



PERFORMANCE COMPARISON OF K-MEANS AND DBSCAN METHODS FOR AIRLINE CUSTOMER SEGMENTATION

Kevser ŞAHİNBAŞ^{1*}


¹Istanbul Medipol University, Faculty of Business Administration and Management Sciences, Department of Management Information Systems, 34810, İstanbul, Türkiye

Abstract: Organizations are now fully embracing ideas such as customer success, customer loyalty, customer experience management and customer satisfaction. The application of these concepts must be based on three pillars of technology, process and people, to ensure that the organization ultimately has satisfied, loyal and successful customers. In today's competitive environment, as in all sectors, gaining great services in the aviation industry can provide a competitive advantage. With this study, it is aimed to help aviation companies to know how their services should meet the needs of customers and to obtain passenger satisfaction. Customer segmentation is widely used, which groups objects according to the similarity difference on each object and provides a high level of homogeneity in the same cluster or a high level of heterogeneity between each group. The aim of this study is to examine airline passenger satisfaction by using data mining methods including K-Means and Density-based spatial clustering of applications with noise (DBSCAN) clustering algorithms to reveal the service quality importance for customer satisfaction. K-Means algorithm achieved slightly better results than DBSCAN algorithm with a Silhouette value of 0.1450671.

Keywords: Clustering, Customer segmentation, K-means, DBSCAN, Data mining, Data management

*Corresponding author: Istanbul Medipol University, Faculty of Business Administration and Management Sciences, Department of Management Information Systems, 34810, İstanbul, Türkiye

E mail: ksahinbas@medipol.edu.tr (K. ŞAHİNBAŞ)

Kevser ŞAHİNBAŞ  <https://orcid.org/0000-0002-8076-3678>

Received: September 05, 2022

Accepted: September 19, 2022

Published: October 01, 2022

Cite as: Şahinbaş K. 2022. Performance comparison of K-Means and DBSCAN methods for airline customer segmentation. BSJ Eng Sci, 5(4): 158-165.

1. Introduction

The aviation, which is one of the most used modes of transportation due to its benefits such as safety, speed and comfort, has become one of the most important sectors in recent years with different factors such as developing technology, increasing aircraft companies and rising demand and decreasing flight prices. Airlines' efforts to attract and retain consumers are increasing competition in the aviation industry. Airlines must realize that price changes alone cannot win the competitive position of price competition for their competitors in the long run (Chang and Yeh, 2002). According to studies, when selecting an airline service, the passenger takes into account both the price and the quality of the service (Jou et al., 2008). Many previous studies have shown that airline service quality is an important factor in passenger satisfaction. (Jiang and Zhang, 2016; Ariffin et al., 2020). The fact that transportation services are shaped according to the wishes and needs of passengers today has caused the companies that provide this service to shift their focus from making profit to providing customer satisfaction. The fact that repeat purchasers are in the most valuable customer category for most businesses highlighted the importance of customer retention. This study examines the full-service airline business model for passenger

satisfaction. Because in today's competitive market, airline firms' ability to provide excellent customer service is a key competitive advantage. When a customer is dissatisfied with the level of service they receive, they are likely to change their mind and choose a different airline for their subsequent travels (Archana and Subha, 2012). Because the frequency of negative events will influence a customer's opinion of the business, the quality of the service provided is considered a significant factor in providing customer satisfaction (Munusamy et al., 2011).

Clustering has tremendous application. DBSCAN is especially useful for large databases and datasets including noisy objects (Cassisi et al., 2013). Besides DBSCAN efficiently discovers random sets of sizes, shapes and numbers in a large dataset (Jahirabadkar and Kulkarni, 2014). Large datasets can be handled by K-means with ease because it is simple to apply and has linear time complexity. K-Means and DBSCAN are used in many fields such as fast clustering of big data (Hanafi and Saadatfar, 2022) and data-intensive applications (Ajin and Kumar, 2016; Saeed et al., 2020), healthcare (Santhanam and Padmavathi, 2015), social networks (Hao et al., 2020), anomaly detection (Chen and Li, 2011), bioinformatics (Masood and Khan, 2015; Bustamam et al., 2017), customer segments of charging stations (Straka and Buzna, 2019), detecting grain inventory



modes (Cui et al., 2021) and so on.

Various conceptual and empirical studies have been conducted to find service quality issues in the aviation industry. Yelmen et al. (2020) provide a customer segmentation analysis using air ticket sales data with the self-organizing map method. In their study, they obtained 15 clusters by grouping customers who had similar sales behaviors. Leon and Martin (2020) used fuzzy logic and fuzzy segmentation to assess airline customer happiness and service quality in the U.S. market. The questionnaire was collected online with 624 respondents using Amazon Mechanical Turk. The findings illustrate that technical quality is not as important to passengers' satisfaction as functional quality, however both factors contribute to overall airline satisfaction. According to a survey, the quality of the food and beverage services is one of the least significant elements affecting how satisfied passengers are with their flights (Deveci and Demirel, 2018). Noviantoro and Huang (2022) proposed a model to investigate airline passenger satisfaction by applying data mining techniques. Their study showed that online/mobile boarding was the most important variables for passenger satisfaction, followed by in-flight wi-fi service second, baggage handling third, and in-flight entertainment fourth. The same constructs are relevant, and they are ranked differently from most to least effective, according to Farooq and Radovic-Markovic (2016). A study is conducted data of more than 5800 airline passengers to provide the segmentation of consumers into business and leisure did not accurately reflect the diversity of customer preferences, which resulted in an incorrect understanding of those choices (Teichert et al., 2008). A preprocessing phase is added to the Fahim (2021) approach before using the k-means algorithm to determine the number of clusters and beginning centers. The DBSCAN algorithm will be used as a preprocessing step in the suggested procedure. His article focuses on DBSCAN and k-means as a result. The final product will be of higher quality because the suggested method will eventually converge to the global minimum. The suggested approach is identical to DBSCAN in that it takes two input parameters and has a time complexity of $O(n \log n)$. In Majhi and Biswal's (2018) study a hybrid clustering approach built on K-means and Ant Lion Optimization was taken into consideration for the best cluster analysis. Based on several performance metrics, the proposed algorithm's performance is contrasted with that of the K-Means, KMeans- Particle Swarm Optimization (PSO), DBSCAN, and Revised DBSCAN clustering algorithms. Eight datasets are used in the experiment, and statistical analysis is done for each of them. The findings demonstrate that in terms of sum of intra-cluster distances and F-measurement, the hybrid of K-Means and Ant Lion Optimization method preferentially outperforms the other four algorithms. In the study of Du (2020) the global parameter selection step of the DBSCAN requires human interaction, and the regional

query procedure is difficult and prone to item loss. Based on maintaining the intrinsic nonlinearity of Internet of Thing (IoT) data, an advanced parameter adaptive and regional query density clustering method is suggested that can efficiently eliminate redundant data in the high-level complex data domain.

In this study, airline passenger satisfaction was investigated by applying data mining methods K-Means and DBSCAN algorithms. This study provides a guide to airlines, and by this way, airlines can use this survey as a reference in developing measures to aid passengers better know how they feel about airline services. This research will help improve airline management and increase the caliber of some services over competitors to give airlines a competitive advantage.

2. Materials and Methods

2.1. Dataset

In this paper, the dataset was collected from the publicly accessible Kaggle website (<https://www.kaggle.com/code/frixinglife/airline-passenger-satisfaction>). The dataset depicts customer satisfaction based on data acquired from 25.976 passenger samples who flew on full-service airlines during a survey conducted at the airport following their arrival in 2015 (Table 1).

Table 1. Variables description

Numeric Variables	
Age	
Flight distance	
Departure delay in minutes	
Arrival delay in minutes	
Categorical variables with satisfaction level (0: not rated; 1-5)	
Inflight wifi service	
Departure/Arrival time convenient	
Ease of Online booking	
Gate location	
Food and drink	
Online boarding	
Seat comfort	
Inflight entertainment	
On-board service	
Leg room service	
Baggage handling	
Checkin service	
Inflight service	
Cleanliness	
Categorical Variables	
Gender	male or female
Customer type	regular or non-regular airline customer
Type of travel	personal or business travel
Class	business, economy, economy plus
Air passenger satisfaction	1- Satisfied, 0- Neutral or dissatisfied

2.2. Clustering Algorithms

Clustering, which is one of the purposes of Data Mining, is frequently implemented in many areas such as pattern recognition and statistical data analysis (Goharnejad et al., 2019). Data mining relies heavily on clustering algorithms, which group objects with related attributes together and organize data in databases into groups or clusters. Data clustering is making strong progress. Clustering analysis has recently great attention in the research of data mining.

2.2.1. K-means

One of the most widely utilized algorithms is the well-known K-means algorithm. It is a clustering method used for classifying data. The main aim of use is to divide the data to be classified into k classes determined by the researcher or clusters in terms of their properties. In the K-Means algorithm (Table 2), k, which represents the number of clusters sought, is a previously known constant and its value does not change until the clustering process is finished (Kaufman and Rosseeauw, 1990). This approach needs a predefined number of clusters, which can be determined using techniques like the elbow method or expert opinion (Han et al., 2011).

Table 2. Algorithm steps of K-means (Hartigan and Wong, 1979)

Algorithm: K-Means

- 1: The value of k, which represents the number of clusters, is read by the algorithm. This value read is given to the algorithm as ready from the outside.
- 2: The cluster center is determined randomly. There are k cluster centers. The first k point can be the center.
- 3: The closeness of the points to the determined centers is calculated.
- 4: According to the calculated values, the points are clustered based on the centers they are close to.
- 5: In order to determine new cluster centers, the averages of the clusters are calculated.
- 6: If there are other points to be clustered, the process is repeated. But if there is no point to be clustered, the process is completed

2.2.2. DBSCAN

DBSCAN algorithm is a density-based spatial clustering algorithm and enables to reveal the neighborhood of data points in two or multidimensional space. This approach can find clusters of different shapes and does not need prior knowledge of the number of clusters (Mahesh, 2020; Hanafi and Saadatfar, 2022). Due to its focus on spatial perspectives, the database is mostly employed in the analysis of spatial data (Ester et al., 1996). Eps, MinPts, core object, direct density accessible point, density accessible point, and density connected point are the fundamental concepts for DBSCAN method (Figure 1; Table 3). Eps and MinPts are taken as input parameters. Starting with any object in the database, it examines every object. If the controlled item has previously been a

part of a cluster, it moves on without being processed to the other object. It uses a Region Query to find the Eps neighbors of the object if it has never been clustered before. This item and its neighbors are referred to as a new cluster if the number of neighbors exceeds MinPts. The process then searches the new region for each neighbor who was previously unclustered. The region query is a part of the cluster if it has more neighbors than MinPts. In the DBSCAN method, the density function of a point neighborhood eps is thus defined as $N_{eps}(p)$, which is depicted in Equation 1.

$$N_{eps}(p) = \text{card}(a \in D \mid \text{dist}(a, b) \leq eps) \quad (1)$$

where D stands for the dataset, b is a random point within D, dist(a,b) is the function measuring the distance between points a and b, and card(*) is the function calculating the number of items in a set.

After that Equation 1 and the specified threshold mpt are used to determine three different types of points.

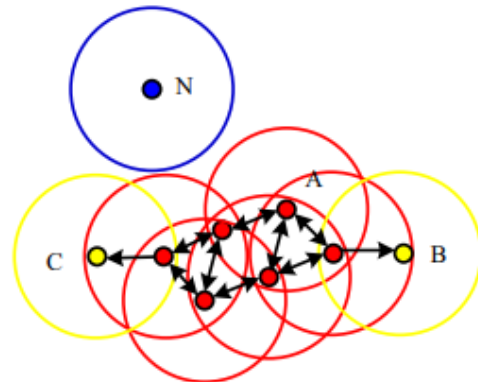


Figure 1. DBSCAN clustering method.

Table 3. Algorithm steps of DBSCAN

Algorithm: DBSCAN (D, eps, mpt)

Inputs: D, dataset of points;
eps, the neighborhood size;
mpt, the density threshold;

- 1: Start
- 2: Arbitrary select a point a 2D;
- 3: Compute the density $N_{eps}(a)$ and decide the types of point a according to mpt;
- 4: If a is a seed point, a cluster is created;
- 5: If a is a boundary point and density is accessible from other points, expand the set and visit the next point;
- 6: Continue processing until all data points have been visited;
- 7: Stop

DBSCAN is superior to other traditional clustering methods in many ways. This technique allows the management of noise patterns in the data and allows the identification of clusters of various shapes. DBSCAN typically gives reliable results. It is also used to minimize the number of computational operations.

2.3. Determination of k Cluster Number by Elbow Method

The elbow method is another simple and meaningful decision solution looking for the optimal K number. The Elbow method, which is calculated by the sum of the squares of the distance of each point from the cluster centers (WCSS: Within Clusters Sum of Square), is another simple and meaningful decision solution that seeks the optimal K number. According to this method, the point where the amount of change in WCCS decreases is the bend point and this bend point represents the best number of k clusters (Ketchen and Shook, 1996).

2.4. System Overview

The research methodology of the study is presented in Figure 2.

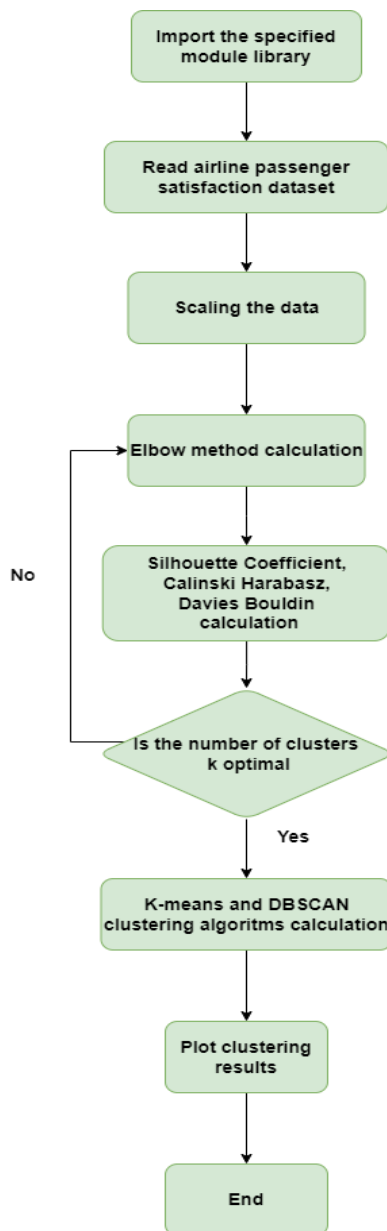


Figure 2. Workflow of cluster analysis.

Dataset was obtained from Kaggle web site. Data is preprocessed, then is scaled. Elbow method is executed to find the optimal k value for K-Means. DBSCAN and K-

Means algorithms are applied for airline passenger data. Performance metrics were found to select the algorithm. Lastly, clustering results are plotted.

2.5. Performance Metrics

In this study, Silhouette Coefficient, Calinski Harabasz, Davies Bouldin and elbow method were used to quantitatively evaluate the cluster result.

2.5.1. Silhouette Coefficient

The silhouette coefficient is a parameter that can be used to evaluate cluster performance (Rousseeuw, 1987). The clustering effect increases as the value increases; the range of this number is between -1 and 1. The solution of the resulting silhouette coefficient is shown in Equation 2.

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1, & a(i) > b(i) \end{cases} \quad (2)$$

Here $a(i)$ specifies the distance between element i and other samples in the same cluster. $b(i)$ is the average distance of the sample i and the color samples in other clusters. $S(i)$ indicates the average value of the entire silhouette coefficient.

2.5.2. Calinski Harabasz

The Calinski-Harabasz (CH) criterion shows a ratio between the within-cluster distribution and the between-cluster distribution. The CH criterion is used to evaluate the compactness and segregation of clusters (Equation 3).

$$ch = \frac{tr(B_k)x(n_E - k)}{tr(W_k)x(k - 1)} \quad (3)$$

$tr(B_k)$: sum of squares within-clusters

$tr(W_k)$: between-cluster sum of squares

The highest ch value represents the best cluster (Caliński and Harabasz, 1974).

2.5.3. Davies Bouldin

The following equation is used to measure the clustering validity with this method, which aims to make the distance between the cluster's minimum and the distance between clusters maximum (Equation 4):

$$db = \frac{1}{k} \sum_{i=1}^k R_i \quad (4)$$

$i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k$. and the following equation determines the maximum comparison ratio between other clusters (Equation 5):

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (4)$$

d_{ij} : distance between centers in clusters

S_i and S_j : the average distance to the centers of the

cluster where the cluster observations are located. Small *db* values indicate good clustering (Davies and Bouldin, 1979).

3. Results and Discussion

In this section, dataset variables are visualized,

correlation matrix results are shown and findings from algorithms are detailed. Figure 3 indicates the visualization of data. In Figure 4, number of satisfied and dissatisfied passengers are shown. 11403 passengers are satisfied, and 14573 passengers are dissatisfied.

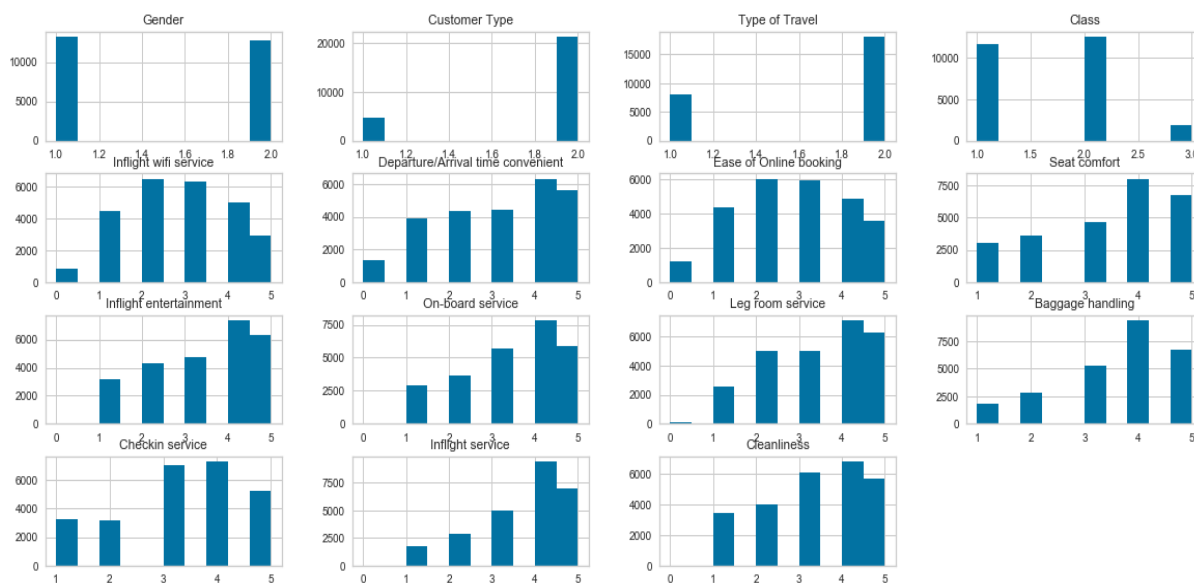


Figure 3. Data visualization.

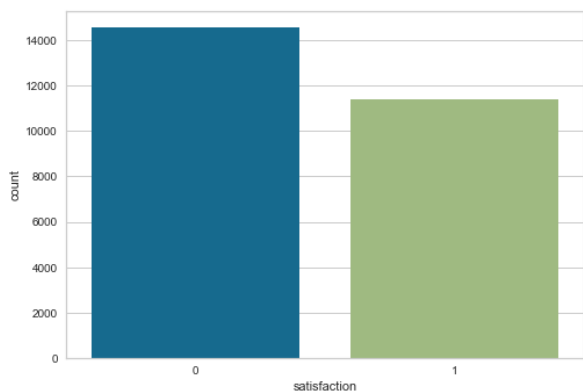


Figure 4. Number of satisfied and dissatisfied passenger.

Figure 5 illustrates correlation between variables. In correlation analysis, the aim is to see in which direction the other variable will change when the value of one variable changes (Taylor, 1990). The findings from Figure 5 indicate that the correlation coefficient between the 'age' attribute and the 'flight distance' variable is 0.099 with 0.000048 p value, indicating a weak and positive relationship between them. It means that while age increased, the flight distance increased slightly as well. Weak and positive relationship is seen between 'flight distance' and 'departure delay in minute' with 0.0034 correlation coefficient and 0.00065 P value. It shows that while flight distance increased, the departure delay in minute increased slightly. Lastly, there is low and negative relationship between 'age' and 'departure delay in minutes' with the value of -0.0043 and 0.00096 P

value. It means that while age increased, the departure delay in minute decreased slightly. P values are less than 0.05 that means our correlation coefficient is statistically significant. The relationship that emerged as a result of the correlation analysis should not be interpreted as a causality relationship.

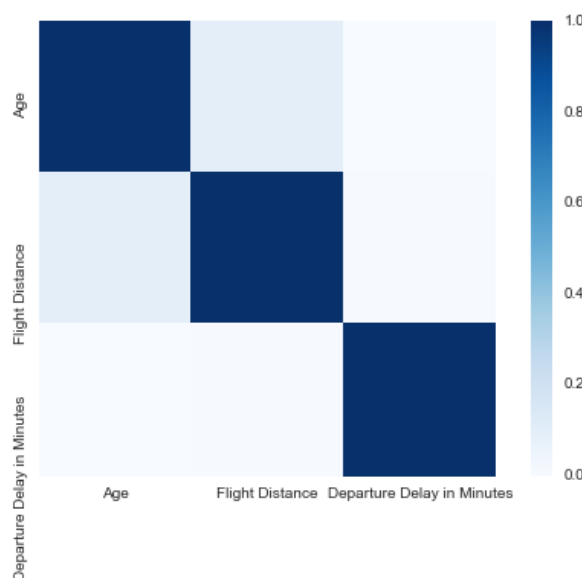


Figure 5. Correlation matrix.

As shown in Figure 6, the optimal k number is 8. Table 4 indicates the metrics values of each clustering algorithm in terms of Silhouette Coefficient, Calinski Harabasz and Davies Bouldin.

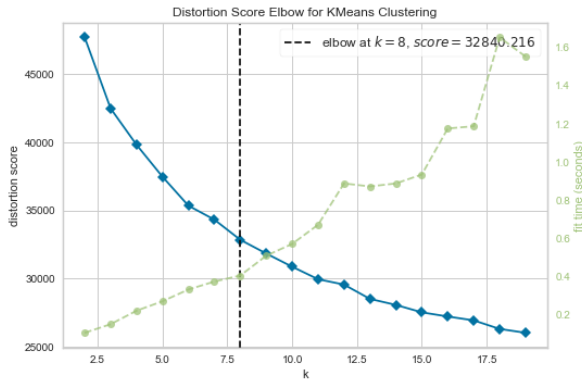


Figure 6. Elbow method.

Table 4. Clustering performance scores

Algorithm	Sihoutte Coefficient	Calinski Harabasz	Davies Bouldin
K-Means	0.145	2854.193	2.078
DBSCAN	0.133	1672.774	2.287

The findings in Table 4 demonstrate that the K-Means algorithm achieve the best performance in terms of Sihoutte Coefficient, Calinski Harabasz and Davies Bouldin with a small margin of 0.145, 2854.193 and 2.078, respectively with a small difference.

Figures 7 and 8 demonstrate the Silhouette analysis of the K-Means and DBSCAN algorithms. The Silhouette range is between -1 and 1 as described in Section 2.5.1. The silhouette coefficient graph indicates that the K-Means clustering algorithm performs better.

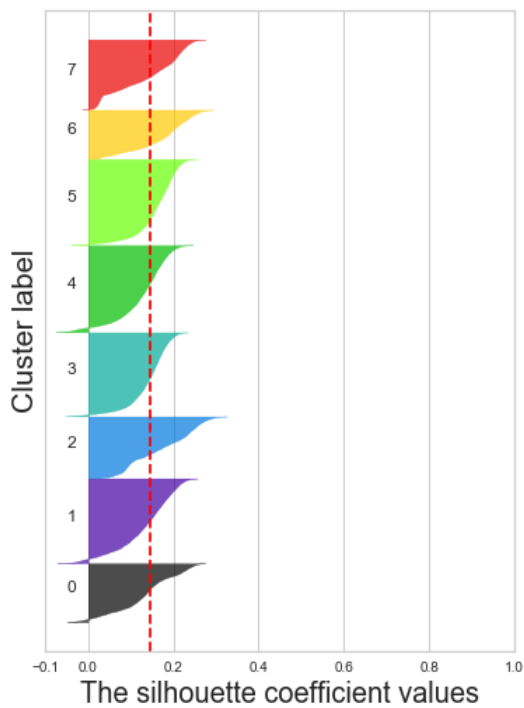


Figure 7. K-means Silhouette coefficient values.

In Figure 9, PCA is applied to the dataset. The number of clusters for K-Means is 8.

The results obtained with the application of K-means and DBSCAN algorithms are compared and the following evaluations can be clarified. These two algorithms generate similar results. The findings from analysis present that K-Means outperformed with a small difference. In fact, both algorithms provide high performance. It can be stated that K-Means and DBSCAN approaches perform quite well to cluster airline customer. In order to achieve good results in the DBSCAN algorithm, it is necessary to change the algorithm with different parameters. For example, eps value can be changed to obtain good results.

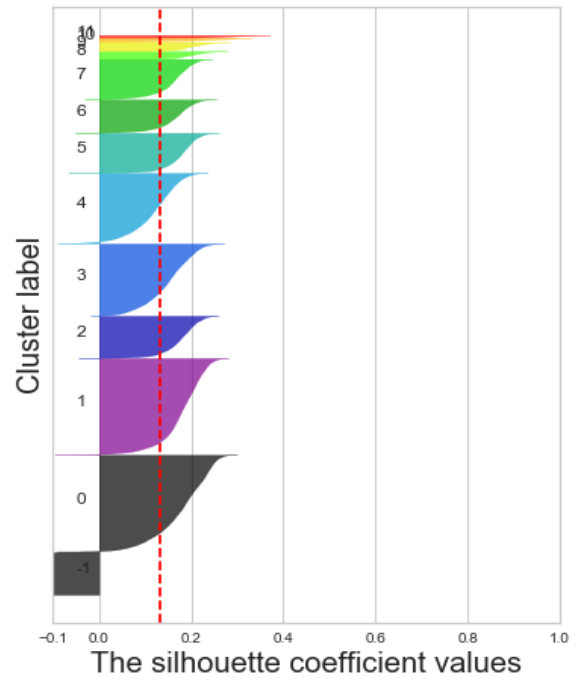


Figure 8. DBSCAN Silhouette coefficient values.

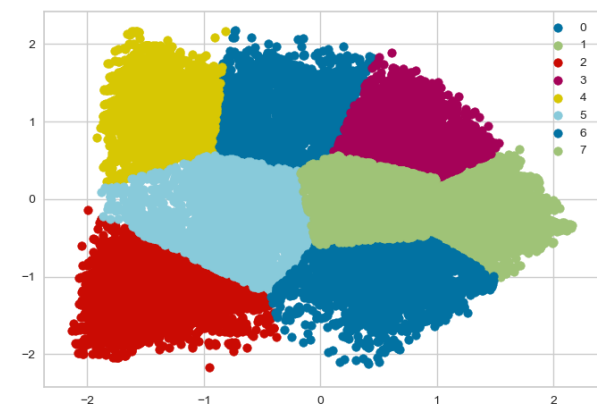


Figure 9. PCA of K-means algorithm.

Table 5. Clustering analysis results

Cluster	Member	Satisfied (25.976)		High score	Low score
1	3899	Yes	2426	Baggage Handling	Gate Location
		No	1473	Inflight wifi service	Checkin service
2	3012	Yes	2940	Inflight service	Gate Location
		No	72	Inflight service	Inflight wifi service
3	3439	Yes	3105	Seat comfort	Gate Location
		No	334	Inflight wifi service	Cleanliness
4	2456	Yes	341	Baggage Handling	Cleanliness
		No	2115	Inflight service	Inflight wifi service
5	3374	Yes	1360	Inflight entertainment	Gate Location
		No	2014	Inflight entertainment	Inflight wifi service
6	2191	Yes	509	Online Boarding	Inflight wifi service
		No	1682	Departure, Arrival time convenient	Baggage handling
7	4635	Yes	1181	Online boarding	Gate Location
		No	3454	Departure, Arrival time convenient	Gate Location
8	2970	Yes	494	Inflight service	Departure/Arrival time convenient
		No	2676	Departure, Arrival time convenient	Ease of Online booking

The definition of columns is as follow: Cluster= cluster number, Member= class member count, Satisfied= satisfied member count and dissatisfied member count, High score= the category name with the highest score from the headings used as the data source, Low score= the category name with the lowest score from the headings used as the data source.

The findings from Table 5 present the members of each cluster, satisfied and dissatisfied number for each cluster. The findings of the study can provide information that can help airline management, professionals and those related to the aviation industry see the most important and unimportant criteria and help them to develop strategies and decision-making in this direction.

4. Conclusion

Ensuring the satisfaction of passenger in air travel, which is one of the most popular transportation modes because it is a faster means of transportation, plays a significant role in increasing the number of competitors day by day.

In this paper, a publicly dataset for airline passenger is used by applying K-Means and DBSCAN clustering algorithms to cluster the airline customer dataset. The reason behind the selection of these two algorithms is that DBSCAN is the first and best-known density-based clustering algorithm and K-means handles huge datasets well since it is straightforward to implement and has linear time complexity.

This study presents a proof of concept on how data analytics can be used in customer segmentation for airline passengers. When the findings of K-Means algorithm and DBSCAN are compared, it is observed that K-Means achieved slightly better results in Sihoutte Coefficient, Calinski Harabasz and Davies Bouldin metrics in all three performances, 0.145, 2854.193 and 2.078, respectively. The study's findings can offer information that airline management, professionals, and other parties with an interest in the aviation industry can use and make decisions about. More different clustering algorithms and the cost implication of each cluster can be considered for further studies.

Author Contributions

All task was done by the single author: K.Ş. (100%). The author reviewed and approved the manuscript.

Conflict of Interest

The author declared that there is no conflict of interest.

References

- Ajin VW, Kumar LD. 2016. Big data and clustering algorithms. International conference on research advances in integrated navigation systems (RAINS) IEEE, 6-7 May 2016, Bangalore, India, pp: 1-5.
- Ariffin Mohd IA, Yajid SA, Johar MGM. 2020. Consumer preferences of airline choice: A comparison of Air Asia and Malaysia Airlines System. *Syst Rev Pharm*, 11(1): 817-826.
- Archana R, Subha MV. 2012. A study on service quality and passenger satisfaction on Indian airlines, *Int J Multidis Res*, 2(2): 50-63.
- Bustamam A, Tasman H, Yuniarti N, Mursidah I. 2017. Application of K-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). *AIP Conf Proc*, 1862(1): 030134.
- Caliński T, Harabasz J. 1974. A dendrite method for cluster analysis. *Commun Stat Theo Meth*, 3(1): 1-27.
- Cassisi C, Ferro A, Giugno R, Pigola G, Pulvirenti, A. 2013. Enhancing density-based clustering: parameter reduction and outlier detection. *Inf Syst*, 38(3): 317-330.
- Chang YH, Yeh CH. 2002. A survey analysis of service quality for domestic airlines. *European J Oper Res*, 139(1): 166-177. DOI: 10.1016/S0377-2217(01)00148-5.
- Chen Z, Li YF. 2011. Anomaly detection based on enhanced DBScan algorithm. *Procedia Eng*, 15: 178-182.
- Cui H, Wu W, Zhang Z, Han F, Liu Z. 2021. Clustering and application of grain temperature statistical parameters based on the DBSCAN algorithm. *J Stored Prod Res*, 93: 101819.
- Deveci M, Demirel NÇ. 2018. A survey of the literature on airline crew scheduling. *Eng App Artif Intel*, 74: 54-69.
- Ester M, Kriegel HP, Sander J, Xu X. 1996. A density based algorithm for discovering clusters in large spatial databases.

- Int. Conference of Knowledge Discovery and Data Mining (KDD'96), Portland, USA, pp: 226-231.
- Davies DL, Bouldin DW. 1979. A cluster separation measure. *IEEE Transact Pattern Analysis Machine Intel*, 2: 224-227.
- Du Z. 2020. Energy analysis of Internet of things data mining algorithm for smart green communication networks. *Comp Commun*, 152: 223-231.
- Fahim A. 2021. K and starting means for k-means algorithm. *J Comput Sci*, 55: 101445.
- Farooq MS, Radovic-Markovic M. 2016. Modeling entrepreneurial education and entrepreneurial skills as antecedents of intention towards entrepreneurial behaviour in single mothers: a PLS-SEM approach. *ETCTFP*, 2016: 198-216.
- Goharnejad H, Shamsai A, Zakeri Niri M. 2019. Prediction of sea level rise in the south of Iran coastline: evaluation of climate change impacts. *Water Res Eng*, 12(42): 1-17.
- Jiang H, Zhang Y. 2016. An investigation of service quality, customer satisfaction and loyalty in China's airline market. *J Air Trans Manag*, 57: 80-88.
- Han J, Pei J, Kamber M. 2011. *Data mining: concepts and techniques*. Elsevier, New York, US, pp: 703.
- Hao F, Zhang J, Duan Z, Zhao L, Guo L, Park DS. 2020. Urban area function zoning based on user relationships in location-based social networks. *IEEE Access*, 8: 23487-23495.
- Hanafi N, Saadatfar H. 2022. A fast DBSCAN algorithm for big data based on efficient density calculation. *Expert Sys App*, 203: 117501.
- Hartigan JA, Wong MA. 1979. Algorithm AS 136: A k-means clustering algorithm. *J Royal Stat Soc Series c*, 28(1): 100-108.
- Jou RC, Lam SH, Hensher DA, Chen CC, Kuo CW. 2008. The effect of service quality and price on international airline competition. *Transport Res Part E*, 44(4): 580-592.
- Jahirabadkar S, Kulkarni P. 2014. Algorithm to determine ϵ -distance parameter in density based clustering. *Expert Sys App*, 41(6): 2939-2946.
- Kaufman L, Rosseeuw PJ. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons Inc., New York, US, pp: 335.
- Leon S, Martín JC. 2020. A fuzzy segmentation analysis of airline passengers in the US based on service satisfaction. *Res Transport Busin Manag*, 37: 100550.
- Ketchen DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strat Manag J*, 17(6): 441-458.
- Masood MA, Khan MNA. 2015. Clustering techniques in bioinformatics. *IJ Modern Educ Comp Sci*, 1: 38-46.
- Munusamy J, Chelliah S, Pandian S. 2011. Customer satisfaction delivery in airline industry in Malaysia: a case of low cost carrier. *Australian J Basic App Sci*, 5(11): 718-723.
- Noviantoro T, Huang JP. 2022. Investigating airline passenger satisfaction: Data mining method. *Res Transport Busin Manag*, 43: 100726.
- Majhi SK, Biswal S. 2018. Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. *Karbala Int J Modern Sci*, 4(4): 347-360.
- Mahesh B. 2020. Machine learning algorithms-a review. *Int J Sci Res*, 9: 381-386.
- Straka M, Buzna LU. 2019. Clustering algorithms applied to usage related segments of electric vehicle charging stations. *Transport Res Proc*, 40: 1576-1582.
- Teichert T, Shehu E, von Wartburg I. 2008. Customer segmentation revisited: The case of the airline industry. *Transport Res Part A*, 42(1): 227-242.
- Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J Comput App Math*, 20: 53-65.
- Saeed MM, Al Aghbari Z, Alsharidah M. 2020. Big data clustering techniques based on spark: a literature review. *Peer J Comp Sci*, 6: e321.
- Santhanam T, Padmavathi MS. 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Comp Sci*, 47: 76-83.
- Taylor R. 1990. Interpretation of the correlation coefficient: a basic review. *J Diag Medic Sonograp*, 6(1): 35-39.
- Yelmen İ, Üstebay S, Zontul M. 2020. Customer segmentation based on self-organizing maps: a case study on airline passengers. *J Aeronautics Space Technol*, 13(2): 227-233.