data can be overwhelming to analyze manually, and some essential client segments can quickly go unidentified.

2 Data collection and preprocessing

The data set used in our research contains the service, payment, and temporal data of 19 million individual transactions during the time period from January 2017 to August 2019. Mobilly [15] collected the data from a mobile payment service provider in Latvia. All the client data was anonymized to ensure EU data privacy standards.

The data set contains information about various services that can be paid for through the mobile phone app – parking, bus and train tickets, theatre tickets, and donations. Even though the company offers various services, parking is the most important business component; only the parking data subset was analyzed in detail. Both the data preprocessing and analysis was performed in Jupyter notebook environment using Python programming language.

The tables contain the transaction data about payments that clients make for the received services (payments/costs) as well as data about the funds they store in their accounts (income). This data differs significantly from other data sets used for parking analysis [6, 7, 8, 9, 10] in that it doesn't contain location data, but the emphasis is instead on payments and various services. Most transactions in the data set relate to parking in various cost parking zones, but there are also transactions related to other mobile payment services.

2.1 Attribute extraction

As the transaction data does not contain any aggregated information about the clients, additional data aggregation is necessary to achieve the goal of client segmentation. Several groups of attributes are extracted from the data set that represents various characteristics of the clients:

- Absolute measures – this group contains attributes representing comprehensive data about the clients during the analyzed period, e.g., Total payment amount over two years and eight months.
- Average measures – this group contains attributes representing averaged data about the client (usually monthly), e.g., Average payment for parking in one month.
- Relative measures – this group contains attributes representing some relative relationship of two other attributes, e.g., Rate of parking payments compared to total costs.

An exhaustive list of attributes is given in **Table 1**. *Costs* represent the client's amount for services, while *income* shows how much the client has paid in his account. Different parking zones are available with standard zones on the streets (A zone is the most expensive and D zone is the cheapest) and parking places with plot-specific costs

(e.g., parking zones by supermarkets or private properties). Clients can make payments for the services as individuals or company funds (represented by *private payment rate*).

Table 1. The measures extracted from the data set

Absolute measures	Average measures	Relative measures
Total months	Avg costs / month	Private payment rate
Total costs	Avg payment count / month	Parking rate
Total payment count	Avg amount paid	A parking rate
Vehicle count	Avg income / month	B parking rate
Total income	Avg income count/month	C parking rate
Total income count		D parking rate
Total parking expenses		Other parking rates
Different service count		

2.2 Data preprocessing

Most machine learning algorithms benefit from the preprocessing of data. To prepare the data for the segmentation, the following steps are performed:

1. Missing value handling.
2. Removal of extreme values in the data.
3. Data standardization.
4. Principal Component Analysis for data dimensionality reduction.

Missing value identification

The original data set is of high quality, and none of the records in the tables contain null or NaN values. However, clients who have never used parking services lack corresponding vehicle counts, parking rates, and other parking-related information. For those clients, the missing values are replaced with zeros.

Removal of extreme values in the data.

Extreme values are removed to further adapt the data for the clustering part of the outliers. This approach has significant benefits because large clients are considered a completely separate group and receive individual offers when they are a target of marketing strategies. Additionally, if their data is left in the data set, the data analysis is made significantly more difficult due to the skewing of the data. As an example, consider **Fig. 1**.

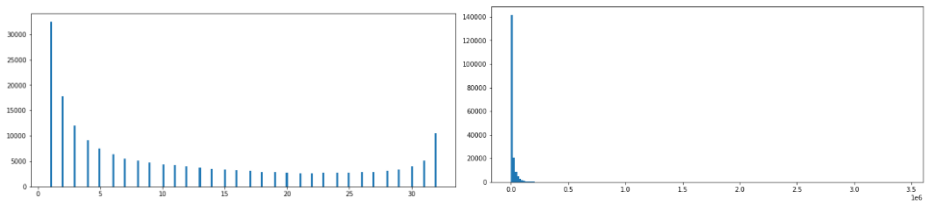

Fig. 1. Histograms of two attribute values. Left: Total month histogram; Right: Total costs histogram.

On the left, the histogram of the client total active months is shown, and it is no surprise that there are many new clients (1 month active), but a significant part of clients have also been active for a longer time. However, a histogram of total costs is shown on the right side of Fig. 1. Although almost all clients are on the far left side of the histogram, some exceptions (large corporate clients) skew the histogram significantly.

To reduce the problem of outliers, some values are removed from the dataset by computing the 1st and 3rd Quartiles of the dataset (values 0.1 and 0.9) and removing all values below lower fence (1st Quartile – 1.5 Interquartile range) and above upper fence (3rd Quartile + 1.5 Interquartile range). For most attributes, no values are removed – only some absolute and average measures are affected.

Data standardization.

The Minmax scaling method is chosen for the data standardization, which transforms all attribute values in the range [0; 1]. While the Minmax scaling method is sensitive to data extremes and outliers, the data extreme removal in the previous preprocessing step reduces this problem and, at the same time, allows it to retain the natural scale of relative attributes.

Principal Component Analysis

The data set doesn't contain any categorical data; therefore, Principal Component Analysis (PCA) reduces data set dimensionality, which extracts the essential details from the data and represents them as a reduced set of variables called principal components [16]. PCA is performed for each data subset separately.

3 Experimental results

Two different clustering methods are chosen for the experiments: K-Means and DBScan. Each of these algorithms has its advantages and drawbacks, and their comparison is beneficial when the data characteristics are unknown.

K-Means is a well-known clustering method, and it is often used for client segmentation [17, 18, 19, 20]. It is a partition-based algorithm that considers the centre of the data points as a centre of the cluster. It has high computational efficiency, but it has several drawbacks – it is unable to detect non-convex shapes in the data, and it is sensitive to cluster count, which is a required parameter for this algorithm [21].

As an alternative to K-Means clustering DBScan algorithm can be used for client data segmentation [17, 22, 23]. DBScan is a density-based clustering method requiring two parameters – the radius of the neighbourhood and the minimum number of neighbour points. Compared to K-Means, DBScan is efficient when the data shape is non-convex, but the results suffer when the data density is not balanced [21].

To compare how different attribute groups impact the client segmentation, five attribute subsets are used for the experiments (refer to **Table 1** for contents of each attribute group):

- Data set with all attributes. Ideally, this group would have comparable segmentation results with attribute subsets.

- Data set with only absolute measure attributes. Represents general information about clients but contains no specific parking data.
- Data set with only relative measure attributes. It primarily contains information about client parking behaviour with the addition of private/company payment rates.
- Data set with only average measure attributes. Represents how much and how often, on average, the payments are made.
- Data set with a selected subset of attributes. All relative and average attributes are included here, along with some absolute attributes that contain information not represented in any other attribute groups – total months, vehicle count and different service count.

All five attribute subsets are used and compared for each clustering algorithm. For the implementation, Jupyter notebook was used [24] that provides a simple and interactive way to process data and develop live code.

3.1 K-Means clustering

The only parameter for K-Means clustering is the cluster count. To find the best cluster count for each attribute subset Silhouette coefficient for 2-15 clusters was calculated. The silhouette coefficient represents how close points in one cluster are to points in another cluster [25]. It can take values [-1;1] where value close to 1 indicates points being far away from each other (good for clustering), 0 – points from different clusters are close and negative values – possible assignment of points to wrong clusters.

Each data point in **Fig. 2** represents an average value of 3 clustering attempts.

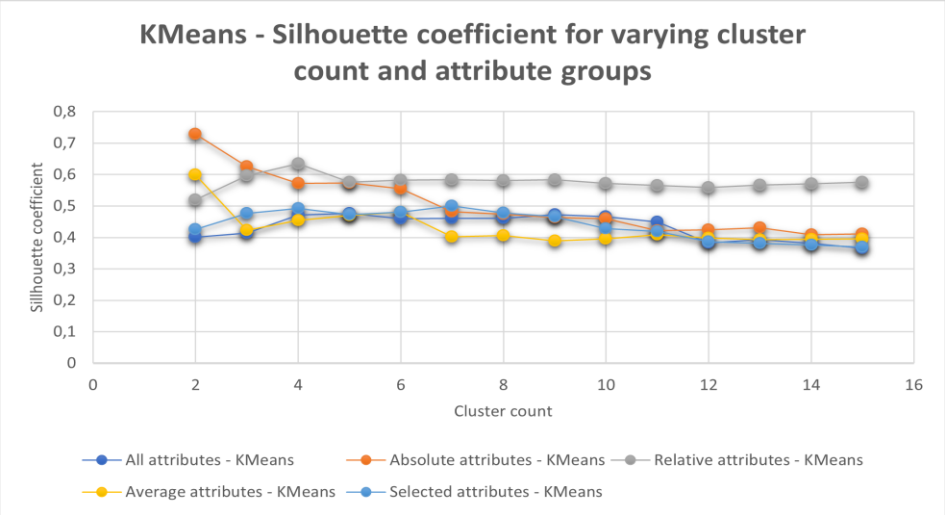

Fig. 2. K-Means clustering results for 2-15 clusters in five data subsets

Although they indicate the clustering quality, the overall highest Silhouette score does not guarantee that it is the best clustering result. The decision trees were used on

the cluster count with highest Silhouette scores for each attribute group to interpret clustering results. Illustrations of decision trees are given in **Fig. 3**.

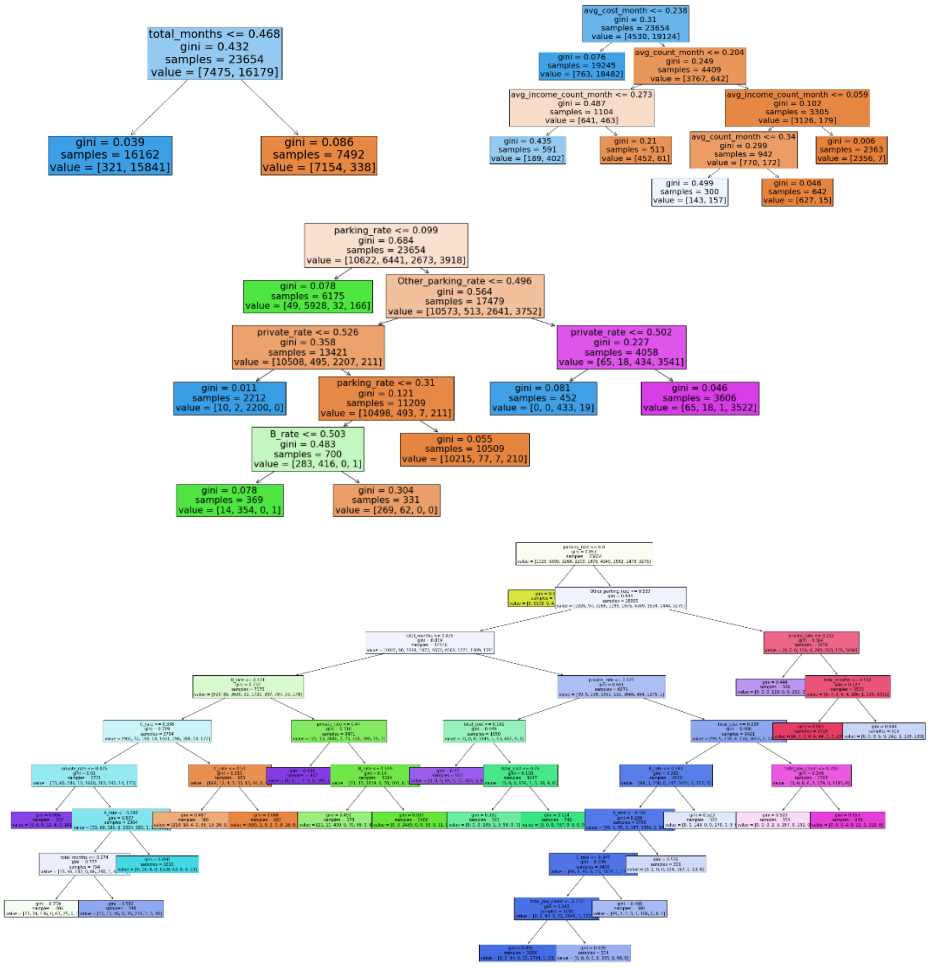

Fig. 3. Visual representation of decision trees for best Silhouette scores in each attribute group. Colours indicate data subset belongs to a particular class. Top left: Absolute measures – 2 clusters; Top right: Average measures – 2 clusters; Middle: Relative measures – 4 clusters. Bottom: All attributes – 9 clusters (also included as additional file for readability). The decision tree for Selected attributes is not included (it has similar characteristics to All attribute classification).

When each decision tree is analyzed in detail, it can be concluded that there is little new knowledge to be found about clients in decision trees with only two classes: Absolute attributes clustering (**Fig. 3** top left) divides the clients by one attribute only – total months they have been active, and Average attribute clustering (**Fig. 3** top middle)

does the same by average costs per month (further classification has an impact on a very small part of data set).

More meaningful results are given by Relative, All and Selected attribute datasets with 4, 9 and 7 clusters as highest Silhouette scores. E.g., some client group examples that can be found in the decision tree analysis are (**Fig. 3** bottom):

- Clients who don't use parking services at all.
- Corporate clients who mainly use parking services and mainly park in parking plots instead of streets.
- Long term private clients who often use street parking services without a strong preference for one specific parking zone.

3.2 DBScan clustering

The DBScan clustering algorithm requires two parameters – *min_samples* that characterize the minimum amount of data examples in any cluster and ε – radius of neighbourhood. The *min_samples* parameter was set to *100* in all experiments. This parameter value was based on the premise that there is little incentive for businesses to identify very small client groups in the data set for targeted marketing purposes. The best value of ε is completely unknown; therefore, for all five data subsets, ε values [0.01-0.3] were tested with a step of 0.01.

Three metrics are used for ε value evaluation (**Fig. 4**) – Silhouette coefficient (same meaning as for K-Means), noise rate (data samples that do not belong to any clusters) and cluster count.

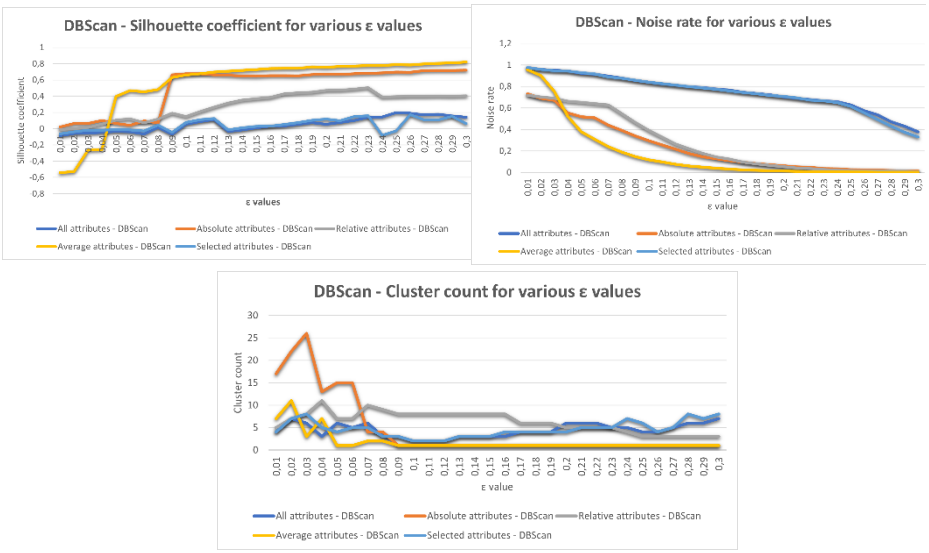

Fig. 4. DBScan clustering results in five data subsets for ε values [0.01-0.3]. Top left: Silhouette coefficient; Top right: Noise rate; Bottom: Cluster count.

The three metrics in **Fig. 4** must be analyzed together to make any conclusions. While one metric in separation may show promising results for a particular ε value and data subset combination, in many cases, other metrics demonstrate that the acquired result is poor. The metric results for the five data subsets can be interpreted in the following way:

- All attributes. Cluster count, in this case, is irrelevant. The Silhouette scores (<0.2) and noise rates (>0.4) are consistently poor for all ε values.
- Absolute attributes. The combination of Silhouette scores (>0.6) and noise rates (<0.1) are suitable for ε values 0.17 and higher. However, cluster count shows that only one cluster is left for the average attribute subset, which is useless for business purposes.
- Relative attributes. Relative attributes have a relatively high Silhouette score (0.5) and noise rate (0.03) combination for ε value 0.23. The cluster count for this ε value is 5, which warrants further analysis of this parameter and data subset combination.
- Average attributes. Similar to absolute attributes, there is a range of ε values (0.1 and higher) that have a good combination of Silhouette scores (>0.6) and noise rates (<0.1). However, there is only one cluster left for this ε parameter range as before.
- Selected attributes. Like all attribute data sets, the metric results for the selected data subset are consistently poor through all tested ε values.

The overall conclusion about the DBScan results for all five attribute sets is that only the relative attribute subset gives any results considered for additional analysis. By far, the largest identified group is clients who use private funds and park at least sometimes. No significant new information about the clients is found by interpreting the results with a decision tree (Fig. 5).

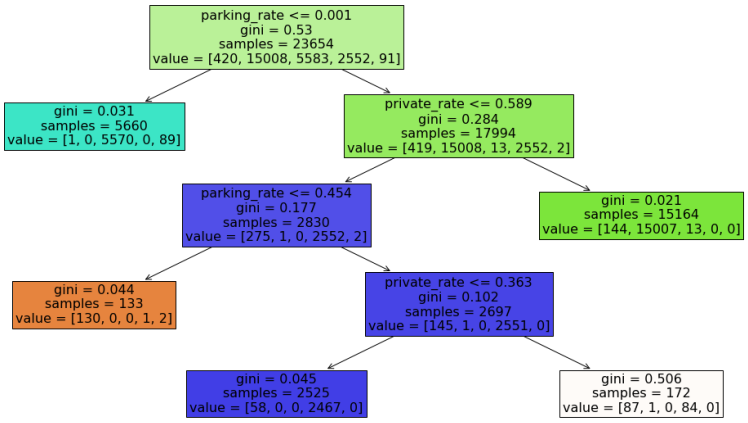

Fig. 5. Visual representation of decision tree for DBScan epsilon 0.23 (Relative attributes)

4 Conclusions

In this research, mobile payment company data was analyzed to target client segmentation. For this purpose, three attribute types (absolute, average and relative) were extracted from the company transaction data, five data subsets were created, and two clustering algorithms – K-Means and DBScan – were compared with different parameters to find the most suitable approach for the acquisition of new information.

Based on the K-Means clustering results with different parameters on five data subsets, it was concluded that although the highest Silhouette scores were acquired in the Absolute, Average and Relative data sets separately, the decision tree interpretation of the clustering gives more meaningful results for All attributes and Selected attributes, where decision tree analysis can yield useful information about the client groups and their proportion in the data set. Unfortunately, no meaningful results were acquired with the DBScan algorithm in any data subsets.

As future work, use of geographical data to support parking suggestions based on the history of user preferences is suggested. This would allow to identify overall client behaviour and give individualized parking suggestions based on regular parking behaviour history. E.g. suggestions of cheaper or safer parking places might not be important for clients who consistently park in the same spots, but more relevant to clients who often park in unfamiliar places.

Acknowledgement

The research leading to these results has received funding from the project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 1.2.1.1/18/A/003 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 1.3 "Research, development prototyping of financial analysis tool based on document management system".

References

1. Li, Y., & Lin, F. Customer segmentation analysis based on SOM clustering. In 2008 IEEE International Conference on Service Operations and Logistics, and Informatics (Vol. 1, pp. 15-19). IEEE. (2008).
2. Maryani, I., & Riana, D. Clustering and profiling of customers using RFM for customer relationship management recommendations. In 2017 5th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-6). IEEE. (2017).
3. Yoseph, F., & Heikkila, M. Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. In 2018 International Conference on Machine Learning and Data Engineering (iCMLDE) (pp. 108-116). IEEE. (2018).
4. Mihova, V., & Pavlov, V. A customer segmentation approach in commercial banks. In AIP Conference Proceedings (Vol. 2025, No. 1). AIP Publishing LLC. (2018).
5. Yuping, Z., Jílková, P., Guanyu, C., & Weisl, D. New Methods of Customer Segmentation and Individual Credit Evaluation Based on Machine Learning. In New Silk Road: Business

Cooperation and Prospective of Economic Development, (pp. 925-931). Atlantis Press. (2020).
6. Alsafery, W., Alturki, B., Reiff-Marganiec, S., & Jambi, K. Smart car parking system solution for the internet of things in smart cities. In 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), pp. 1-5. IEEE. (2018).
7. Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., & Krcmar, H. Predicting the availability of parking spaces with publicly available data. Informatik 2016. (2016).
8. Rong, Y., Xu, Z., Yan, R., & Ma, X. Du-parking: Spatio-temporal big data tells you realtime parking availability. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 646-654. (2018).
9. Liu, K. S., Gao, J., Wu, X., & Lin, S. On-street parking guidance with real-time sensing data for smart cities. In 2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1-9. IEEE. (2018).
10. Hilvert, O., Toledo, T., & Bekhor, S. Framework and model for parking decisions. Transportation research record, 2319(1), pp 30-38. (2012).
11. Millard-Ball, A. The autonomous vehicle parking problem. Transport Policy, 75, pp 99-108. (2019).
12. Piovesan, N., Turi, L., Toigo, E., Martinez, B., & Rossi, M. Data analytics for smart parking applications. Sensors, 16(10), 1575. (2016).
13. Bonsall, P., & Palmer, I. Modelling drivers' car parking behaviour using data from a travel choice simulator. Transportation Research Part C: Emerging Technologies, 12(5), pp 321-347. (2004).
14. Gomari, S., Knoth, C., & Antoniou, C. Cluster analysis of parking behaviour: A case study in Munich. Transportation Research Procedia, 52, pp 485-492. (2021).
15. Mobilly, SIA, (accessed June 2021), https://mobilly.lv/en/about-mobilly/
16. Abdi, H., & Williams, L. J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp 433-459. (2010).
17. Monalisa, S., & Kurnia, F. Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour. Telkomnika, 17(1), pp 110-117. (2019).
18. Ezenkwu, C. P., Ozuomba, S., & Kalu, C. Application of K-Means algorithm for efficient customer segmentation: a strategy for targeted customer services. (2015).
19. Ye, L., Qiu-ru, C., Hai-xu, X., Yi-jun, L., & Zhi-min, Y. Telecom customer segmentation with K-means clustering. In 2012 7th International Conference on Computer Science & Education (ICCSE), pp. 648-651. IEEE. (2012).
20. Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS), pp. 135-139. IEEE. (2018).
21. Xu, D., & Tian, Y. A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), pp 165-193. (2015).
22. Zakrzewska, D., & Murlewski, J. Clustering algorithms for bank customer segmentation. In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05), pp. 197-202. IEEE. (2005).
23. Wang, X., Zhou, C., Yang, Y., Yang, Y., Ji, T., Wang, J., & Zheng, Y. Electricity Market Customer Segmentation Based on DBSCAN and k-Means:—A Case on Yunnan Electricity Market. In 2020 Asia Energy and Electrical Engineering Symposium (AEEES), pp. 869-874. IEEE. (2020).
24. Project Jupyter, https://jupyter.org/ , last accessed 02.2022
25. Kaufman, L., & Rousseeuw, P. J. Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons. (2009).