Outlier Detection of Transaction Data Using DBSCAN Algorithm

[1]Sunjana, [2]Azizah Zakiah

Abstract---*The supermarket is one means of marketing the company's products. Marketing activities undertaken with supermarket provides a wide range of types of products from different companies (as producers). Consumers prefer to go to the supermarket than traditional markets due to promo. For example, the products offered were given discounted half price of normal price. Consumers tend to buy more of their needs so that existing stock items in the supermarket can be drastically reduced. Therefore, the supermarket had to anticipate in order to not shortage of stock in the warehouse. Various techniques in data mining can be used, one that is an outlier detection. The role of an outlier detection is needed in order to detect abnormal transactions including candidate anomalies and normal transactions and will be help the supermarket in anticipation of running out of stock items. Outlier detection is an outlier search process on dataset and is one of the first steps to be able to perform analysis of data coherent. The main objective in outlier detection is to detect data with properties/state data with different data or are most of the anomalies found in multidimensional datasets. One of the formidable algorithms for detecting outlier is DBSCAN. Therefore, in this study, the author will use the technique of outlier detection algorithm with expected DBSCAN can help supermarket in anticipation of running out of stock items. The result from research that has been done by calculating 1862 products there was no product data that classified as outlier, whereas by calculating 100 first products there are 4 products data that classified as outlier, products with id 80069449, 80015728, 82024920, 80021527.*

Keywords---*Data Mining, Outlier Detection, Euclidean Distance, Clustering , DBSCAN.*

I. INTRODUCTION

Supermarkets are one of the company's product marketing tools. Supermarkets use a variety of ways to get consumers by way of prices that are not too far from traditional markets, there are various promos, more complete and many products, and various facilities. With the promo, consumers tend to buy products more than their needs so that the stock of goods can be reduced drastically. Therefore, the supermarket must anticipate that there is no shortage of stock in the warehouse. Various techniques can be used, one of which is detection outliers.

Outlier detection is the process of finding outliers in a dataset, and is one of the first steps to be able to do a coherent data analysis. The main objective in outlier detection is to detect data with data properties that are different from most data, or are anomalies found in multidimensional datasets [2]. The method that can detect anomalous data is by using the Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm.

II. ATURE STUDY

A. Clustering Analysis

[1]*Computer Science, Faculty of Engineering, Widyatama University, Jln. Cikutra 20124 A, Bandung 40125, Indonesia.*
[2]*Computer Science, Faculty of Engineering, Widyatama University, Jln. Cikutra 20124 A, Bandung 40125, Indonesia.*
sunjana@widyatama.ac.id

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

Cluster analysis is the work of grouping data (objects) based only on information found in data that describes the relationship between these objects [6]. The purpose of cluster analysis is that objects that have similarities join in a group and are different from objects in other groups. In this study, clustering is used for supermarket transaction data segmentation to separate anomalous objects from objects that are normal.

Based on the data compactness theory, clustering method is divided into two, namely complete and partial. All data can be said to be compact into one group if all data can be combined into one group (in the context of insulation), but if there is little data that does not join in the majority group, then the data is said to have deviant behavior. Data that has this deviant behavior, known as noise. Based on the results of previous studies a fairly good method for detecting this noise is DBSCAN [8].

B. Outlier Detection

Outlier detection is a process to find outlier data in a group of data, and is one of the first steps used to analyze coherent data [2]. Outliers are things that are considered different from the data in general. Outliers in a group of data can arise due to various things, among others, errors when data is inputted, errors during measurement and data collection, or indeed the original characteristics of the data [8]. In conducting outlier detection, it can be done using several approaches including:

1. Graphic Approach
   Suppose using a plot box (1D), scatter plot (2D) and spin plot (3D). To do detection outliers using graphical methods, it can be done by plotting the data with the i-observation (i = 1,2,3, ..., n). If the data obtained from the plot data is located far from all data sets, then the data can be indicated as an outlier.

2. Statistical approach
   To be able to approach the statistical method, assume the data distribution function that is owned, then use a statistical test that depends on:
   a) Data distribution
   b) Distribution Parameters
   c) Number of acceptable outliers (confidence interval)
3. Nearest-Neighbor Based
   To detect anomalies using the nearest-neighbor approach method, first determine the distance from each pair of data points. A data point is said to be an outlier if:

   a) The number of neighboring points is less than p in the distance D.
   b) This point is the top n point which is the farthest distance from the nearest neighbor.
   c) This point is the top n average point of the largest distance from the nearest neighbor.

4. Density Based
   Another approach that can be used to perform detection outliers is density based methods. Based on the density based approach, outliers are data points that are in low density areas. This density based method can be used for data with different densities, but what needs to be considered in this method is the selection of parameters. Because the selection of parameters in the density based method determines the density value.

C. DBSCAN

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a clustering method that builds an area based on density connected. DBSCAN is a type of partitioning cluster where areas of high density are considered as clusters while those with low density or not incorporated in the cluster are considered noise [10]. DBSCAN requires two input parameters, epsilon (Eps) and minimum points (MinPts). EPS around point is defined as:

$$N_{Eps}(x) = \{ y \in D \mid dist(x, y) \leq Eps \} \text{ ……………....(1)}$$

Where :

$N_{Eps}(x)$: the surrounding point of x within the Eps radius

D: Data Cluster

$dist(x, y)$: euclidean distance from objects x and y

Eps : radius or threshold

The algorithm from DBSCAN is as follows:

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

1) Select random starting point r.
2) Initialize input parameters: MinPts and Eps
3) Calculate Eps or all distances of affordable density to r using euclidean distance.
4) If the point meets the Eps more than MinPts, then point r is the center point (core) and the cluster is formed by following the following conditions:
   i. $\forall x, y$: if x ∈ C (cluster) and y affordable density of x, then y ∈ C. (Maximality)
   ii. $\forall x$, y∈ C: x density is connected to y. (Connectivity)
5) Repeat steps 3-4 until all points are processed.
6) If r is a border point and there is no point with an affordable density to r, then the process continues to another point [4].

D. Euclidean Distance

Euclidean distance is an equation commonly used in calculating the distance of two objects. The distance calculated using this Euclidean distance formula, which is the square root of the difference in distance between the two objects of the rank of two. This distance calculation is often used to determine the similarity between two objects [12].

$$D = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad (2)$$

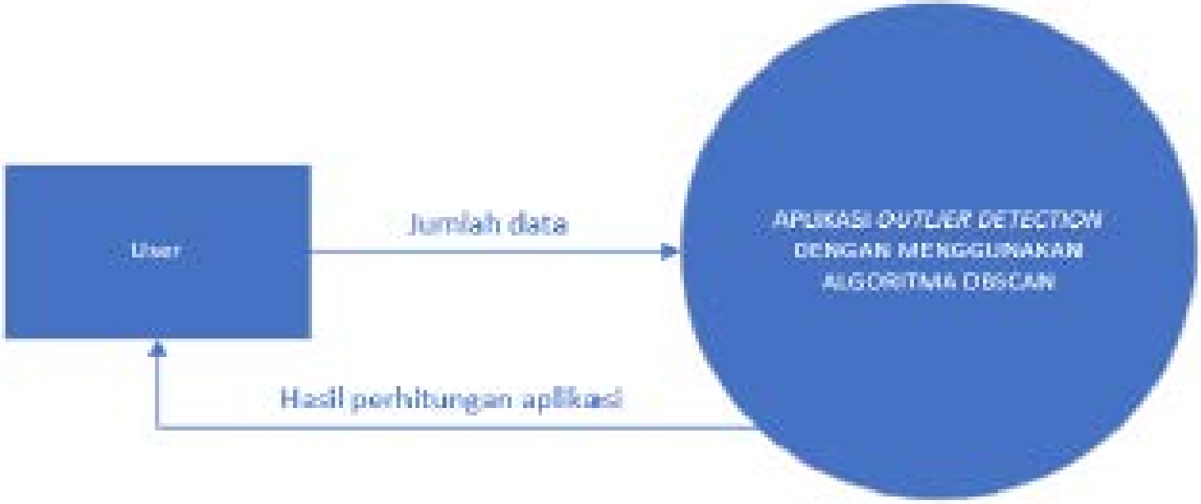
Euclidean distance is often used in calculating the distance between two objects [13]. Euclidean distance is often used in clustering techniques, because this formula is easy to understand but sensitive to outliers [12].

III. N AND IMPLEMENTATION

E. System Design

The system built in this research is outlier detection using the Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm. The algorithm does not require input in the form of cluster numbers, because this algorithm will automatically form a cluster according to the characteristics of the data. In the cluster formation process, this algorithm only requires input in the form of epsilon (eps) and minimum point (minpts) values. In this study, epsilon and minimum point values are obtained from the preprocessing results by calculating the distance between products using the euclidean distance formula.

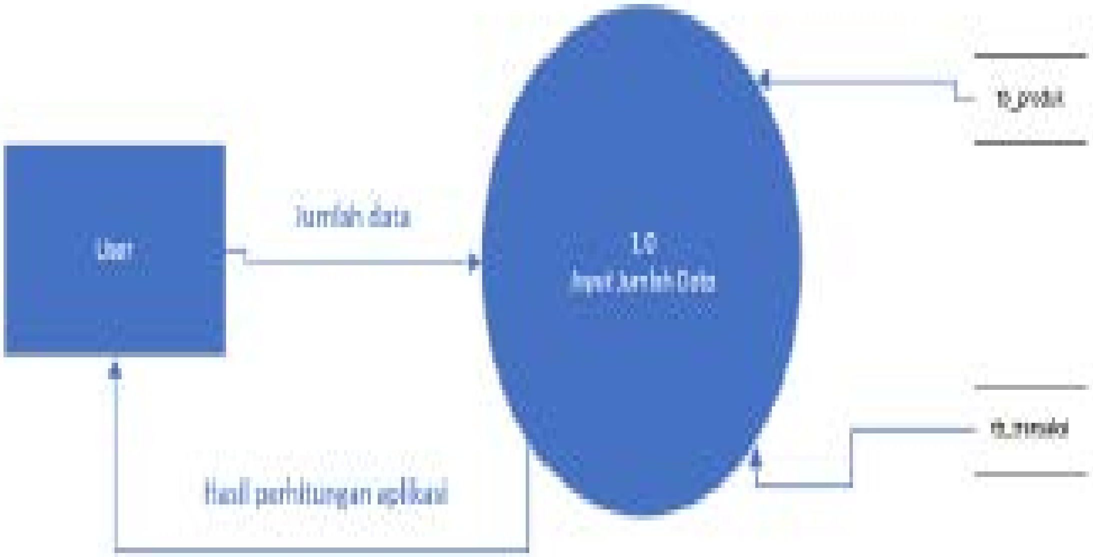
The following is a Context Diagram of the system.



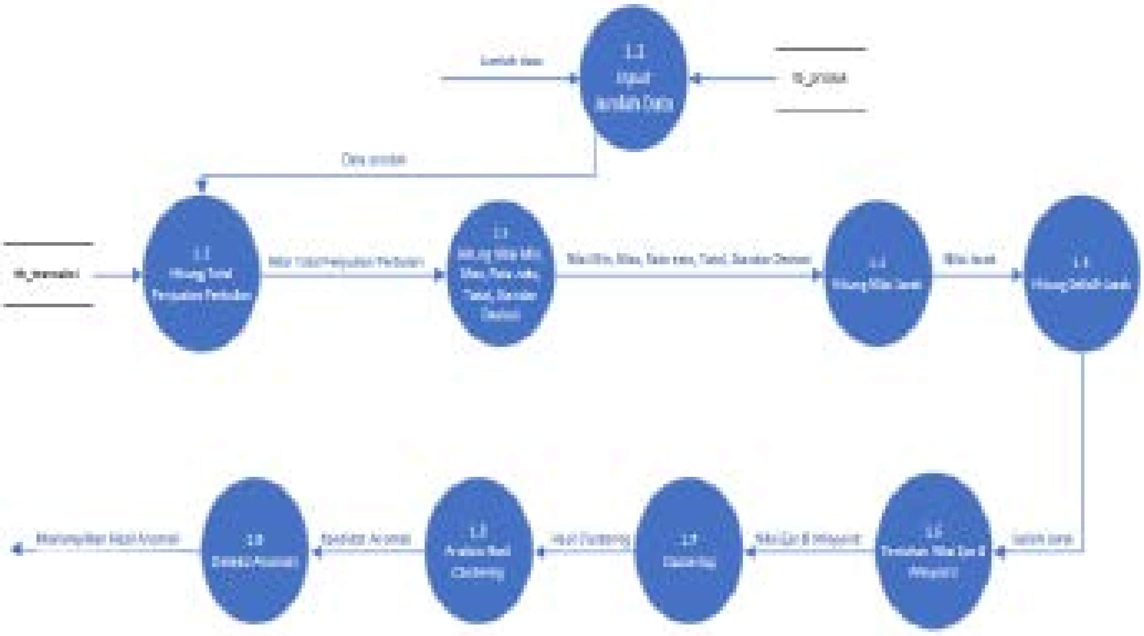
Ficture3.1*ContextDiagram*

After forming a context diagram, then the next process design is made using Data Flow Diagrams (DFD).

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020



Ficture 3.2 *Data Flow Diagram Level* 1



Ficture 3.3 *Data Flow Diagram Level* 2

F. System Implementation

In this study, discussed about the data that is used as input for the mining process using the Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm. Where in this study using a web-based system with PHP programming language and using MySQL as a place to create a database. The system is divided into three, namely input, process, and output. Each part of the system will be explained as follows.

a) Input

The data used in this study is transaction data from products in a supermarket. The number of transactions used was 9999 transactions from 1862 products.

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

b) Process

The process carried out in this study is to calculate the value of eps and minpts by using the euclidean distance formula to determine the transactions of which products are normal and including anomalous candidates.

c) Output

The system will show the results of system calculations in the form of clustering results of each product. Which products include normal transactions and which products are included in anomalous candidates.

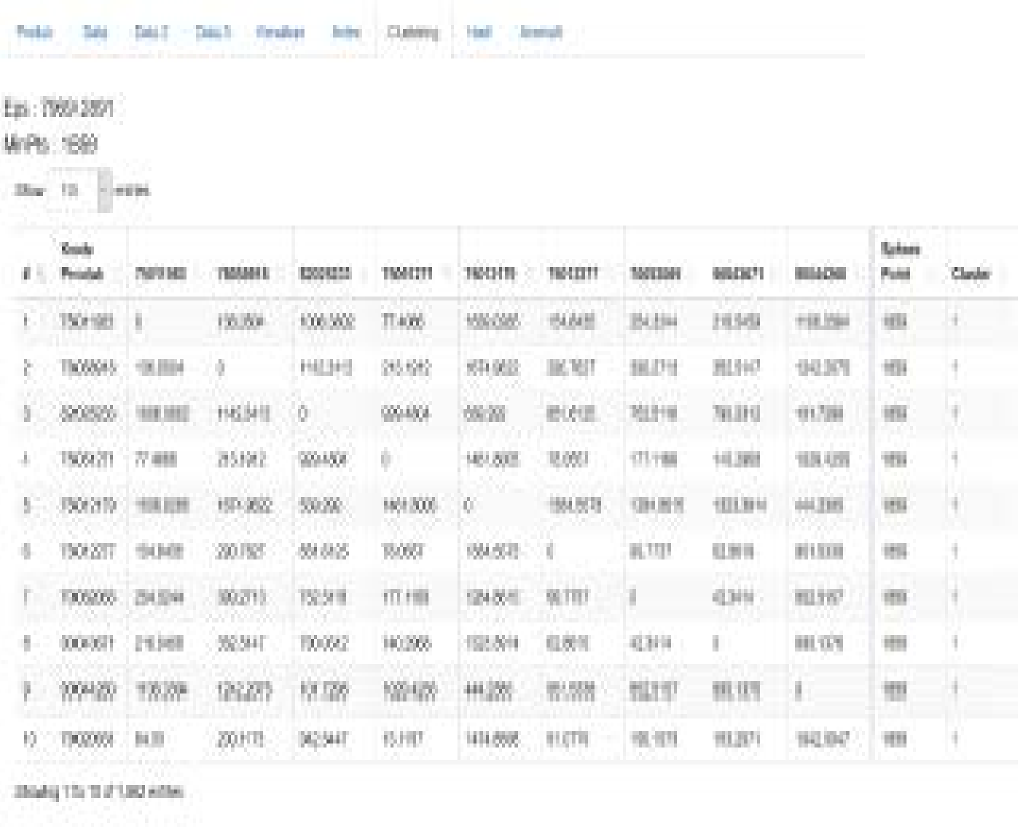
IV. ING AND ANALYSIS OF RESULTS

A. Scenario Testing

The performance of the Density Based Spatial Clustering algorithm of Applications with Noise (DBSCAN) is that it does not do the total input of the cluster to be formed. Where in this algorithm the total cluster is formed according to the epsilon value and the minimum point used. The following are the scenarios in this study.

1. Scenario 1

In scenario 1, the transaction data of 1862 products taken from January to December are used. The value of eps obtained is 7969,2891 and the value of the minpts obtained is 1859.



Eps : 7969,2891

MinPts : 1859

Show 10 entries

#	Kode Produk	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] Point	Cluster
1	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1859	1
2	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1859	1
3	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1859	1
4	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1859	1
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	1859	1
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	1859	1
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	1859	1
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	1859	1
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	1859	1
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1859	1

Showing 1 to 10 of 1,862 entries

Figure 4.1: *Eps and MinPts values*

By using eps value of 7969,2891 and minpts value of 1859, the results shown in the system indicate that there is no data from the data used including noise.

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

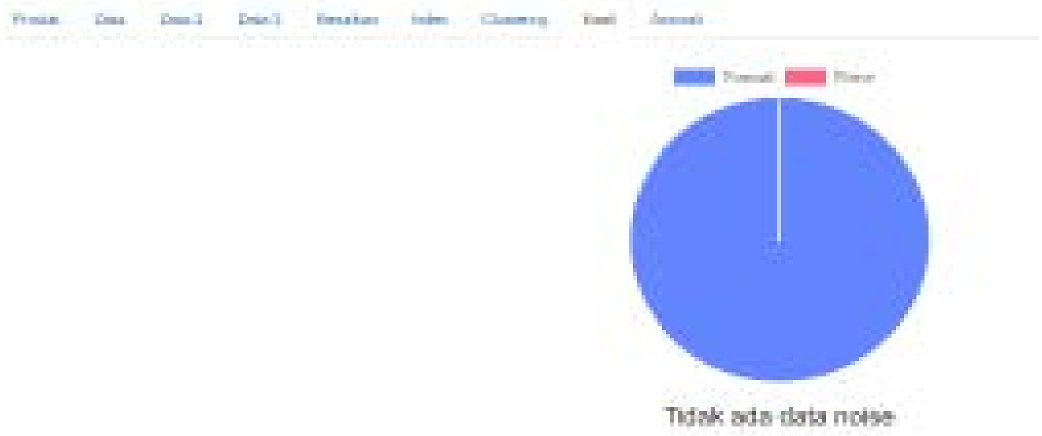
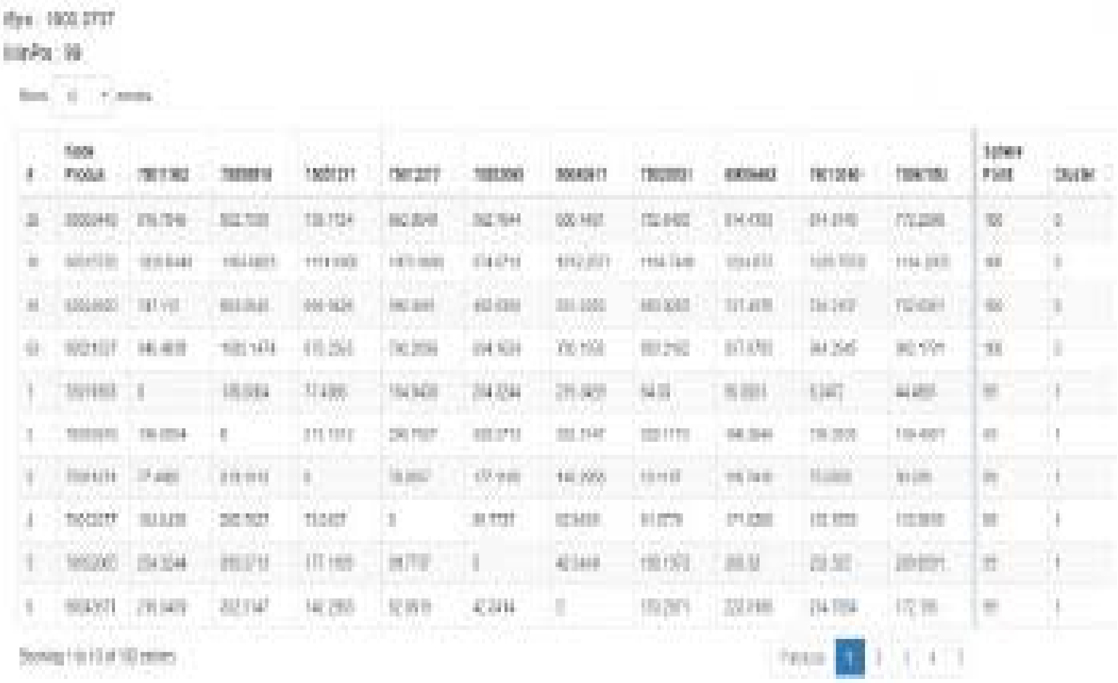


***Figure* 4.2.** Clustering Results

2. Scenario 2
   In scenario 2, the transaction data from the first 100 products sold are taken from January to December. The value of eps obtained is 1500.0737 and the value of the minpts obtained is 99.



Eps : 1500.0737

MinPts : 99

Figure* 4.3.Eps andMinPts values*

By using the Eps value of 1500.0737 and the MinPts value of 99, the results displayed in the system indicate that the data used contained product data including anomalies, namely products with id 80069449, 80015728, 82024920, 80021527.

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

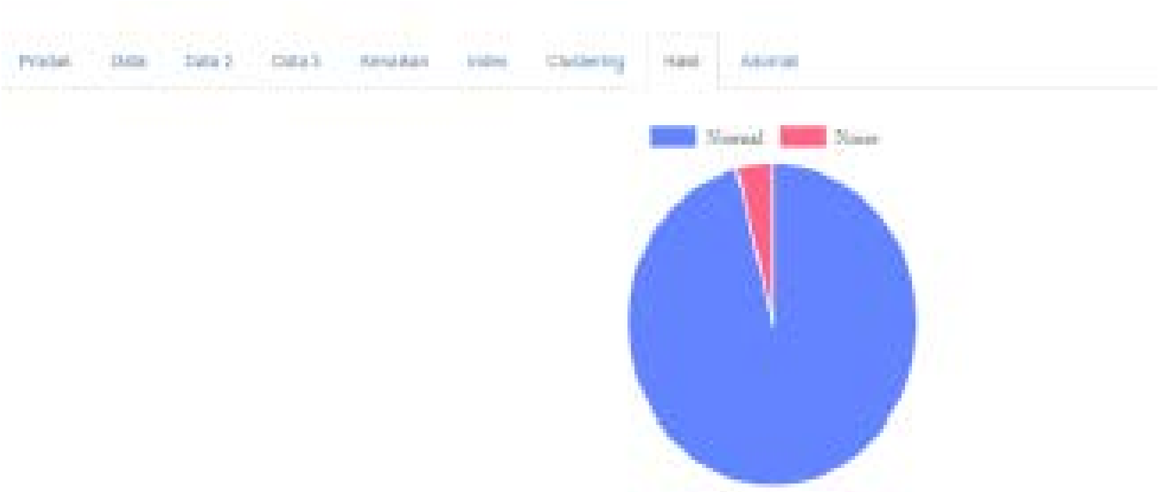
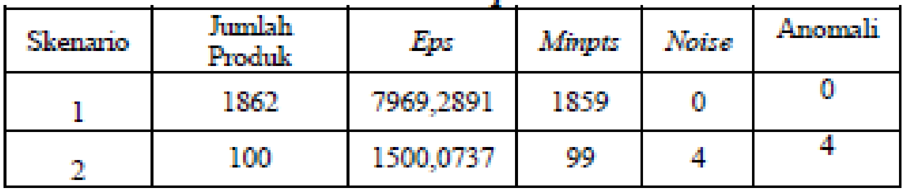


Figure* 4.4.Clustering Results*

B. Analysis of Test Results

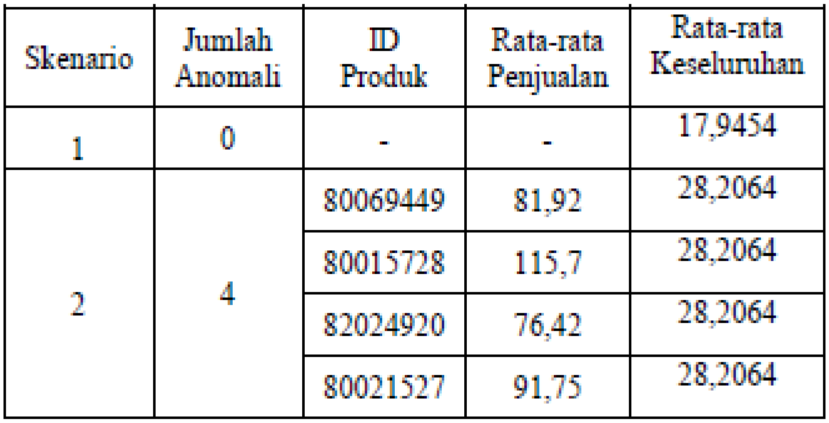
In this study the data used is data in a period of 1 year and is made in 2 scenarios. Where in scenario 1 consists of 9999 transactions and 1862 products calculated. While scenario 2 consists of 9999 transactions but only the first 100 products are counted. The following table of detection results in this study:

Table 4.1:*Tput Results 2 Scenario*



Skenario	Jumlah Produk	*Eps*	*Minpts*	*Noise*	Anomali
1	1862	7969,2891	1859	0	0
2	100	1500,0737	99	4	4

Table 4.2:*Anomaly Details*



Skenario	Jumlah Anomali	ID Produk	Rata-rata Penjualan	Rata-rata Keseluruhan
1	0	-	-	17,9454
2	4	80069449	81,92	28,2064
		80015728	115,7	28,2064
		82024920	76,42	28,2064
		80021527	91,75	28,2064

V. CLUSIONS AND RECOMMENDATIONS

A. Conclusions

Based on the results of the testing and analysis that has been carried out regarding " OUTLIER DETECTION TRANSACTION DATA USING DBSCAN ALGORITHM
" can be concluded as follows:

1. Detecting anomalous candidates can be used with the Density Based Spatial Clustering of Applications algorithm with Noise (DBSCAN) and the euclidean distance distance calculation formula. By using transaction data for 1 year in 2015, including:

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

a) Scenario 1 by calculating all products totaling 1862 products produced epsilon value of 7969.2891 and minimum point value of 1859. By using the epsilon and minimum point values, 1 cluster was produced and there was no data noise or anomalous data.
b) Scenario 2 by calculating the first 100 products sorted by sales from 1st month resulted in epsilon value of 1500.0737 and minimum point value of 99. By using the epsilon value and minimum point, 4 anomalous candidates were generated.

2. By using the Density Based Spatial Clustering Application with Noise (DBSCAN) algorithm, the anomaly values obtained from the candidates generated from each scenario are:
   a) In scenario 1 uses transaction data for 1 year, the number of products counted is 1862 products. The results obtained are no anomalous data.
   b) In scenario 2 uses transaction data for 1 year, the number of products counted as many as the first 100 products sorted by sales from the 1st month. There are 4 anomalies, product id 80069449, product id 80015728, product id 82024920, product id 80021527.

B. Recommendations

The implementation of the Density Based Spatial Clustering Applications with Noise (DBSCAN) algorithm is not yet a perfect implementation, so it needs an implementation improvement as needed. Some of the things suggested for further development in this implementation are as follows:

a) For further development, a number of data scenarios can be added to make them more accurate in determining anomalous candidates.
b) Further development can be implemented using other algorithms to compare the accuracy of anomalous candidate detection.

REFERENCES

[1] Asih, Nur dkk. 2016. Metode Peng*cluster*an Berbasis Densitas Menggunakan Algoritma DBSCAN. Bandung: Universitas Islam Bandung.

[2] Devi, Ni Made Anindya Santika dkk. 2015. Implementasi Metode *Clustering* DBSCAN pada Proses Pengambilan Keputusan. Bali: Universitas Udayana.

[3] Fitriany, Indah Ayu. 2017. *Anomaly Detection* Pada Data Konsumsi Listrik Pelanggan Menggunakan Algoritma *Density Based Spasial Clustering Application with Noise*, Studi Kasus : PT PLN (persero) Distribusi Jabar Area Purwakarta. Universitas Widyatama.

[4] Handriyadi, Dedy dkk. 2009. Analisis Perbandingan *Clustering-Based*, *Distance-Based,* dan *Density-Based,* Dalam Mendeteksi *Outlier*. Bandung: IT Telkom.

[5] Hussain, H.I., Kamarudin, F., Thaker, H.M.T. & Salem, M.A. (2019) Artificial Neural Network to Model Managerial Timing Decision: Non-Linear Evidence of Deviation from Target Leverage, *International Journal of Computational Intelligence Systems,* (forthcoming).

[6] Jariah, Nur. 2007. Analisis *Brand Switching* Untuk Memprediksi *Market Share* Dan Segmentasi Terhadap Jenis Merek *Shampoo* Dengan *Marcov Chain* Dan *Cluster Analysis* Studi Kasus: Toserba Swalayan MITRA Kartasura. Surakarta: Universitas Muhammadiyah.

[7] Jiawei, Han dkk. 2011. *Data mining: Concept and Tecniques Third Edition* USA: Elsevier Inc

[8] Lailasari, Siti Nur Elia dkk. 2009. Implementasi Dan Analisis *Distance-Based Outlier Detection* Pada Kumpulan Artikel Web Berita Berbahasa Indonesia. Bandung: Universitas Telkom.

[9] Mumtaz, K. and Duraiswamy, K., (2010). *An analysis on density based clustering of multi dimensional spatial data. Indian Journal of Computer Science and Engineering*, 1(1), pp.8-12.

[10] Nagpal, P. B., & Mann, P. A. (2011). *Comparative study of density based clustering algorithms. International Journal of Computer Applications*, 27 (11), 44-47.

[11] Prasetyo, Eko. 2014. "DATA MINING-Mengolah Data Menjadi Informasi Menggunakan Matlab". Yogyakarta: Andi Yogyakarta.

[12] Sinwar dan R. Kaushik, "Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering". *INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET),* vol. 2, no. 5, 2014.

[13] Solimun (2002), *Structural Equation Modeling* LISREL dan Amos, Fakultas MIPA Universitas Brawijaya, Malang.

[14] Tan, dkk. 2006. "TAHAPAN *KNOWLEDGE DISCOVERY in DATABASE*".

[15] Vitalievichaveryanov, S., Khairzamanova, K.A., Kudashkina, N.V., Hasanova, S.R., Tuygunov, M. Efficiency of clinical application of phytofilm in treating patients with traumatic lesions of oral mucosa(2018) International Journal of Pharmaceutical Research, 10 (4), pp. 611-615.

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062407112&partnerID=40&md5=c790c8507f3c4f39fc6e07ef73e66f55

[16] M. I. Niyas ahamed (2014) ecotoxicity concert of nano zero-valent iron particles- a review. Journal of Critical Reviews, 1 (1), 36-39.

[17] Gangurde HH, Gulecha VS, Borkar VS, Mahajan MS, Khandare RA, Mundada AS. "Swine Influenza A (H1N1 Virus): A Pandemic Disease." Systematic Reviews in Pharmacy 2.2 (2011), 110-124. Print. doi:10.4103/0975-8453.86300

DOI: 10.37200/IJPR/V24I2/PR200632
Received: 02 Jan 2020 | Revised: 15 Jan 2020 | Accepted: 28 Jan 2020