JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 6, ISSUE 2, DECEMBER-2023
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

OPTIMIZATION OF K VALUE IN CLUSTERING USING SILHOUETTE SCORE (CASE STUDY: MALL CUSTOMERS DATA)

Heti Mulyani[1], Ricak Agus Setiawan[2], Halimil Fathi[3]

[1,2,3]Politeknik Enjinering Indorama, Purwakarta, Indonesia
Heti.mulyani@pei.ac.id

***Abstract*--Clustering is an important phase in data mining. The grouping method commonly used in data mining concepts is using K-Means. Choosing the best value of k in the k-means algorithm can be difficult. In this study the technique used to determine the value of k is the silhouette score. Then, to evaluate the k-means model uses the Davies Bouldin Index (DBI) technique. The best DBI value is close to 0. The parameters used are total consumer income and spending. Based on the results of this study it can be concluded that the silhouette score method can provide a k value with optimal results. For mall customer data of 200 data, the most optimal silhouette score is obtained at K = 5 with a DBI = 0.57.**

***Keywords*: Cluster; Davies Bouldin Index; K-Means; Market; Silhouette Score**

I. INTRODUCTION

In line with the rapid development of technology, data growth has also experienced a very high increase. Every day transactions occur on various platforms such as e-commerce, hospitals, government institutions, etc. This has led to a flood of data in various fields. The data will not provide useful information if it is not processed. Therefore, the concept of data mining emerged, where this concept provides a solution so that data can be processed so as to produce useful information and can be used as a reference for stakeholders to refer to in making policies. One of the widely used data mining techniques is clustering, i.e. grouping of data based on their similarity [1]. One of the most widely used cluster methods is K-Means clustering [2]. Selecting the number of clusters in a clustering algorithm, e.g. choosing the best value of k in the various k-means algorithms can be difficult [3].

The K-Means algorithm is one of the non-hierarchical data clustering methods that partitions the existing data into two or more groups [4]. The quality of data clustering results depends on the input of the number of clusters or the value of K. To evaluate the clustering results, use the Davies Bouldin Index (DBI). The DBI is a method for calculating the average distance ratio within a cluster and calculating the average distance between clusters and their closest data. [5].

Some researches related to calculating the optimization of the k value has been carried out by several previous researchers, including optimal clustering with the elbow method for clustering traffic accident data in the city of Semarang, in this study the optimal k value was obtained using the elbow method, where the optimal K according to the elbow is as many as 3 pieces on the research data [4]. The next is research with the title Clustering Mall Visitors Using the K-Means Method and Particle Swarm Optimization. This study obtained mall customer clusters with k-means and evaluated with the Davies Bouldin Index with the results [6]. The next study, that is entitled Comparative Analysis of the Elbow and Silhouette Methods on the K-Medoids Clustering Algorithm in Grouping Balinese Craft Production, shows that the results of grouping data using the silhouette technique have a smaller DBI value, so the silhouette score technique can produce better clusters [7]. The next research is entitled the Application of K-Means Clustering Algorithm for Grouping Road Construction at the Public Works and Spatial Planning Office. In this study, 5 groups of data were obtained and evaluated using the Davies Bouldin index [8]. The next research is entitled The Use of the K-Means Algorithm in the Tourist Area Cluster Mapping Application. This research shows the optimal cluster using the silhouette score method by dividing the tourist

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

area into 2 groups [9]. Based on the background above, in this study, data analysis will be carried out with case studies of mall customers to obtain optimal K values and high DBI values.

II. METHOD

The data mining method used in this study is CRISP-DM. CRISP-DM has 6 steps, namely: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The framework of this research can be seen in Fig. 1.

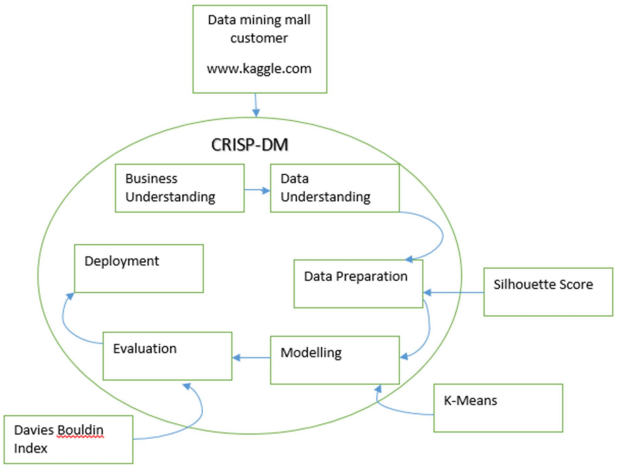

Fig. 1. Research Framework

The following is the explanation of the CRISP-DM Framework [10].

A. Business Understanding

Business understanding is the stage of understanding what to be achieved from a business perspective. Determining project objectives and needs in detail within the scope of the business or research unit as a whole. This stage also translates the objectives and constraints into a formula for the data mining problem. This stage requires mastery of two different parts, namely understanding the business process including the regulations that govern it and understanding how to process it.

B. Data Understanding

Data understanding is the stage to identify and to collect relevant data for the project; including what data is available, how it is collected, and how it can be used to achieve business goals. The data in this study were taken from the Kaggle.com source with the file name mall_customer.csv. The number of datasets is 200 and has 5 columns consisting of customer ID, gender, age, annual_income in $, and spending score with a score of 1-100.

C. Data Preparation

Data preparation is the stage of cleaning, integrating, and preparing data for analysis. This process includes processing missing data, dealing with outliers, and changing the data format if necessary. For data preparation in this study, in addition to clean the data, it also calculates the silhouette score to determine the optimal K value before the modeling process is carried out. The silhouette score steps are as follows:

1. Calculate Distance Between Objects:
   For each object in the dataset, calculate the average distance between that object and all other objects in the same cluster, called the average intra-cluster distance (a).
   Also calculate the average distance between that object and all objects in the closest different cluster or called the inter-cluster average distance (b).
2. Calculate the silhouette value for each object. For each object in the dataset, calculate the Silhouette value as follows:
   $$S(object) = (b - a)/\max(a, b) \quad (1)$$
3. Calculate the Global Silhouette Score Value:
   $$SS = (\sum S(object))/number\_objects \quad (2)$$

D. Modeling

The modeling stage is the process of developing data analysis and evaluating using specific models, such as techniques like regression, classification, clustering, or other machine learning models [11]. The steps for the K-Means method can be described as follows [12]:

1. Specify the number of groups (K).
2. Calculate the group center or centroid value of the data in each group.
   $$x = \frac{1}{M} = \sum_{j=1}^{M} Xj \quad (3)$$
   where M is the amount of data in a group.
3. Allocate each data to nearest centroid/average. using the Euclidian as follow:
   $$d = \sqrt{\sum xi^2 - yi^2} \quad (4)$$
4. Group data based on the closest distance.
5. Repeat step 2 until the data does not change.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

E. Evaluation

The evaluation in this research uses Davies Bouldin Index method. Evaluation using DBI has an internal cluster evaluation scheme, in which the quality of the cluster results are seen from the quantity and closeness between the cluster result data. The way to measure DBI is by looking at the both inter-cluster and intra-cluster distance. If the distance between clusters is farther, then it is said better. Meanwhile, the closer the intra-cluster distance, then it is also said better [13]. The steps for calculating Davies Bouldin Index are as follows [14]:

1. The formula for the sum of squares in clusters (SSW) as a cohesion metric in cluster i is as follows.
$$SSWi = \frac{1}{mi}\sum_{j=1}^{mi} d(xj, ci) \quad (5)$$
where mi is the number of data in cluster i and $d\ (x, c)$ is the distance of data x to centroid c.
2. Sum of squares between clusters (SSB) by measuring the distance between centroids ci and cj as in the following equation:
$$SSBi, j = d(ci, cj \quad (6)$$
where (ci, cj) is the distance between centroid ci and centroid cj.
3. Calculating Rij, which is a measure of the ratio of how good the comparison value is between cluster I and cluster j. The value is obtained from the cohesion and separation components. A good cluster is one that has the smallest possible cohesion and the greatest possible separation.
$$Rij = \frac{SSWi+SSW}{SSBi,j} \quad (7)$$
4. Calculating Davies Bouldin Index (DBI).
$$DBI = \frac{1}{K}\sum_{j=1}^{K} \max(Rij) \quad (8)$$

F. Deployment

Model implementation may involve application development, integration with existing systems, or other necessary steps. Evaluation of the K value test in this study, carried out as many as 6 experiments, namely k with a value of 2 to 6.

III. Result and Discussion

This section will explain the results and discussion of customer data clusters as well as the results of testing the optimization of the K value. The results and discussion of this study are as follows:

A. Business Understanding

Currently, the proliferation of online stores has caused the number of mall visitors to drop drastically. Therefore, data grouping is needed to find out the cluster of mall visitors, to find out the group of people who still like to visit the mall and to conduct the right target market to promote products and discounts that apply to consumers in order to increase sales.

B. Data Understanding

At this stage, the data reading process is carried out using Google Colab. The goal is to find out the details of the data, attributes, and the amount of data. Table I shows the dataset that has been displayed on Google Colab.

TABLE I
Mall Customer Dataset

Customer ID	Gender	Age	Annual Income	Spending Score
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
…	…	…	…	…
200	Female	35	19	99

The K value generated by the Silhouette Score method will be a reference in grouping mall customer data in Table I. The determination of the value will be calculated first so that the cluster results are more optimal.

C. Data Preparation

At this stage, data preparation is carried out, starting with cleaning the data so that it is ready to be processed. The data attributes used for the clustering process use the annual income and spending score; therefore customer ID, age, and gender must be removed from the dataset. Table II shows the view of the dataset that has been cleaned, while Fig. 2 shows the data plot.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

TABLE II
Dataset Visitors by Annual Income and Spending Score

CustomerID	Annual_Income	Spending_Score
1	15	39
2	15	81
3	16	6
4	16	77
5	17	40
6	17	76
7	18	6
8	18	94
9	19	3
10	19	72
…	…	…
200	19	99

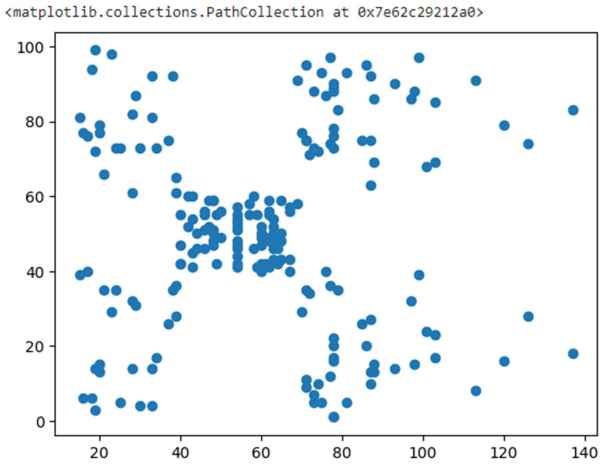

Fig. 2. Data Distribution

The process for finding the optimal K value with the Silhouette Score in Python is shown in Fig. 3. Based on it, the silhouette score results are obtained in Fig. 4. The results of the silhouette score calculation is illustrated in Fig. 5.

```
sil=[]
#start the cluster range from 2
range_n_clusters = range(2,10)

for n_clusters in range_n_clusters:
    clusterer = KMeans(n_clusters=n_clusters,random_state=14)
    cluster_labels = clusterer.fit_predict(new1)
    silhouette_avg = silhouette_score(new1, cluster_labels)
    print("For n_clusters =", n_clusters,
     "The average silhouette_score is :", silhouette_avg)
    sil.append(silhouette_avg)

plt.plot(range_n_clusters,sil)
plt.xlabel('Values of K')
plt.ylabel('Silhouette score')
plt.title('Silhouette analysis For Optimal k')
plt.show()
```

Fig. 3. Finding the Optimal K Value

```
For n_clusters = 2 The average silhouette_score is : 0.2968969162503008
For n_clusters = 3 The average silhouette_score is : 0.46761358158775435
For n_clusters = 4 The average silhouette_score is : 0.4931963109249047
For n_clusters = 5 The average silhouette_score is : 0.553931997444648
For n_clusters = 6 The average silhouette_score is : 0.53976103063432
For n_clusters = 7 The average silhouette_score is : 0.5270287298101395
For n_clusters = 8 The average silhouette_score is : 0.4558493609925033
For n_clusters = 9 The average silhouette_score is : 0.44966289417722194
```

Fig. 4. Result Silhouette Score

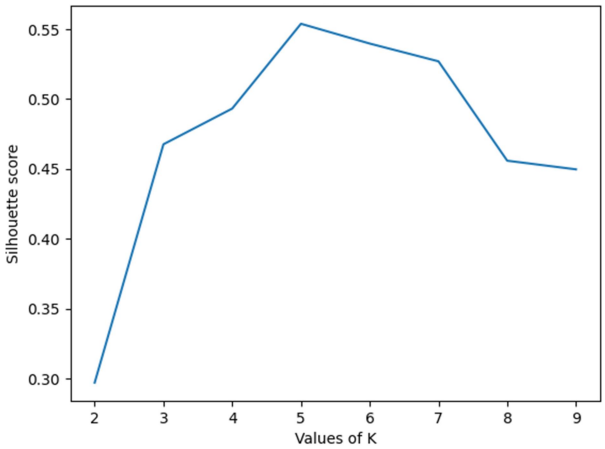

Fig. 5. Graphic Silhouette Score

Fig. 5 shows the highest graph when K = 5, with a silhouette score of 0.553931997444648. This shows that the most optimal K for the mall customer dataset is 5.

D. Modeling

At the modeling stage, the data will be grouped using the optimal K value that has already been obtained by the Silhouette Score technique. The source code to determine the K value in Python in Fig. 6. The results of customer mall data clustering can be seen in Fig. 7.

```
km = KMeans(n_clusters=5)
km
```

```
KMeans
KMeans(n_clusters=5)
```

```
y_predicted = km.fit_predict(df_scaled[['Annual_Income','Spending_Score']])
y_predicted
new1['Kelompok'] = y_predicted
```

Fig. 6. Source Code Clustering

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

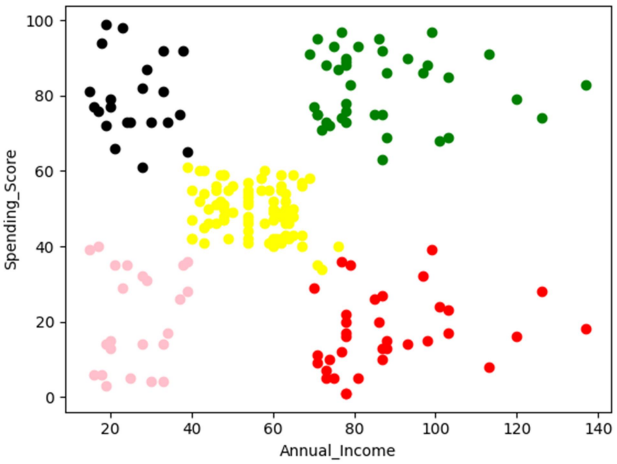

Fig. 7. Result of Clustering

Fig. 7 shows that there are 5 groups, and each group color is explained as follows:

1. The first group, that is green, shows customers who have high income and expenses. Consumers in this group are the main target of the mall for offering products.
2. The second group in red shows customers who have high income but low expenses. Consumers with these characteristics are mall targets for offering discounted products to make them more interested in shopping.
3. The third group in yellow shows customers who have balanced expenses and income.
4. The fourth black group shows customers who have low income but have high expenses. This kind of consumers tend to be consumptive. They do not consider the income earned. These consumers are good targets for offering products and discounts.
5. The fifth group in pink shows customers who have low income and low expenses. This kind of customer characteristics should not be a priority.

E. Evaluation

The evaluation process was carried out using the Davies Bouldin Index (DBI) technique. The closer the DBI value is to 0, the better the cluster results [15]. The following is a comparison of cluster results using K values from 2 to 6. K value testing is done using Google Colab, with the command in Fig. 8. Table III is a table of results from the comparison of Davies Bouldin Index values for variations in K values.

```
km = KMeans(n_clusters=2)
km
```

```
KMeans(n_clusters=2)
```

```
from sklearn.metrics import davies_bouldin_score

db_index = davies_bouldin_score(new1, y_predicted)
print(db_index)
```

```
1.2324797540771337
```

Fig. 8. Test DBI using K=2

```
km = KMeans(n_clusters=3)
km
```

```
KMeans(n_clusters=3)
```

```
from sklearn.metrics import davies_bouldin_score

db_index = davies_bouldin_score(new1, y_predicted)
print(db_index)
```

```
0.7148942288806714
```

Fig. 9. Test DBI using K=3

```
km = KMeans(n_clusters=4)
km
```

```
KMeans(n_clusters=4)
```

```
from sklearn.metrics import davies_bouldin_score

db_index = davies_bouldin_score(new1, y_predicted)
print(db_index)
```

```
0.7092573763454366
```

Fig. 10. Test DBI using K=4

```
km = KMeans(n_clusters=5)
km
```

```
KMeans(n_clusters=5)
```

```
from sklearn.metrics import davies_bouldin_score

db_index = davies_bouldin_score(new1, y_predicted)
print(db_index)
```

```
0.5710238111525201
```

Fig. 11. Test DBI using K=5

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5243. SUBMITTED: SEPTEMBER 11, 2023; REVISED: OCTOBER 23, 2023; ACCEPTED: NOVEMBER 9, 2023

```
km = KMeans(n_clusters=6)
km

        KMeans
KMeans(n_clusters=6)

from sklearn.metrics import davies_bouldin_score

db_index = davies_bouldin_score(new1, y_predicted)
print(db_index)

0.6570245967994689
```

Fig. 12. Test DBI using K=6

TABLE III
Result Value DBI based on K Value

Number of K	Value of DBI
2	1,232479754
3	0,714894229
4	0,709257376
5	0,571023811
6	0,657024596

Table III shows that the lowest DBI value is obtained when K = 5, which is 0.571023811. This shows that the silhouette score calculation results are able to produce the best K value, so that the cluster results are more optimal.

Based on Fig. 8 to 12, DBI testing shows that the minimum value is obtained when K=5. These results are in accordance with the results of the Silhouette Score calculation which suggests the best K value is 5.

IV. Conclusion

Based on the results of the research above, it can be concluded that determining the value of K in K-Means using the Silhouette Score method produces the smallest DBI value, so that the cluster results are more optimal. For the next study, the results of the Silhouette Score method can be applied to different datasets, with more data.

V. References

[1] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[2] M. Orisa, "Optimasi Cluster pada Algoritma K-Means," *Pros. SENIATI*, vol. 6, no. 2, pp. 430–437, 2022, doi: 10.36040/seniati.v6i2.5034.

[3] A. Z. Faridee and V. P. Janeja, "Cluster Quality Analysis Using Silhouette Score. *J. o*, vol. 15, no. 2, pp. 7–22, 2020.

[4] V. A. Ekasetya and A. Jananto, "Klusterisasi Optimal Dengan Elbow Method Untuk Pengelompokan Data Kecelakaan Lalu Lintas Di Kota Semarang," *J. Din. Inform.*, vol. 12, no. 1, pp. 20–28, 2020, doi: 10.35315/informatika.v12i1.8159.

[5] Y. Sopyan, A. D. Lesmana, and C. Juliane, "Analisis Algoritma K-Means dan Davies Bouldin Index dalam Mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1464–1470, 2022, doi: 10.47065/bits.v4i3.2697.

[6] T. M. Dista and F. F. Abdulloh, "Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1339, 2022, doi: 10.30865/mib.v6i3.4172.

[7] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompok-an Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.

[8] D. Kurniadi, Y. H. Agustin, H. I. N. Akbar, and I. Farida, "Penerapan Algoritma k-Means Clustering untuk Pengelompokan Pembangunan Jalan pada Dinas Pekerjaan Umum dan Penataan Ruang," *Aiti*, vol. 20, no. 1, pp. 64–77, 2023, doi: 10.24246/aiti.v20i1.64-77.

[9] L. F. Marini and C. D. Suhendra, "Penggunaan Algoritma K-Means Pada Aplikasi Pemetaaan Klaster Daerah Pariwisata," *J. Media Inform. Budidarma* , vol. 7, no. 2, pp. 707–713, 2023, doi: 10.30865/mib.-v7i2.5558.

[10] P. Chapman *et al.*, "Step-by-step Data Mining Guide," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72%0Ahttp://www.crisp-dm.org/CRISPWP-0800.pdf

[11] A. Rohmah *et al.*, "Analisis Penentuan Hambatan Pembelajaran Daring Dengan Algoritma K-Means 1," *J. Rekayasa Teknol. Nusa Putra*, vol. 4, no. 2, pp. 30–35, 2018.

[12] W. M. P. Dhuhita, "Clustering Metode K-Means Untuk Menentukan Status Gizi Balita," *J. Inform.*, vol. 15, no. 2, pp. 160–174, 2015.

[13] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 918, 2019, doi: 10.30645/senaris.v1i0.100.

[14] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 1, p. 48, 2020, doi: 10.26418/jp.v6i1.37606.

[15] S. Butsianto and N. Saepudin, "Penerapan Data Mining Terhadap Minar Siswa dalam Mata Pelajaran Matematika dengan metode K-Means" *Angew. Chemie Int. Ed. 6(11), 951–952.*, vol. 3, no. 1, pp. 10–27, 2018,