Detecting Data Outliers with Machine Learning

Article *in* Al-Salam Journal for Engineering and Technology · May 2023

DOI: 10.55145/ajest.2023.02.02.018

3 authors, including:

Ghalia Nasserddine
Rafik Hariri University
31 PUBLICATIONS **188** CITATIONS
SEE PROFILE

Joumana Younis
44 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Al-Salam Journal for Engineering and Technology
Journal Homepage: http://journal.alsalam.edu.iq/index.php/ajest
e-ISSN: 2790-4822p-ISSN: 2958-0862

Detecting Data Outliers with Machine Learning

Thaer Falahi[1], Ghalia Nassreddine[1*], Joumana Younis[1]

[1]Faculty of Business, Jinan University of Lebanon, Tripoli, Lebanon.

*Corresponding Author: Ghalia Nassreddine

DOI: https://doi.org/10.55145/ajest.2023.02.02.018
Received February 2023; Accepted April 2023; Available online May 2023

ABSTRACT: Anomalies are instances or collections of data that occur very rarely in a dataset and where features differ significantly from most of the data. In the age of technology, data is widely used in all sectors. Thus, anomalies in the data may produce problems if they are not detected. Anomaly detection involves examining specific data points and detecting rare occurrences that seem suspicious because they are different from the established pattern of behaviors. In this study, an approach to anomaly detection is built using a machine learning technique. The clustering distance-based method (k_means) is adopted. First, the anomaly existence is tested using p_value. After that, the anomaly data is detected using the clustering method. The proposed method was tested using real data collected from Kaggle. The results show the good performance of the k_means algorithm in the detection of outlier data.

Keywords: Outlier detection, machine learning, clustering, K-means, performance

1. INTRODUCTION

In the last decade, our world has witnessed a major technological revolution. Almost all trading and commercial matters are done through an electronic device. The traded data is saved for further use. For this reason, reliance on electronic data has increased. Indeed, it has become the main factor in the establishment and development of companies. It has also become a key factor in making basic decisions in most commercial businesses. However, this data often suffer from several problems, such as missing data, outliers, and anomalous data [1].

Therefore, real data is never as good as we would like it to be. It suffers from several problems that can affect its interpretation and processing. Systems and models that are created based on available data, like decision systems, are affected by their quality. Corruption or damage can be caused by several factors, including human errors, the inherent variability of the field, rounding errors, transcription errors, and tool glitches. Anomalous data can also be the result of rare but valid behavior, which often leads to interesting results. As we mentioned, the importance of detecting anomalies comes before processing data for further use [2].

This process is known as "outlier detection" in data analysis. It is the process of identifying rare elements, events, or observations that deviate significantly from the majority of the data and do not conform to the determinant's concept of the normal behavior of this data [3].

With a huge boom in the field of technology that the world has witnessed recently, the number of technologies that may contribute to addressing the problem of data anomalies has greatly increased.

Data is a collection of numbers, letters, or words that special computer programs process to obtain the necessary information. Data is an important resource in our time. Information and data have a close relationship in decision-making and constitute an essential resource for companies in our current era. For the company to maintain its survival, it must collect, purify, store, and use a huge amount of data and information. Data quality issues can stem from data redundancy, unstructured data, anomalous data, incomplete data, different data formats, or difficult data access [4].

*Corresponding author: gnassred@gmail.com

Over the past decade, we have seen an enormous amount of research effort dedicated to designing techniques for detecting anomalies in the data. Artificial intelligence techniques are effective in terms of efficiency and accuracy in detecting extremism and deviations in data [5].

There are several techniques for detecting outliers in data, such as cluster analysis, Bayesian networks, and hidden Markov. One of the most important of these techniques is cluster analysis [6-8].

Cluster analysis is a technique of grouping data in such a way that the data in the same set (called a cluster) has similar features. This technique is a sub-part of data mining. It is a well-known technique for a statistical data analyst. Cluster analysis is used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics, and machine learning.

In this paper, a distance-based clustering algorithm is used to cluster the available data. A real dataset extracted from Kaggle.com will be used to test the performance of this technique. Silhouette and Dunn's scores are used to evaluate the clustering performance.

This paper is organized as follows: First, the outlier detection problem is described in Section 2, followed by reviewing the clustering techniques. Section 3 presents a literature review. Section 4 illustrates the suggested method and performance parameters, followed by results analysis in Section 5. In Section 6, the authors offer a list of recommendations to conclude this paperwork.

2. OUTLIER DETECTION

Detecting outliers or anomalies in data can be defined as identifying items that are unexpected or that differ significantly from the norm. It is applied to unclassified data by data scientists. Any type of outlier detection depends on two basic assumptions [9].

- Data outliers are rare.
- The features of outliers' data are very different from those in normal data.

Outlier data is usually associated with a problem or rare events, such as hacking, bank fraud, equipment failure, structural defects, infrastructure failures, or scripting errors. For this reason, identifying actual anomalies is essential from a business perspective.

In this section, the process of detecting outliers in data will be described. First, outlier data will be defined. Then, the techniques used to find outliers will be represented. Later, the outlier types will be illustrated.

2.1 OUTLIERS

In data science, outliers are referred to values that do not conform to the expected pattern for other values in the data set (See Fig. 1). Outliers are elements or records whose characteristics differ from other data. They are usually not close to the clustering of other elements [10].

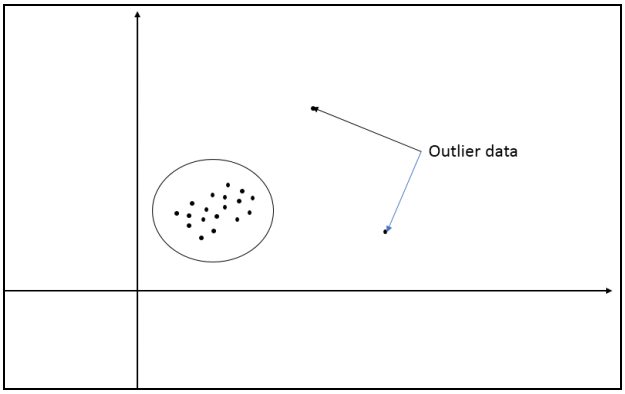

FIGURE 1. - Outliers are shown to be far from the neighboring data points

Outliers are formed as a result of errors [10], such as:

- Measurement error.
- Data distribution error.
- Human error.
- Sensor or machine error.

These outliers can affect the analysis method if they are not detected and managed. In many cases, the analysis of these outliers has a high value of importance like in Banks where outliers may present fraud or theft transactions [11].

There are several reasons for the appearance of outliers in the available data [12-13]:

- Human reasons: entering wrong data, such as entering the age of 500 instead of 50 when entering the age of the new employee in the Human Resources Department.

- Technical reasons: such as during data collection from sensors, a malfunction in these devices can occur. This may lead to wrong and outliers data (the weather thermometer gives a temperature of 100 degrees).
- Intentional Errors: In this case, dummy outliers are intentionally introduced to test fraud detection methods.
- Errors as a result of sampling, such as extracting or mixing data from the wrong or diverse sources.
- Natural errors, such as developments in the data (mutations).

2.2 PROCESS OF FINDING OUTLIERS

Several techniques are used to find outliers. In this paper, the authors present two methods [14-15]: graphs and interquartile range that are described in details.

Method 1: Graphs as histograms, box plot, or scatter plot are used, as shown in Fig. 2, 3, and 4. For example, the red line on the far left in Fig. 2 indicates the presence of outliers.

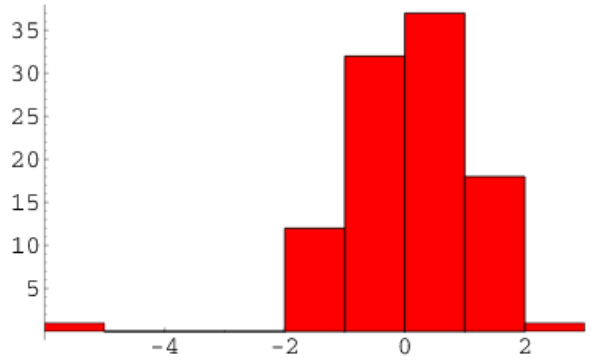

FIGURE 2. - Outliers are shown as histograms

Figure 3 shows the weight of a primary school student's backpack (each dot represents a student). We notice that two points (Brad and Sharon) do not fit the pattern well. Sharon can be considered an anomaly because she carries a much heavier backpack. As for Brad, he is an outlier because he is holding a much lighter backpack than one would expect.

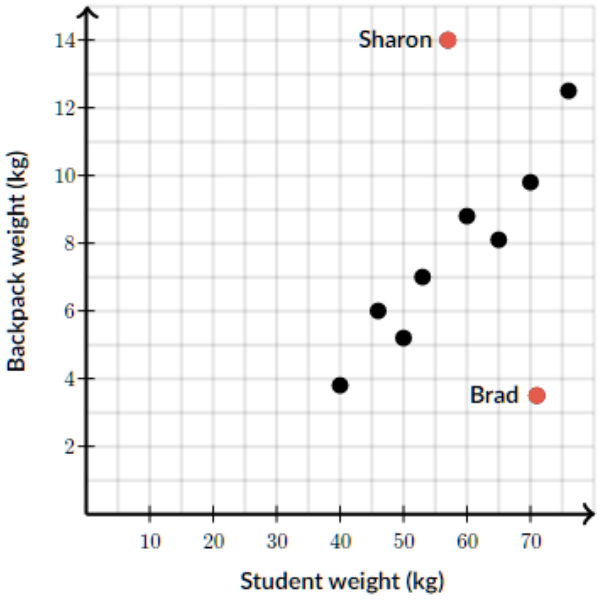

FIGURE 3. - Outliers are shown as scatter plot

A box plot is a way to depict groups of numerical data by their quadrants (see Fig. 4). It has extending vertical lines that represent the maximum and minimum values [16].

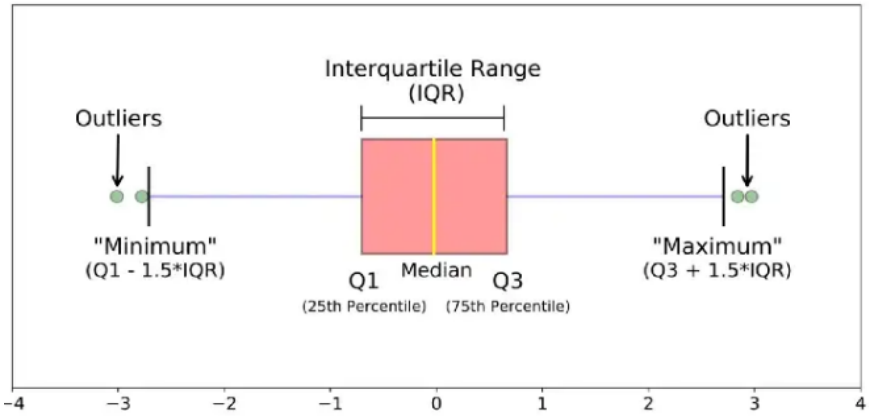

FIGURE 4. - Outliers are shown as box plot

Method 2: Interquartile range (IQR): The authors present an example using the Python programming language. In this example, a measurement of temperature in Iraq will be presented using:

$$T = [25, 30, 35, 4, 75, 18, 45, 35, 25, 70, 12, 15]$$

The temperature is normal, between 0 and 40 degrees. Figure 4 shows the results of detecting outliers' data using the interquartile range (IQR). According to this picture, there are three extremes: [75, 45, 70]. Figure 5 shows the code in Python for illustrating IQR.

```
[12]
    T = [25,30,35,4,75,18,45,35,25,70,12,15]
    Q1 = np.percentile(T, 10, interpolation = 'midpoint')
    Q2 = np.percentile(T, 20, interpolation = 'midpoint')
    Q3 = np.percentile(T, 40, interpolation = 'midpoint')

    IQR = Q3 - Q1
    print('Interquartile range is', IQR)
    low_lim = Q1 - 1.5 * IQR
    up_lim = Q3 + 1.5 * IQR
    print('low_limit is', low_lim)
    print('up_limit is', up_lim)
    outlier =[]
    for x in T:
        if ((x> up_lim) or (x<low_lim)):
             outlier.append(x)
    print(' outlier in the dataset is', outlier)

    Interquartile range is 11.5
    low_limit is -3.75
    up_limit is 42.25
     outlier in the dataset is [75, 45, 70]
```

FIGURE 5. - Code for interquartile range (IQ)

2.3 OUTLIER TYPES

There are several types of outliers [17]:

- General or normal: they are outliers without being linked to other features, such as the salaries of employees, whose value is considered fixed throughout the months of the year.
- Private or Individual: The values go to outliers in certain conditions, such as the temperature being 30 in the winter.
- Collective: a group of outliers, such as delays in home delivery.
- High-dimensional data: This type spreads in high-dimensional data. It is subject to multiple standards and spreads according to the multi-dimensionality.

2.4 OUTLIER DETECTION TECHNIQUES

Figure 6 shows the primary types used in the outliers detection problem, including distance-based algorithms, statistical methods, density-based algorithms, and clustering. These methods are described briefly.

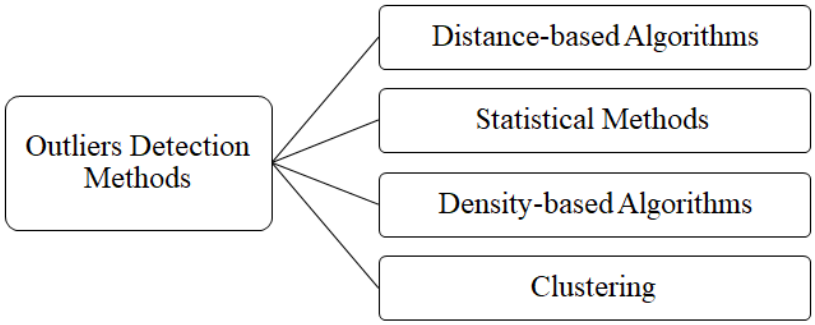

FIGURE 6. - Outliers detection techniques

Distance-based Algorithms [18] scan the vicinity of an object, which is defined by a given radius. Then, it is considered an outlier (or single) if it does not have enough points in its neighborhood. Distance-based methods usually rely on a multidimensional index, which is used to retrieve the neighborhood of each object to see if it has enough points. If there are not enough points, it is called an outlier. Distance-based methods fit better in multidimensional space and can be calculated more efficiently than the statistical method. Identifying outliers based on distance is a significant and helpful activity for examining the data. The main disadvantage of distance-based methods is that detecting distance-based extremes depends on a single value for the assigned process. This can cause problems if the dataset contains dense and sparse regions.

Statistical Methods are used to detect outliers based on statistical tests or procedures. Using Z-score, outliers can be detected based difference between outliers' values and the mean. If the z-score value is higher than three or lowers than -3, it is considered an outlier. In addition, IQR can also be used [19]. The IQR is the value of the middle half of the data set. IQ is used to create "fences" around data and then define outliers as any values outside those fences (see Fig. 7).

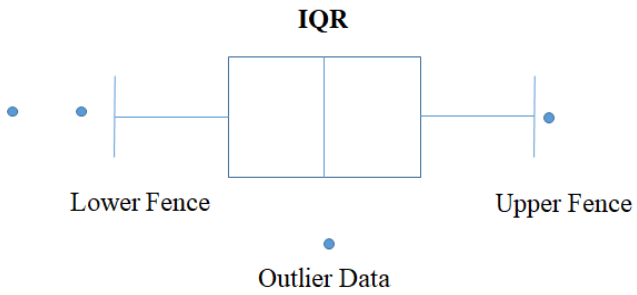

FIGURE 7. - IQR technique

Density-based Algorithms investigates the density of the object and the density of its neighbors [20]. Here, an object is defined as an outlier if its density is relatively less than that of its neighbors. Many real-world datasets show a more complex structure in which objects can be considered outliers in their local neighborhood, rather than in the global data distribution.

Clustering assumes that normal data objects belong to large, dense groups, while outliers belong to small, sparse, or none groups [16]. Clustering-based methods detect outliers by extracting the relationship between objects and mass. The object is abnormal by answering the following questions:

- Does the object belong to any cluster? If not, it will be marked as an outlier.
- Is there a large distance between the object and the cluster that is closest to it? If the answer is yes, then it is an outlier value.
- Is the object part of a small or separate block? If yes, then all objects in that set are outliers.

Clustering is a technique of Machine learning [20]. It belongs to the category of unsupervised learning. Typically, it is a procedure that consists of discovering important structures and groupings of data points in a dataset [21]. Thus, clustering is the division of a population or data point into several groups. Therefore, data points in one category are different from those in other categories. It is a technique for grouping things based on their similarities and differences. There are many ways to perform clustering [22]:

- Distance-based technique: in this technique, the distances are used to assign a point to a cluster. A distance (or, equivalently, a similarity) measure can be considered the main function that quantifies the similarity between two objects. K-means is one of the clustering distance-based techniques [23].
- Density-Based Methods: These methods consider clusters of dense regions with some similarities and differences to the less dense regions of space. These methods are precise and can combine two groups: density-based spatial clustering for applications with noise (DBSCAN) and optical clustering for applications with noise (OPTICS) (order of points to define clustering structure) [24].
- Hierarchical-Based Methods: Based on the hierarchy, the groups formed in this way form a tree-like structure. The previously formed block is used to form new clusters. It is classified into two types: conglomerate (bottom-up approach) and divisional (top-down approach) [25].

3. LITERATURE REVIEW

Many researchers examine the problem of outlier detection in many application fields. The authors in [26] aimed to study the effectiveness of an artificial neural network (ANN) in predicting fraud in the financial reports of small capital firms in Malaysia. Based on the concepts of ANN, the researcher developed a mathematical model to compare non-fraudulent and fraudulent reports selected among small capital firms. In Malaysia, some companies had already been accused by the Securities Commission of falsifying financial statements. Ten financial ratios are used as indicators of fraud risk to predict fraudulent financial reporting using ANN. The results prove that the proposed ANN methodology outperforms other widely used statistical methods for predicting fraudulent financial reporting.

The authors in [27] demonstrated how artificial intelligence works innovatively with systems of internal controls to help managers produce high-quality accounting information by reducing information risk. Despite many types of research proposals using AI in accounting and auditing, none of them have directly demonstrated how to reduce information risk using AI. Research shows that companies reduce many costs and losses resulting from failure to produce reliable accounting information, help managers make better decisions, and improve the overall performance of entities. The paper proposed a general model for using artificial intelligence in practice to automate the removal of flaws in internal control systems. This model applies to all types of business entities. This, in turn, reduces control and detection risks and increases audit quality by reducing accounting information risks.

The authors in [4] presented a summary of the techniques used in detecting data anomalies, defining the characteristics of each technique, in addition to the open challenges and future research directions in identifying new outlier detection strategies. They proposed taxonomy of newly designed anomaly detection strategies and the identification of their characteristics (kama). The research also introduced several popular anomaly detection methods designed for high-dimensional data, streaming data, big data, and minimally classified data.

In addition, the authors in [28] aimed to summarize topics related to artificial intelligence and big data. It also comprehensively analyzes the development of the financial field. It also shows the progress of applying artificial intelligence and big data in finance and taxation. This paper summarizes a risk assessment method for financial information management systems in the context of big data. This method prepares logical nodes according to the system module and builds a risk estimation model by searching for the logical relationship between the data. Based on the background of big data, the batch and flow calculation methods of big data are determined to calculate the system software risk index and the financial data dishonesty risk index.

Moreover, the authors in [29] built a survey on anomaly detection. It is a data analysis task. It is used to identify interesting patterns, trends, and anomalies emerging from the data. Anomaly and outlier detection are essential in many fields, including financial fraud detection, computer network intrusion, human behavior analysis, and gene expression analysis. For instance, there has been renewed interest in research to detect fraudulent activities in the financial sector. There has been a lot of work on unsupervised group-based extremism detection in finance. The researchers presented an in-depth survey of various aggregation-based outlier detection techniques and compared them from different perspectives. Additionally, we discuss the lack of real-world data and how synthetic data has been used to validate current detection techniques.

At last, the authors in [30] reviewed anomaly detection techniques in their survey. They aimed to investigate and provide a comprehensive review of the most common and effective anomaly detection methods applied to detect financial fraud, with an emphasis on highlighting recent developments in the areas of semi-supervised and unsupervised learning.

4. PROPOSED APPROACH

This study aims to create an approach based on machine learning that helps detect outliers in data. This system consists of five steps, as illustrated in Fig. 8 [31-33]:

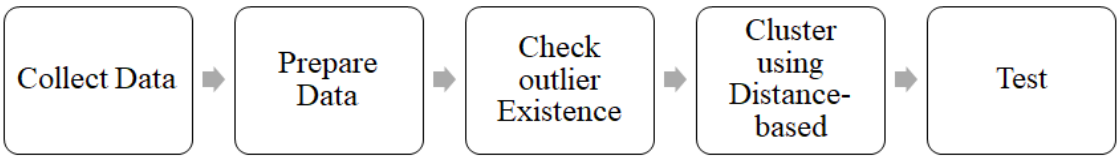

FIGURE 8. - The proposed approach in this paper

Collect Data: the data should be collected first. This can be done from different sources.

Prepare Data: before detecting outliers, data should be preprocessed and cleaned. As data may be collected from different sources, it may be in different formats. Therefore, a transformation step may also be applied.

Check Outlier Existence; this can be performed using a p-value test. Indeed, for each column in the dataset, p-value testing is applied using the Grubbs test in R. According to this test, the existence of outliers can be approved or rejected [34].

Cluster using Distance-based Technique: in this study, clustering based on distance will be used to detect outliers in the data (k_mean).

Test: Here groups are evaluated based on some measure of similarity or difference, such as distance between group points. If the clustering algorithm separates dissimilar and similar points together, then it has done well.

The two most popular metrics for clustering algorithms are the silhouette coefficient and the Dunn index. The Silhouette coefficient is determined for each sample and consists of two degrees:

$$S = \frac{b - a}{\max(a, b)} \tag{1}$$

The variable a the average distance between a point and all other is points in the same group, and the variable b is the average distance between a point and all points in the nearest group.

The silhouette coefficient for a set of samples is given as the average of the silhouette coefficients for each point (Value). The score was set between -1 for incorrect clustering and +1 for high-density clustering. Scores around zero indicate overlapping groups. The score is higher when the clusters are dense and well separated, which relates to the standard concept of mass.

The Dunn Index (DI) is another metric for evaluating a clustering algorithm. Dunn's index is equal to the minimum distance between groups divided by the maximum group size. Note that larger distances between clusters (better separation) and smaller cluster sizes (more compact clusters) lead to a higher DI value. A higher DI indicates better aggregation. Better aggregation is assumed to mean that the blocks are compact and well separated from the other blocks.

5. RESULTS

In this section, the results of applying the proposed technique to a real dataset are illustrated, after describing the dataset.

5.1 DATASET

The dataset was generated from the raw dataset of 600 thousand transaction data collected in 6 years (period 2014 - 2019), indicating the date and time of sale, pharmaceutical brand name, and quantity sold, which is exported from the POS system in individual vs. A selection of drugs from the data set (57 drugs) classified into the Anatomical Therapeutic Chemical (ATC) classification system categories:

- M01AB - Non-steroidal anti-inflammatory and rheumatic products, acetic acid derivatives, and related substances
- M01AE - Anti-inflammatory and rheumatic products, non-stimulants, propionic acid derivatives
- N02BA - Other analgesics and antipyretics, salicylic acid and its derivatives
- N02BE / B - Other Analgesics and Antipyretics, Pyrazolones, and Anilides
- N05B - Antipsychotics, Anxiolytics
- N05C - Neuroleptics, hypnotics, and sedatives
- R03 - Medicines for obstructive airways disease
- R06 - antihistamine for systemic use

Sales data is re-sampled into hourly, daily, weekly, and monthly periods. The data is already pre-processed, which included external detection and processing and the calculation of missing data (see Fig. 9).

	datum	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06
1	1/5/2014	14	11.67	21.3	185.95	41	0	32	7
2	1/12/2014	29.33	12.68	37.9	190.7	88	5	21	7.2
3	1/19/2014	30.67	26.34	45.9	218.4	88	8	29	12
4	1/26/2014	34	32.37	31.5	179.6	80	8	23	10
5	2/2/2014	31.02	23.35	20.7	159.88	84	12	29	12
6	2/9/2014	22.65	29.72	53.3	226.8	98	7	31	6

FIGURE 9. - A screenshot of the dataset studied in this paperwork

This dataset contains nine columns: eight of them represent the features collected in different dates (the ninth column shown as the first one to the left in Fig. 9). In addition, Figure 10 shows a summary of these columns, including minimum, maximum, median, and mean of each column.

```
   datum              M01AB           M01AE           N02BA
Length:302         Min.   : 7.67   Min.   : 6.237   Min.   : 3.50
Class :character   1st Qu.:29.39   1st Qu.:22.387   1st Qu.:21.30
Mode  :character   Median :34.56   Median :26.790   Median :26.50
                   Mean   :35.10   Mean   :27.168   Mean   :27.06
                   3rd Qu.:40.17   3rd Qu.:31.047   3rd Qu.:32.48
                   Max.   :65.33   Max.   :53.571   Max.   :60.12
    N02BE             N05B            N05C             R03
Min.   : 86.25   Min.   : 18.00   Min.   : 0.000   Min.   :  2.00
1st Qu.:149.30   1st Qu.: 47.00   1st Qu.: 2.000   1st Qu.: 21.00
Median :198.30   Median : 57.00   Median : 3.979   Median : 35.00
Mean   :208.63   Mean   : 61.74   Mean   : 4.139   Mean   : 38.44
3rd Qu.:252.47   3rd Qu.: 71.00   3rd Qu.: 6.000   3rd Qu.: 51.00
Max.   :546.90   Max.   :154.00   Max.   :17.000   Max.   :131.00
     R06
Min.   : 1.00
1st Qu.:11.47
Median :17.50
Mean   :20.22
3rd Qu.:26.00
Max.   :65.00
```

FIGURE 10. - Columns summary showing mean, min, max, median, etc.

Figure 10 summarizes this data set obtained by the summary (Summary()) function used in the R programming language, showing the minimum, maximum, median, average, first Qu, and third Qu of each attribute, as follows:

- The lower quartile is also referred to as the first quartile. When sorting in ascending order, the value removes the first 25% of the data.
- The mean is also referred to as the second quartile. It is the value that sheds the first 50%.
- Upper spring is another name for the third spring. It is the value that eliminates the first 75% of the population.
- The median divides the data set into two halves; 50% of the data points are less than or equal to, and the remaining 50% are greater than or equal to the median.

5.2 RESULTS

In this part, the result of clustering will be illustrated. First, the Grupps test will be applied to test the existence of outliers. After that, the boxplot will be visualized. The clustering will be shown later.

5.2.1 Grupps Test

The Grupps test result is illustrated in Table 1.

Table 1. - Grupps test results

Data Column	Hypotheses	P_Value	Result
M01AB	H0: There is no outlier in the data in M01AB	0.1201	H0 approved
	H1: There is an outlier in the data in M01AB		H1 rejected
M01AE	H0: There is no outlier in the data in M01AE	0.0454	H0 rejected
	H1: There is an outlier in the data in M01AE		H1 approved
N02BA	H0: There is no outlier in the data in N02BA	0.0103	H0 rejected
	H1: There is an outlier in the data in N02BA		H1 approved
N02BE	H0: There is no outlier in the data in N02BE	0.001875	H0 rejected
	H1: There is an outlier in the data in N02BE		H1 approved
N05B	H0: There is no outlier in the data in N05B	0.009271	H0 rejected
	H1: There is an outlier in the data in N05B		H1 approved
N05C	H0: There is no outlier in the data in N05C	0.009356	H0 rejected
	H1: There is an outlier in the data in N05C		H1 approved
R03	H0: There is no outlier in the data in R03	0.01275	H0 rejected
	H1: There is an outlier in the data in R03		H1 approved
R06	H0: There is no outlier in the data in R06	0.02056	H0 rejected
	H1: There is an outlier in the data in R06		H1 approved

5.2.2 BoxPlot

Now, the existence of the outliers will be verified using BOXPLOT (see Fig. 11). There is an outlier for the associated features.

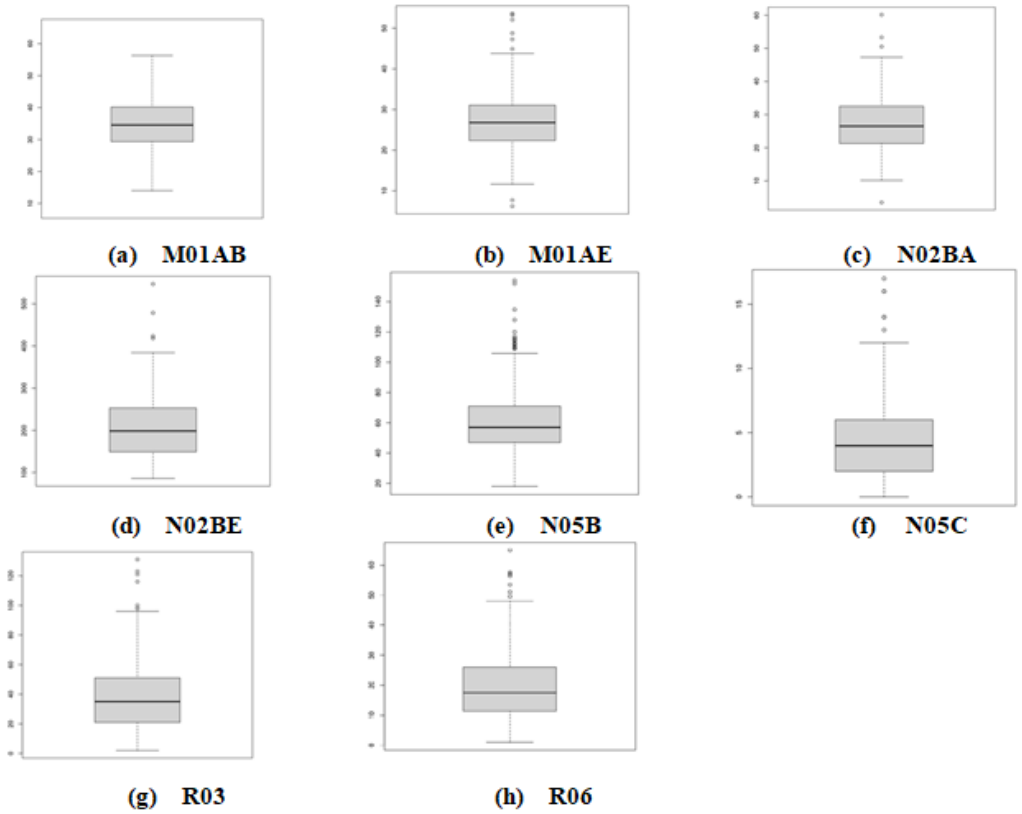

FIGURE 11. - Box plot showing the outliers for each data column

5.2.3 Clustering

The data in the columns (M01AE, N02BA, N02BE, N05B, N05C, R03, and R06) are clustered using k_means methods into two clusters:

- Cluster 1: normal data
- Cluster 2: outliers' data

The results are also illustrated in Fig. 12.

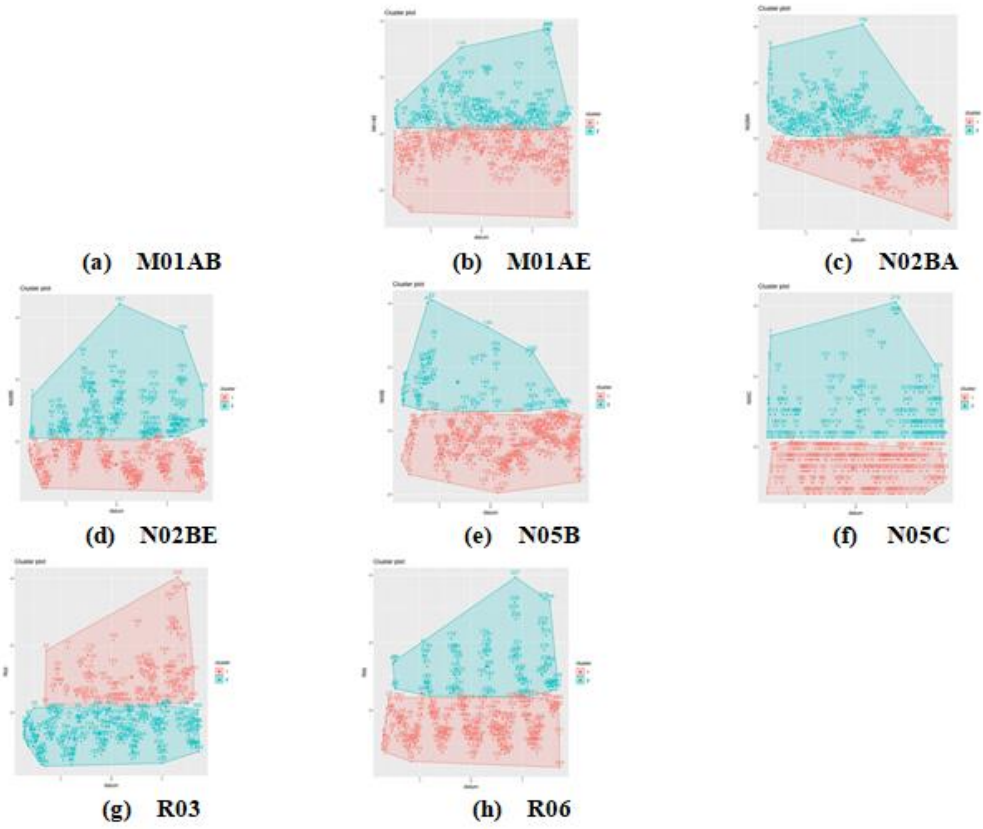

FIGURE 12. - Clustering for each data column

5.2.4 Performance

The performance is measured using two indices: the silhouette value and the Dunn index. The silhouette value is a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters. The values of all silhouettes are between 0.48 and 0.67, as illustrated in Table 2. These values are close to 1. Thus, the performance of clustering is good.

Table 2. -Silhouette values

Silhouette Distance	Cluster 1	Cluster 2	Silhouette Distance	Cluster 1	Cluster 2
M01AB	-	-	N05B	0.55	0.61
M01AE	0.57	0.42	N05C	0.52	0.65
N02BA	0.66	0.52	R03	0.48	0.66
N02BE	0.68	0.49	R06	0.51	0.66

The Dunn index aims at quantifying the compactness and variance of the clustering. A cluster is considered compact if there is a small variance between members of the cluster. A higher Dunn Index will indicate compact, well-

separated clusters, while a lower index will indicate less compact or less well-separated clusters. All the Dunn values are small, as illustrated in Table 3. This is evident as the clusters in this data are not well separated.

Table 3. - Dunn index

	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06
Dunn Value	-	0.0097	0.0066	0.0056	0.0123	0.0312	0.0115	0.0078

6. CONCLUSION

Recently, the volume of data held in the business sector becomes enormous. Therefore, analytics strategies should be used to address this problem. These strategies should cover people, technology, and processes to achieve business goals and objectives.

The huge volume of available data may present many impurities like outliers, missing data, and uncertain data. Outliers' data are wandering data that can skew averages, and should usually be filtered out and excluded before trying to get information from the data. Indeed, outliers' data may affect the data analysis process by providing error decisions or wrong classification, or other problems.

Data anomaly detection is fast becoming a popular topic in the business world as companies look for better ways to predict performance outcomes based on raw data.

This study aimed to build an approach for outliers' detection using machine learning techniques; clustering distance-based method is adopted. After describing anomalous data, the author briefly describes the outlier detection technique. Then, the proposed approach is described. Later, the result of applying this approach to a real dataset collected from Kaggle.com was illustrated. This dataset contains eight rows. After testing existing anomalies using the Grupp test, only seven rows show outliers. Then, the proposed approach is applied to detect anomalies in these rows. Two performance measures were used to measure assembly accuracy: the silhouette and the Dunn index. The silhouette value represents an object's similarities with its group (coherence) compared to other clusters (disjunction). The silhouette values associated with all rows were close to 1. Therefore, clustering performs well on this problem. Dunn's index quantifies the compressibility and anisotropy of aggregation. Dunn's values in all grades were small. It is normal because the groups are not well separated.

After completing this study, the author recommends the following:

- This technique works well, but it should be compared to another one.
- Companies in Arab countries should train their employees to understand what anomalies are in the data and why it is important to spot them.
- Companies in Arab countries should study data on extremism.
- Companies should hire data mining experts to detect outliers.
- More techniques should be tested for the problem of outliers detection.

FUNDING

No funding received for this work

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their efforts.

CONFLICTS OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 2021.

[2] G. Michau and O. Fink, "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer," Knowledge-Based Systems, vol. 216, pp. 106816, 2021.

[3] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaei, "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics," *Journal of Big Data*, vol. 8, no. 1, pp. 1-20, 2021.

[4] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier detection: Methods, models, and classification," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1-37, 2020.

[5] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in internet of things," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4672-4683, 2018.

[6] J. Bian, D. Tian, Y. Tang, and D. Tao, "A survey on trajectory clustering analysis," arXiv preprint arXiv:1802.06971.

[7] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden Markov models and their applications," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 1429-1448, 2021.

[8] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on Bayesian network structure learning from data," *Progress in Artificial Intelligence*, vol. 8, no. 4, pp. 425-439, 2019.

[9] C. C. Aggarwal, "An introduction to outlier analysis," Springer International Publishing, 2017.

[10] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-33, 2021.

[11] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. HU, "COPOD: copula-based outlier detection," in *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2020, pp. 1118-1123.

[12] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, pp. 1, 2020.

[13] A. Degirmenci and O. Karal, "Efficient density and cluster based incremental outlier detection in data streams," *Information Sciences*, vol. 607, pp. 901-920, 2022.

[14] K. Singh and M. Xie, "Bootlier-plot: bootstrap based outlier detection plot," Sankhyā: The Indian Journal of Statistics, pp. 532-559, 2003.

[15] S. Walfish, "A review of statistical outlier methods," *Pharmaceutical Technology*, vol. 30, no. 11, pp. 82, 2006.

[16] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 28, pp. 100306, 2020.

[17] L. Bottmer, C. Croux, and I. Wilms, "Sparse regression for large data sets with outliers," *European Journal of Operational Research*, vol. 297, no. 2, pp. 782-794, 2022.

[18] M. Sugiyama and K. Borgwardt, "Rapid distance-based outlier detection via sampling," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[19] J. Yang, S. Rahardja, and P. Fränti, "Outlier detection: how to threshold outlier scores?," in *Proceedings of the international conference on artificial intelligence, information processing and cloud computing*, pp. 1-6, Dec. 2019.

[20] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, pp. 104743, 2022.

[21] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised And Unsupervised Learning For Data Science*, pp. 3-21, 2020.

[22] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020.

[23] M. Cui, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5-8, 2020.

[24] J. Tobin and M. Zhang, "DCF: an efficient and robust density-based clustering method," in *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 629-638, Dec. 2021.

[25] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmospheric Pollution Research*, vol. 11, no. 1, pp. 40-56, 2020.

[26] N. Omar, Z. A. Johari, and M. Smith, "Predicted fraudulent financial reporting using an artificial neural network," *Journal of Financial Crime*, vol. 24, no. 2, pp. 362-387, 2017.

[27] S. Askary, N. Abu-Ghazaleh, and Y. A. Tahat, "Discuss artificial intelligence and the reliability of accounting information in Challenges and Opportunities," in the *Digital Era: 17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society*, Kuwait, pp. 315-324, Nov. 2018.

[28] N. Wang, Y. Liu, Z. Liu, and X. Huang, "Application of artificial intelligence and big data in modern financial management," in *2020 International Conference on Artificial Intelligence and Education (ICAIE)*, pp. 85-87, IEEE, June 2020.

[29] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," Future Genration Computer Systems, vol. 55, pp. 278-288, 2016.
[30] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," 2022.
[31] D. Dinesh, *Techniques and Challenges in Building Intelligent Systems: Anomaly Detection in Camera Surveillance*. Springer, Switzerland, vol. 2, pp. 11-21, 2016.
[32] A. A. Ghorbani, W. Lu, and A. Mahbod, *Network Intrusion Detection and Prevention: Concepts and Techniques*. Springer, New York, 2010.
[33] P. K. Kumar, "Classification and Genetic based Anomaly Detection," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 2, pp. 504-507, 2016.
[34] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia, "Testing for outliers with conformal p-values," *The Annals of Statistics*, vol. 51, no. 1, pp. 149-178, 2023.