Jurnal Teknik Informatika Vol. 15 No. 1, April 2022 (1-10)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i1.25682

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Outlier Detection in Inpatient Claims Using DBSCAN and K-Means

Panca Oktavia Candra Sari[1], Suharjito[2]

[1,2] Computer Science Department, BINUS Graduate Program
[1,2]Bina Nusantara University
[1,2] Jakarta, Indonesia 11480
E-mail: [1]panca.sari@binus.ac.id, [2]suharjito@binus.edu

Article:
Accepted: April 16, 2022
Revised: May 13, 2022
Issued: June 15, 2022

*Correspondence Address:
panca.sari@binus.ac.id

ABSTRACT

Health insurance helps people to obtain quality and affordable health services. The claim billing process is manually input code to the system, this can lack of errors and be suspected of being fraudulent. Claims suspected of fraud are traced manually to find incorrect inputs. The increasing volume of claims causes a decrease in the accuracy of tracing claims suspected of fraud and consumes time and energy. As an effort to prevent and reduce the occurrence of fraud, this study aims to determine the pattern of data on the occurrence of fraud based on the formation of data groupings. Data was prepared by combining claims for inpatient bills and patient bills from hospitals in 2020. Two methods were used in this study to form clusters, DBSCAN and K-Means. To find out the outliers in the cluster, Local Outlier Factor (LOF) was added. The results from experiments show that both methods can detect outlier data and distribute outlier data in the formed cluster. Variable that high effect becomes data outlier is the length of stay, claims code, and condition of patient when discharged from the hospital. Accuracy K-Means is 0.391, 0.003 higher than DBSCAN, which is 0.389.

Keywords: *Fraud; Data Mining; DBSCAN; K-Means; Outlier;*

I. INTRODUCTION

Health is one of the focuses of the Indonesian government. One solution to provide comprehensive treatment, Indonesian government provide health insurance for all citizens. This facility helps people living in rural areas or the poor to access comprehensive treatment. Processing claims by entry manually into the system according to the diagnostic code and treatment. It is highly susceptible to human error and is considered fraud.

Fraud that occurs in the health care sector is a problem that arises that affects health services. As a result, patients who need treatment for health services are hampered[1]. There are 3 categories of fraud in health insurance, first is giving expensive therapy that is not needed by the patient or giving medication that exceeds the dose, second is avoiding bills exceeding the ceiling value, third is acting as if the patient has gone home and immediately submitting a claim and fraud is carried out by demanding own expenses[2].

This study proposes a method based on outlier detection in clustering data claims. Outlier detection helps to find data vastly different from the normal ones. Because of the variety of combined variables from the dataset, we used data mining techniques to create cluster and detect outlier. Two methods were used in this study to form clusters, DBSCAN and K-Means. To find out the outliers in the cluster, Local Outlier Factor (LOF) was added.

Our main contribution is summarized as follows: first, based on the proposed method, clustering can be easy to make and effective using real-world data. Second, applying Local Outlier Factor (LOF) can greatly improve the efficiency of outlier detection and identification of which variable high cause outlier.

There is research about fraud detection in insurance claims around the world. In support of this research, previous work was collected as based to proposed experiment and chosen method for detection of outlier.

The graphical analysis method, based on network science and graph theory, was developed for the detection of fraud, waste, and abuse (FWA) in health services. This study focuses on finding four anomalies in the graph, namely: suspicious individuals, suspicious relationships, temporal anomalies of changes and geospatial characteristics, and structure. This study takes data from health services that display a graph of heterogeny and identify anomalies that arise from individual data, and relationships between services and communities originating locally and globally. From the research results, it was identified millions of dollars that can be returned to Xerox Services[3].

The fraud detection method uses the anomaly method which is divided into two parts. The first part is the improved local outlier factor (imLOF) based on the spatial density algorithm and is the development of the Simple Local Outlier Factor (simLOF) with DBSCAN algorithm. The second part is Robust Regression which connects the linear relationship between variables. The proposed model, imLOF, can detect inappropriate medical services in hospitals. In addition, Robust Regression can detect illegal claims[4].

Research trials were conducted using the random forest algorithm and logistics regression algorithm in detecting fraud. The test results show that the random forest and logistics regression algorithms can detect fraud in insurance claims better than using the linear regression method[5].

The method used to detect fraud is the pairwise comparison method of analytic hierarchical processing (AHP). The focus on existing problems by detecting using outlier probabilities. Based on an investigation of all relevant information the true positive rate varies between 89.5% and 94.5%. The results of this figure reveal that the sensitivity to the proposed framework is very high and can detect most of the inappropriate data in the data set. Alternatively, specifically the true negative value, varies between 52.8% and 83.4%[6].

Detect outliers to detect fraud in claim payment data with two approach methods. The first method applied is Multivariate Adaptive Regression Splines aimed at producing residual trained data and using residuals as input to detect univariate outliers based on Bayes using probabilistic programming. The development of a method by combining variables for detection and credible intervals, to be able to detect suspicious activities that arise from the smallest outliers[7].

The experiment model used to detect fraud is the SSIsomap activity clustering method, SimLOF outlier detection method, and the Dempster-Shafer Theory. The data used comes from 40,000 patients with more than 40 million health insurance claims. Based on the experimental results, the proposed model can detect fraud in mobile health services[8].

Classifying fraud into two categories, namely claims and anomalous diseases. Anomalous claims are based on the results of outlier identification by statistical process analysis. Anomalous disease is based on associations between variables and frequent patterns. The proposed model is tested using medical data to eliminate fraud that occurs at each layer by including the variables involved, such as policy provider, disease, etc. The results show that the proposed model can detect fraud [9].

The method tested is the cluster score method which is the development of the semi-supervised learning method to detect fraud. This method involves specifically, the methodology involves the transmission of an unsupervised model into a supervised model using Cluster-Scoremetric, which defines the object boundary between evaluating the homogeneity of abnormalities in the cluster construction. So that it can increase the number of detected fraudulent claims and reduce the proportion of investigated claims that are not actually fraudulent[10].

The experiment to detect fraud using the Decision Support System (DSS) which automatically processes claims and the Genetic support vector machines (GSVMs) method is applied in processing data originating from hospitals in Ghana to detect fraud. The results of the study found that GSVM was able to detect and classify fraud. Faster processing time when processing claims and increasing classification accuracy based on linear SVM classification (80.67%), polynomial (81.22%) and kernel radial basis function (RBF) (87.91%)[2].

Use of Decision Tree, Bagging, Random Forest and Boosting models to detect fraud in health insurance. The performance of the proposed model is evaluated based on accuracy, error rate, sensitivity, and specificity. The best results were obtained using the Bagging method [11].

The paper is organized as follows. The literature review and theory based on this research are described in part one. Part two presents the methodology used in this study. An in-depth analysis of the test results is presented in the third section. The fourth section contains conclusions and future work.

II. METHODOLOGY

Data mining is a series of processes to explore added value in the form of information that is not yet known manually from a database. The resulting information is obtained by extracting and recognizing important or interesting patterns from the data from database[12]. Knowledge or information obtained from processing data that is collected periodically using certain methods is defined as data mining[13]. The resulting information is a collection of several stored data. Data mining is part of the natural development of information technology. By using data mining techniques, data patterns can be identified from large datasets in the field of health insurance in search of important information[14]. Data mining techniques used to detect fraud are divided into supervised and unsupervised. Usually, random forest algorithms and logistic regression are used to detect fraud in health insurance claims. Data mining automatically filters based on suitable or unknown patterns to get new perceptions that help the possibility of fraud[5].

The definition of outlier in data mining is an abnormality, deviation, or anomaly. Outliers contain useful information about the abnormal characteristics of the system process in data generation[15]. To detect fraud, an effective way is to detect outliers. Data deviations from the pattern can usually be known to reduce losses that can occur [6].

According to experiment to find outlier using two techniques, first, classical outliers are used to analyze transaction datasets which consist of a compilation of various variables. Most of the classical outlier techniques are used for business transaction research. The algorithm used in the Classical Outlier approach is a statistical approach, namely the formation of outliers by means of a statistical approach. density-based allocation information for low-density areas. The second outlier detection is spatial outlier, which is the identification of outliers by considering the closest value in the population with a size that is not too large. There are two approaches that include spatial outliers: the space-based approach is to calculate the Euclidean distance as a description of the spatial value of neighboring points and the graph-based approach uses a graph connection to describe the neighboring spatial points [16]. To achieve the research objectives and design flow for experiment, it is necessary to design a model according to the proposed method.

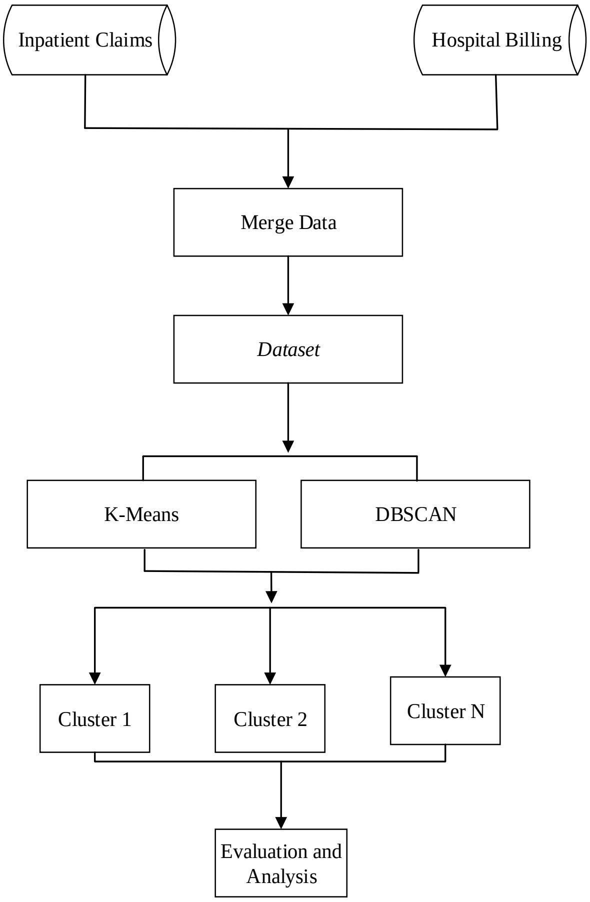


Figure 1. Proposed model

According to Figure 1, the experiment was conducted using datasets from two different sources. Inpatient Claims from insurance application, consist of 14 variables, such as registration code, membership code, participant name, registration date, total claim, claim rate code, etc. The second data source is Hospital Billings, which focus on inpatient bills, and consist of 25 variables, such as registration number, treatment period, patient medical record number, total cost of care, etc. Period data is from January to December 2020.

To combine the two tables derived from Inpatient Claims and Hospital Billing, identification unique variables contained in both tables. From the identification results, the connection between the two data is the patient registration code.

From the combination of the two tables with a total of 39 variables. 14 variables from inpatient claims and 25 variables from hospital billing, the total data is 12817 rows. Incomplete data was removed from the dataset, the total data rows were reduced to 12724 rows. There were 8134 rows of data that matched the claim and 3593 rows that were suspected of being fraudulent. The dataset that will be tested using the proposed method is 3593 rows with variable reduction to 6 variables, consisting of the participant registration code during the treatment period, patient discharge method, patient condition when leaving the hospital, code of claim rates paid, case severity affected by comorbidities. or complications during the treatment period and the length of the patient's treatment period

DBSCAN algorithm automatically identifies deviant data. The result of cluster formation by DBSCAN is knowing the pattern of the dataset and the relationship between data based on the density of objects in the cluster [17]. The formation of clusters using K-Means is based on determining the number of clusters that you want to form from the dataset[13]. So, this research is proposed to use DBSCAN and K-Means methods to find out the right method in forming clusters and identifying outlier data in the clusters that are formed.

To evaluate the proposed method based on the Precision, Recall and F measure (F1 or F-score) values[18][19]

Accuracy: a performance measure that will give the level of accuracy of the overall data, by calculating:

$$Accuracy = \frac{TP}{Total\ Rows} \tag{1}$$

Precision: shows the comparison between the number of True Positive (TP) with the number of True Positive (TP) and False Positive (FP), with the following calculation:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

Recall: shows the comparison between the number of True Positive (TP) with the number of True Positive (TP) and False Negative (FN), with the following calculation:

$$recall = \frac{TP}{TP + FN} \tag{3}$$

F-measure (F1 or F-score): harmonic mean value of precision and recall:

$$F = \frac{2\ x\ precision\ x\ recall}{precision + recall} \tag{4}$$

III. RESULTS AND DISCUSSION

The dataset that comes from the merging of two data sources, 6 variables were selected:

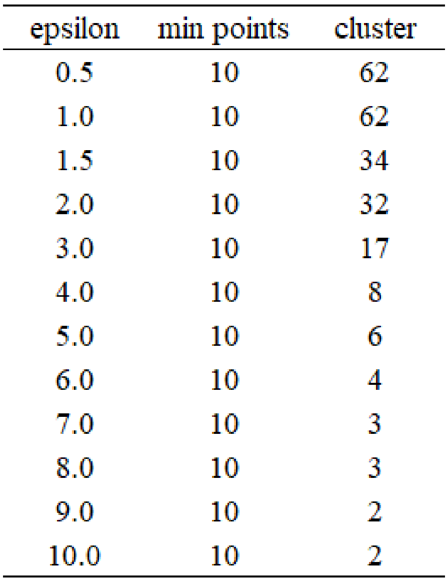
1) V1: Registration code during the treatment period
2) V2: How to get out of the patient
3) V3: Patient's condition when leaving the hospital
4) V4: Code of claims
5) V5: The severity of the case is influenced by the presence of comorbidities or complications during the treatment period
6) V6: The length of the patient's treatment period

V1 variable is used as an ID when export to RapidMiner application. Detailed data, it can be easily identified through the ID contained in the dataset.

Clustering Using DBSCAN

There are many techniques in carrying out the data mining process, including clustering [8]. Clusters created by the Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm are based on density-based data[20]. DBSCAN algorithm about clusters is based on the density range. DBSCAN searches for core objects that have neighboring densities, then core objects are connected to other objects to form dense groups [21]. Referring to previous research, two methods were used to detect outliers in health insurance claims, using DBSCAN by combining the improved local outlier factor (imLOF) algorithm and robust regression. The test results of the imLOF algorithm can detect the value of inpatient care claims and long-term care claims, while robust regression detects claims that do not match the bill of care [4]. The difference from this study, this study uses a dataset that is a combination of hospital bills and claim payments. DBSCAN method used in this study is to determine the outlier data from the resulting cluster formation. The results of the cluster using DBSCAN method, the smaller of epsilon, more difficult to create cluster. Data that is difficult to group is considered an outlier because it does not match the number of points that make up the area density. On the other hand, if the epsilon value is too large, some data will merge into one cluster. Using DBSCAN method, epsilon value starts from 0.5 to 10 and min point = 10.

Table 1. Clustering using DBSCAN



epsilon	min points	cluster
0.5	10	62
1.0	10	62
1.5	10	34
2.0	10	32
3.0	10	17
4.0	10	8
5.0	10	6
6.0	10	4
7.0	10	3
8.0	10	3
9.0	10	2
10.0	10	2

In Table 1, the experiment was carried out with min points = 10. The epsilon value was entered starting from 0.5 to produce 62 clusters, the epsilon value was changed to 1.0, the number of clusters remained at 62. When epsilon increased by 1.5, the number of clusters decreased to 34 and the higher the epsilon value was 10.0, the fewer clusters are formed, which is 2.

The next test is to find out the results of outlier detection in the cluster that is formed, using local density calculations. In comparison between local densities with one another, it can be seen areas of similar density and lower groupings than others. This concept is called an outlier. The test was carried out with DBSCAN using an epsilon value of 10.0, with the result that the cluster formed was 2, connected to the detect outlier (LOF) parameter in RapidMiner.

Table 2. Outlier detection using DBSCAN and LOF

Cluster	DBSCAN	
	Items	Outlier
Cluster 0	10	10
Cluster 1	3583	2719

In Table 2, DBSCAN detects that data outliers in Cluster 1 are 10 items and Cluster 2 are 2719 items. The total outlier data in both clusters is 2729. To find out which outlier data is included in the cluster formed, the data search is as follows:

Table 3. Outlier data identification using DBSCAN and LOF

	DBSCAN		
Cluster	V3: Patient's condition when leaving the hospital	V4: Code of claim	V6: The length of the patient's treatment period
Cluster 0	0	3	7
Cluster 1	109	132	2478
Total	109	135	2485

According to Table 3 data, based on the distribution of outlier data using DBSCAN method, there are three variables that cause outliers, namely V3 (Patient's condition when leaving the hospital), V4 (Code of claims rates paid), V6 (Long period of patient care). Variable V3 (State of the patient when leaving the hospital) based on cluster formation using DBSCAN, is in Cluster 1 with a total of 109. Variable V4 (Code of the claim rate paid) in Cluster 0 is 3 and Cluster 1 is 132. Variable V6 (Long period of time) patient care) in Cluster 0 amounted to 7 and Cluster 1 amounted to 2478.

Variable category V6 (Length of Patient Care Period) 10 highest order is 3 days totaling 327, 4 days totaling 312, 7 days consist of 255 items, 241 items in 5 days, 224 items in 6 days, 174 items in 8 days, 144 items in 9 days, 140 items in 2 days and 10 days with 130 items.

The second-order that causes the data to be outliers is the variable V4 (Code of rates for claims paid) totaling 135. It consists of code rates for claims that are not normally billed.

Finally, the variable V3 (the condition of the patient when leaving the hospital) caused the data to become outliers, totaling 109. The highest category of the variable V3 (the condition of the patient when leaving the hospital) was Improved, the total data was 40. The next sequence was Not Recovered with 24, Died < = 48 Hours totaling 14, Death < 8 Hours totaling 11, Death > 48 Hours and Recovering totaling 7, Died > 8 Hours and Moving Classroom totaling 3.

Clustering Using K-Means

K-Means method defines the cluster as the average value of the points in the cluster, an object assigned to the most similar cluster[13]. The clustering algorithm is grouping similar objects into one cluster, K-Means method increases the cluster results by calculating the distance and number of clusters [9][22]. K-Means algorithm has two steps, the first is to determine the number of clusters (k), the second is to identify the closest object to the cluster [21]. K-Means method is used to cluster data randomly and repeatedly. The difference with DBSCAN, in K-Means model, the cluster to be formed is determined. K-Means can separate attributes well and work efficiently [14]. To detect fraud in insurance claim data, a trial was conducted using statistical decision rules and k-means on the detection of period-based claim anomaly outliers and association rule-based mining with Gaussian distribution applied to disease-based anomaly outlier detection. The proposed method can identify fraudulent claims on insurance data[9].

In this study, the second test uses the proposed method, namely K-Means, the initial parameters are k: 2, max runs:10, measure types: MixedMeasures, mixed measure: Mixed EuclideanDistance.

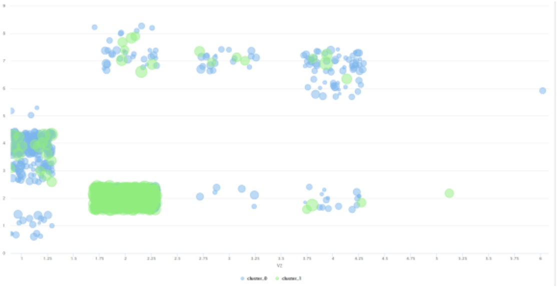


Figure 2. Cluster results using K-Means with settings k= 2, max runs :10, measure types: mixedmeasures, mixed measure: mixedeuclideandistance.

Initial testing with k:2, grouping each item is cluster 0 totaling 3057 and cluster 1 totaling 536. The test is carried out by changing the value of k based on the cluster results generated by DBSCAN method. To find out the right cluster formation, the Davies Boulden Index (DBI) is used. The calculation of the DBI value is based on the validation of the results of data grouping using the quantity and features contained in the dataset[23]. The lowest DBI value will indicate the best cluster formation.

Table 4. DBI value from experiments using K-Means

K	DBI
2	0.498
3	0.525
4	0.536
6	0.572
8	0.619
17	0.642
32	0.819
34	0.811
62	0.894

From Table 4, the results of the cluster experiment by referring to the results of DBSCAN method cluster, the higher the value of k, the higher the DBI value will be. This means that there are outliers that affect the sensitivity to cluster formation. The use of K-Means method is easier to use and makes it easier to create clusters.

The next test, K-Means method, was connected to the detect outlier (LOF) parameter in RapidMiner.

Table 5. Outlier detection using K-Means and LOF

Cluster	K-MEANS	
	Items	Outlier
Cluster 0	3057	2193
Cluster 1	536	536

In Table 5, the results of the cluster using K-Means, resulted in Cluster 0 totaling 3057 items and Cluster 1 totaling 536 items. Outlier data in Cluster 1 is 2193 items and Cluster 2 is 536 items. The total outlier data in both clusters is 2729. To find out the pattern of outlier data included in the cluster that is formed, the data search is as follows:

Table 6. Identification of outlier data using K-Means and LOF

Cluster	V3: Patient's condition when leaving the hospital	V4: Code of claim	V6: The length of the patient's treatment period
Cluster 0	80	43	2070
Cluster 1	29	92	415
Total	109	135	2485

According to Table 6 data, the total distribution of outlier data in both clusters, variable V3 (State of the patient when leaving the hospital) based on cluster formation using K-Means, is spread over two clusters formed, Cluster 0 is 80 and Cluster 1 is 29. Variable V4 (Code of the claim rate paid) in Cluster 0 is 43 and Cluster 1 is 92. Variable V6 (Long period of patient care) in Cluster 0 is 2070 and Cluster 1 is 415.

The first order of variables causing outliers is V6 (Long period of patient care) with a total of 2485. The highest order is 3 days totaling 327, 4 days totaling 312, 7 days totaling 255, 5 days totaling 241, 6 days totaling 224, 8 days totaling 174, 9 days totaling 144, 2 days totaling 140 and 10 days totaling 130.

The second-order that causes the data to be outliers is the variable V4 (Code of rates for claims paid) totaling 135 data. Is the claim rate code that appears first.

Finally, the variable V3 (Patient's condition when leaving the hospital) caused the data to be outliers, totaling 109 in both methods. The highest category of variable V3 (Patient's condition when discharged from the hospital) is Improved, the total data is 40. The next sequence is Not Healed with 24, Died <= 48 Hours totaled 14, Died < 8 Hours totaled 11, Died > 48 Hours and Healed totaled 7, died > 8 hours and moved classrooms totaled 3.

Spread of Outlier Variables in Both Methods

From experiments creating two cluster using two methods, setting for each method:

Table 7. Setting for creating method

Method	Parameter	Cluster 0	Cluster 1
DBSCAN	epsilon: 9.0 min points: 10	10 items	3583 items
K-MEANS	k: 2 max runs: 10 measure types: MixedMeasures mixed measure: MixedEuclideanDistance	3057 items	536 items

According to table 7, the result for creating cluster each method, total items for each cluster is different. Cluster using DBSCAN, Cluster 0 consist of 10 items and Cluster 1 consists of 3583 items. In K-Means, Cluster 0 consists of 3057 items and Cluster 1 consists of 536 items. To find out the best grouping of the proposed method, the criteria for cluster count performance and item distribution performance are used:

Table 8. Setting for creating method

Parameter	DBSCAN	K-Means
Cluster Count Performance	0.999	0.999
Item Distribution Performance	0.994	0.746

From the results of table 8, score for Cluster Count Performance using two methods was same 0.999. K-Means is more precise in classifying data based on Item Distribution Performance; the score is 0.746 compared to DBSCAN score is 0.994.

The next experiment, to find out the results of outlier detection in both methods, using local density calculations. Both methods connect to LOF, here is a result comparison of the outlier detection:

Table 9. Result outlier using LOF

	DBSCAN		K-MEANS	
Cluster	Items	Outlier	Items	Outlier
Cluster 0	10	10	3057	2193
Cluster 1	3583	2719	536	536

In table 9, DBSCAN detects data outliers that cause the number of clusters to match the differences in existing items. K-Means cannot detect outliers in the dataset, so it is assumed that the data are all similar and are grouped according to the similarity of the data.

3.1 Evaluation

To evaluate the two methods, f-measure calculations are used by calculating TP, FP, FN through the Confusion Matrix. Here is a Confusion Matrix based on DBSCAN method.

Table 10. Confusion matrix using DBSCAN

	Prediction				
Actual	V2	V3	V4	V5	V6
V2	6	6	36	2	10
V3	1	5	6	3	3
V4	1	1	93	3	2
V5	1	1	1	3	2
V6	1	1	1	3	2

Based on table 10, the confusion matrix uses DBSCAN method, the accuracy value is 0.006067353, precision is 0.387, recall is 0.391 and f-measure value is 0.389.

The next method to calculate the f-measure is K-Means method:

Table 11. Confusion matrix using K-Means

	Prediction				
Actual	V2	V3	V4	V5	V6
V2	6	6	36	2	10
V3	2	8	18	2	2
V4	4	7	333	2	2
V5	2	2	331	3	12
V6	2	2	35	3	52

Based on table 11, the confusion matrix is based on K-Means method, the accuracy value is 0.02238, precision is 0.374, recall is 0.411 and f-measure value is 0.391.

DBSCAN accuracy calculation is 0.006067353, higher K-Means is 0.02238. The f-measure calculation using K-Means method is

0.391, 0.003 higher than DBSCAN score, which is 0.389.

IV. CONCLUSION

Based on this research by testing the proposed data mining model, it can be concluded as follows:

1. Both methods can detect outlier data and distribute outlier data in the formed cluster. DBSCAN and K-Means can distribute outlier data into clusters that are formed. K-Means spread outliers divided according to the cluster formed.
2. The total outlier data is 2729. The variable that influences the outlier data is V6 (Long period of patient care) totaling 2485 with a treatment period of 3 days totaling 327. The second variable that affects the outlier data is V4 (Code of claims) totaling 135 data and the Improved category totaling 40. The last variable that influences outlier data is V4 (Code of claims) totaling 109 data, which is the claim rate code that is being claimed for the first time.
3. DBSCAN accuracy calculation is 0.006067353, higher K-Means is 0.02238. The f-measure calculation using the K-Means method is 0.391, 0.003 higher than the DBSCAN value, which is 0.389. To simplify data grouping, it is necessary to identify data variables that can be used as labels.

This research shows that outlier behaviors can be detected using datamining method. In the future work, we intend to add some variable, such as code of disease, cost claims and hospital cost. In addition, this experiment was run using old data and offline environments. So, it will be challenging to adapt to an automated method using new-coming data.

BIBLIOGRAPHY

[1] Y. B. Sarwo, "Tinjauan Yuridis Terhadap Kecurangan (Frauds) Dalam Industri Asuransi Kesehatan Di Indonesia," *J. Kisi Huk. Unika*, vol. 14, no. 1, pp. 1–15, 2015.

[2] R. A. Sowah *et al.*, "Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)," *J. Eng. (United Kingdom)*, vol. 2019, no. January 2007, 2019, doi: 10.1155/2019/1432597.

[3] J. Liu *et al.*, "Graph analysis for detecting fraud, waste, and abuse in health-care data," *AI Mag.*, vol. 37, no. 2, pp. 33–46, 2016, doi: 10.1609/aimag.v37i2.2630.

[4] W. Zhang and X. He, "An Anomaly Detection Method for Medicare Fraud Detection," *Proc. - 2017 IEEE Int. Conf. Big Knowledge, ICBK 2017*, pp. 309–314, 2017, doi: 10.1109/ICBK.2017.47.

[5] N. Ghuse, P. Pawar, and A. Potgantwar, "An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques," no. 5, pp. 27–32, 2017, [Online]. Available: www.ijsrnsc.orgAvailableonlineatwww.ijsrnsc.org.

[6] M. S. Anbarasi and S. Dhivya, "Fraud detection using outlier predictor in health insurance data," *2017 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2017*, no. Icices, 2017, doi: 10.1109/ICICES.2017.8070750.

[7] R. A. Bauder and T. M. Khoshgoftaar, *Multivariate outlier detection in medicare claims payments applying probabilistic programming methods*, vol. 17, no. 3–4. Springer US, 2017.

[8] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, and B. Gong, "An Efficient Fraud Identification Method Combining Manifold Learning and Outliers Detection in Mobile Healthcare Services," *IEEE Access*, vol. 6, no. c, pp. 60059–60068, 2018, doi: 10.1109/ACCESS.2018.2875516.

[9] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," *2017 10th Int. Conf. Contemp. Comput. IC3 2017*, vol. 2018-Janua, no. August, pp. 1–7, 2018, doi: 10.1109/IC3.2017.8284299.

[10] S. M. Palacio, "Abnormal pattern prediction: Detecting fraudulent insurance property claims with semi-supervised machine-learning," *Data Sci. J.*, vol. 18, no. 1, pp. 1–15, 2019, doi: 10.5334/dsj-2019-035.

[11] R. Kunickaitė, M. Zdanavičiūtė, and T. Krilavičius, "Fraud detection in health insurance using ensemble learning methods," *CEUR Workshop Proc.*, vol. 2698, 2020.

[12] R. T. Vulandari, *Data Mining Teori Dan Aplikasi Rapidminer*, I. Yogyakarta: Gava Media, 2017.

[13] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts And Techniques*, Third. Morgan Kaufmann, 2011.

[14] S. Kareem, R. B. Ahmad, and A. B. Sarlan, "Framework for the identification of fraudulent health insurance claims using association rule mining," *2017 IEEE Conf. Big Data Anal. ICBDA 2017*, vol. 2018-Janua, pp. 99–104, 2018, doi: 10.1109/ICBDAA.2017.8284114.

[15] C. C. Aggarwal, *Outlier Analysis*, Second., vol. 24, no. 2. Yorktown Heights, New York: Springer US, 2017.

[16] R. Bansal, N. Gaur, and S. N. Singh, "Outlier Detection: Applications and techniques in Data Mining," *Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016*, pp. 373–377, 2016, doi: 10.1109/CONFLUENCE.2016.7508146.

[17] A. R. Ajiboye, A. G. Akintola, and A. O. Ameen, "Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm," *African J. Comput. ICT*, vol. 8, no. 1, pp. 39–46, 2015.

[18] A. M. Khalimi, "Perhitungan Confusion Matrix Multi-Class Clasification 3x3," 2020. https://www.pengalaman-edukasi.com/2020/11/menghitung-confusion-matrix-3-kelas.html (accessed Dec. 10, 2021).

[19] V. Mallawaarachchi, "Evaluating Clustering Results," 2020. https://towardsdatascience.com/evaluating-clustering-results-f13552ee7603 (accessed Dec. 10, 2021).

[20] T. Wahyono, *Fundamental Of Python For Machine Learning*, I. Gava Media, 2018.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "Density-Based Clustering Methods," *Compr. Chemom.*, vol. 2, pp. 635–654, 2009, doi: 10.1016/B978-044452701-1.00067-3.

[22] P. Patel, S. Mal, and Y. Mhaske, "A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques," *Int. Res. J. Eng. Technol.*, vol. 6, no. 1, pp. 591–594, 2019, [Online]. Available: http://www.academia.edu/download/58335030/IRJET-V6I1104.pdf.

[23] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012128.