IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.15, No.4, October 2021, pp. 391~402
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.68160

Anomaly Detection in Hospital Claims Using K-Means and Linear Regression

Hendri Kurniawan Prakosa.*[1], Nur Rokhman[2]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia,
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: *[1]**hendri.k@mail.ugm.ac.id**, [2]nurrokhman@ugm.ac.id

Abstrak

BPJS Kesehatan yang telah berdiri selama hampir satu dekade masih mengalami defisit dalam proses penjaminannya terhadap peserta. Salah satu faktor yang menyebabkan hal tersebut adalah adanya ketidaksesuaian pada proses klaim yang cenderung merugikan BPJS Kesehatan. Misalnya saja dengan menaikkan coding diagnosa sehingga klaimnya menjadi lebih besar, melakukan klaim ganda dan pencatatan klaim palsu. Tindakan tersebut berdasarkan peraturan pemerintah termasuk pada tindakan fraud. Tindakan fraud dapat dideteksi dengan melihat adanya anomali yang muncul pada data klaim. Selain itu anomali juga dapat digunakan untuk mengetahui perubahan pola pembiayaan dan perencannaan rumah sakit.

Penelitian ini bertujuan untuk mengetahui anomali data klaim rumah sakit kepada BPJS Kesehatan. Data yang digunakan adalah data klaim BPJS tahun 2015-2016. Sedangkan algoritma yang digunakan adalah kombinasi algoritma K-Means dan Regresi Linear. Agar hasil klasterisasi optimal digunakan algoritma density canopy pada penentuan centroid awalnya.

Evaluasi klaster menggunakan silhouete index menghasilkan nilai sebesar 0.82 dengan jumlah klaster 5 dan nilai RMSE hasil pemodelan simple linear regression sebesar 0.49 untuk biaya tagih dan 0.97 untuk variabel lama dirawat. Berdasarkan hasil pemodelan tersebut terlihat titik anomali sebanyak 435 dari 10.000 data atau sebesar 4.35 %. Diharapkan dengan diketahuinya anomali tersebut dapat dilakukan tindak lanjut yang lebih efektif.

Kata kunci*—Deteksi, Anomali, BPJS-Kesehatan, K-Means, Regresi Linear*

Abstract

BPJS Kesehatan, which has been in existence for almost a decade, is still experiencing a deficit in the process of guaranteeing participants. One of the factors that causes this is a discrepancy in the claim process which tends to harm BPJS Kesehatan. For example, by increasing the diagnostic coding so that the claim becomes bigger, making double claims or even recording false claims. These actions are based on government regulations is including fraud. Fraud can be detected by looking at the anomalies that appear in the claim data.

This research aims to determine the anomaly of hospital claim to BPJS Kesehatan. The data used is BPJS claim data for 2015-2016. While the algorithm used is a combination of K-Means algorithm and Linear Regression. For optimal clustering results, density canopy algorithm was used to determine the initial centroid.

Evaluation using silhouete index resulted in value of 0.82 with number of clusters 5 and RMSE value from simple linear regression modeling of 0.49 for billing costs and 0.97 for length of stay. Based on that, there are 435 anomaly points out of 10,000 data or 4.35%. It is hoped that with the identification of these, more effective follow-up can be carried out.

Keywords*—Detection, Anomaly, BPJS Kesehatan, K-Means, Linear Regression*

Received August 2nd,2021; Revised October 31th, 2021; October 31th, 2021

1. INTRODUCTION

The current era of National Health Insurance (JKN) encourages people to more easily go to health service facilities (Fasyankes) even in mild conditions. This has led to an increase in the number of participants seeking treatment, so the bills to be paid by BPJS Kesehatan are also increasing.

The number of bills originating from health facilities, especially hospitals, is not proportional to the amount of premiums paid by insurance participants so that BPJS Kesehatan runs a deficit. Through an analysis of BPJS Kesehatan expenses, it was found that the contribution income was always lower every year when compared to the expenses incurred. Figure 1 shows contribution income is always smaller than claims. In addition to this, the results from interviews with internal parties stated that the cause of the deficit at BPJS Health was also due to fraud [1].

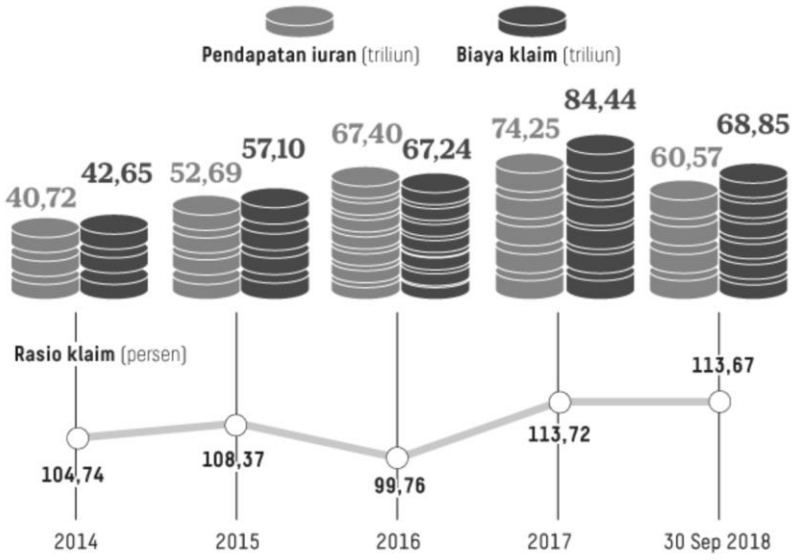


Figure 1 Contribution income and claim expenses incurred[2]

Insufficient funding (deficit) has broad implications, one of which is hospital claims that are late being paid to encourage health facilities to make claims that are not in accordance with procedures [2]. The inaccuracy of claims can potentially lead to fraud based on Pasal 5 Number 36 of the Regulation of the Minister of Health of the Republic of Indonesia concerning the Prevention of Fraud in the Implementation of the Health Insurance Program in the National Social Security System [3]. Fraud can be indicated by the presence of anomalies in the data. In addition to fraud, other things that can be detected from anomalies are changes in hospital financing patterns, assessment of the quality of health facility services, health fund investment planning and logistical supply problems[4].

BPJS Kesehatan has released BPJS Kesehatan Data Sampel Tahun 2015-2016 for use by the public in scientific works[5]. The available data has a large volume and consists of many unlabeled variables. So we need the right unsupervised learning algorithm on big data to find out any anomalies[6]. In this study, an unsupervised learning algorithm, namely K-Means, was used to determine the data cluster as the basis for determining the anomaly point with Linear Regression modeling.

Several studies have used K-Means to detect anomalies with good accuracy, even one of them with a hybrid method can be used to detect fraud on credit cards [6, 7]. Optimization of K-Means has also been carried out at the initial centroid point based on distance density [8]. In fact, anomalies that appear are not only in numerical data, but also in categorical data. Anomaly detection in numeric and categorical data (mixed attribute data) can be classified into 4 types, namely: categorized, enumerated, combined and mixed [9].

2. METHODS

In this study, a combination of the K-Means algorithm and Linear Regression was carried out to determine the anomaly point in the data. Before doing the modeling, the steps that must be done is the data preprocessing. The data preprocessing includes feature selection, codification, variable creation and normalization. After the normalization process is carried out, the first experiment begins by carrying out a dimension reduction process using the Principal Component Analysis (PCA) method. The results of the PCA will be the basis for finding the initial centroid point with the density canopy algorithm. The experiment was conducted by comparing the results of the K-Means cluster using the density canopy and the random method in determining the initial centroid point. The best cluster results will then become a reference for the anomaly cluster. The best cluster value is determined by the silhouette index which is close to 1.

After achieving the best value on the silouete index, then modeling with simple linear regression is carried out using the coordinates of the cluster to the verification cost variable (dependent variable). Because the regression used is 1 independent variable, the modeling is done 2 times, namely for x0 and x1 separately. Each regression model will be evaluated by calculating the Root Mean Square Error (RMSE) value. Based on the RMSE value, only then can the anomaly point in the data be known which is then confirmed with the cluster. The steps taken are shown in Figure 3.

2.1 Data Preprocessing

The data obtained consisted of membership data, visits to Primary Level Health Facilities (FKTP) for Capitation and Non Capitation and visit data for Advanced Level Referral Health Facilities (FKRTL). From some of these data, FKRTL data is taken because it contains information on hospital claims. The data is still raw and needs to be processed at the data preprocessing stage before modeling. The stages are as follows as shown in Figure 2.

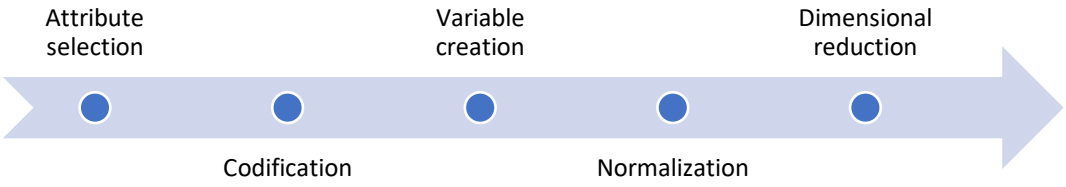


Figure 2 Data preprocessing stage

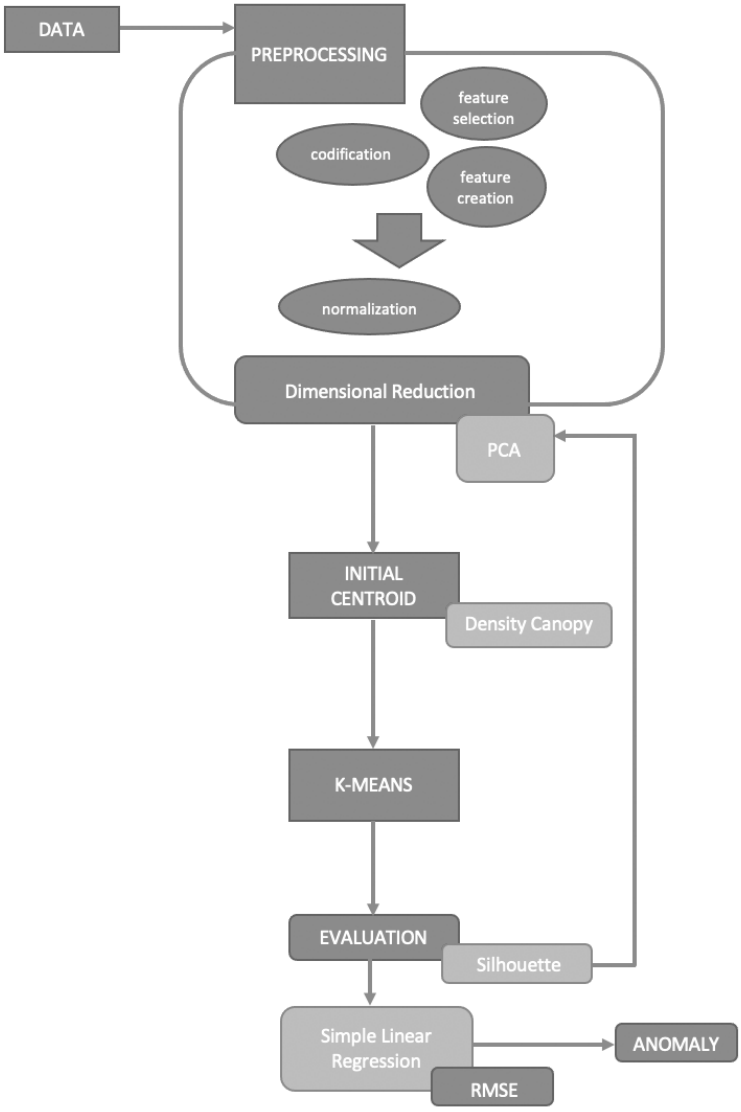


Figure 2 Step by step for research

2. 1.1 Attribute Selection

There are 53 variables in the FKRTL visit data so it is necessary to make a selection. The results of the selection were 10 variables, namely: arrival date (FKL03), return date (FKL04), number of procedures (FKL31), billing costs (FKL48), severity level (FKL23), class of care (FKL13), CMG code (FKL19), special procedure rates (FKL38), grouping rates (FKL33), prosthesis rates and verification costs (FKL49). These variables are considered to have an effect on the claim value, while the other variables only contain demographic data from patients and health facilities. Verification costs will be used in linear modeling as the dependent variable

2.1.2 Codification/Encoding

When processing input data of categorical data type, it is likely to convert a categorical variable to a numerical variable or a vector that its elements are numerical data type. Three common methods used are: 1) Label Encoding; 2) One-hot Encoding and its modification; 3) "Learned" Embedding encoding. In the Label encoding method each label of a categorical data

variable is assigned to a most suitable integer number[10].

Variables that require label encoding are severity level, treatment class, CMG code. The process is convert the variable value in to suitable integer number. For example, the severity level variable, which contains the values "Level I", "Level II" and "Level III" is changed to values 0, 1,2 and 3. Likewise with the treatment class variable and CMG code.

2.1.2 Variable Creation

Variable creation is the process of forming new variables based on existing variables or can be said to be derived variables. Like the age variable which is derived from the date of birth variable. Similarly, the variable length of stay can be obtained from the variable date of arrival and date of return which is sought for the difference.

2.1.3 Normalization

The data obtained has a wide range of variations. For example, the old data being treated has a range of 1 or 2 digits only, but for billing costs it can be up to 6 digits (millions). This will cause problems during the analysis process. Therefore, by using the data transformation method, the data can be normalized. This normalization process makes all variables have the same range. There are several methods that can be used for data normalization, namely min-max normalization, z-score standardization or decimal scaling standardization [11].

Normalization of z-score based on the mean and standard deviation of a dataset. The equation can be seen in equation 2.

$$Z(x_{ij}) = \frac{x_{ij} - \overline{x_j}}{\sigma} \quad (2)$$

Where $\overline{x_j}$ is the mean of the data, x_{ij} is the original value and is the standard deviation.

2.1.4 Dimensional reduction

The K-Means algorithm will not work optimally at high dimensions so it is necessary to do efficiency on features that are not correlated. Principal Component Analysis (PCA) algorithm is one method to reduce features from high to low dimensions [12]. One of the functions of data reduction with PCA is that it can be seen that the variables that have a high variation are considered to have an effect on the dataset. PCA with no reduction resulted in the ratios as shown in table 1.

Table 1 Component ratio with PCA without any variable reduction

Component	Ratio Variation
0	0.45577302
1	0.12946284
2	0.12112074
3	0.10601911
4	0.07950325
5	0.03346452
6	0.03346452
7	0.00188621
8	0.0

2.1.5 Initial Centroid with Density Canopy

The conventional K-Means algorithm has a weakness that lies in the initial centroid selection process which is carried out randomly. Determination of the initial centroid is very sensitive to its influence on the quality of cluster accuracy and computation time. So that by selecting the right centroid, the K-Means algorithm can be more optimal [13].

Many algorithms have been developed to determine the initial centroid, the density canopy algorithm is one of the algorithms that can optimize the selection of the initial centroid because of its superiority, which is robust against outliers. Besides that, it can reduce the interference of annoying points and can reduce iterations in the K-Means process and can also avoid selecting the same points in determining the centroid [14]. In general, the density canopy algorithm will choose the centroid that has the highest number of neighbors based on the Euclidean distance [15].

2.2 K-Means

K-Means is one of the unsupervised learning algorithms for model clustering. Compared to the DBScan algorithm, K-Means has better performance on high-dimensional data [16]. In addition, K-Means also has a higher silhouette coefficient score than DBScan [17]. K-Means algorithm is a machine learning algorithm that can sort data into several clusters. So that data that is outside the cluster can be seen as an anomaly (abnormal). The way the K-Means algorithm works in general is to form the initial centroid in a number of clusters which is then recalculated to determine the next centroid in the clusters that are formed. So that the important point in the algorithm is at the initial centroid.

In this study, after knowing the initial centroid and the number of clusters, the next step is to build a cluster model using the K-Means algorithm. The program code line is shown in Figure 4. The modeling uses the library from sklearn which has been provided by the juptyter netbook tools. With the parameter n_cluster is the number of clusters (k) and init is the initial centroid (ncc).

```
from sklearn.cluster import KMeans

k=len(canopy_centroids)
XX = X.to_numpy()
E = XX.tolist()

new_canopy_centroids = []
for x in canopy_centroids:
   new_canopy_centroids.append(canopy_centroids[x].tolist())
ncc = np.array(new_canopy_centroids)

kmeans = KMeans(n_clusters = k, init = ncc, max_iter = 300, n_init = 1, random_state = 0)
kmeans.fit(E)
```

Figure 4 Code for K-Means using sklearn library

2.3 Linear Regression

Linear Regression is a data modeling on a straight line. This modeling is usually used to determine the relationship between the dependent variable (influenced) and the independent variable (which affects). In addition, linear regression is also often used to perform value prediction analysis. For example, a random variable y (response variable) can be modeled with a linear function by the random variable x, which is called predictor variable shown in equation 1.

$$y = b_1 x + b_0 \quad (1)$$

Where the variance of y is assumed to be response variable and the coefficients of b1 and b0 are called the regression coefficients and x is the predictor variable [10]. Each cluster that has been formed will then look for a straight line equation with a linear regression model as a basis for detecting anomaly data [18].

Root Mean Square Error (RMSE) is a method of measuring the difference in the value of an estimated value over the observed value. By using RMSE, the accuracy of a prediction model can be calculated. The smaller the RMSE value, the more accurate the model is. The line of program code to calculate the RMSE value is shown in Figure 5.

```
# Standalone simple linear regression example
from math import sqrt

# Calculate root mean squared error
def rmse_metric(actual, predicted):
        sum_error = 0.0
        for i in range(len(actual)):
                prediction_error = predicted[i] - actual[i]
                sum_error += (prediction_error ** 2)
        mean_error = sum_error / float(len(actual))
        return sqrt(mean_error)

# Evaluate regression algorithm on training dataset
def evaluate_algorithm(dataset, algorithm):
        test_set = list()
        for row in dataset:
                row_copy = list(row)
                row_copy[-1] = None
                test_set.append(row_copy)
        predicted = algorithm(dataset, test_set)
        print(predicted)
        actual = [row[-1] for row in dataset]
        print(actual)
        rmse = rmse_metric(actual, predicted)
        return rmse

# Simple linear regression algorithm
def simple_linear_regression(train, test):
        predictions = list()
        b0, b1 = coefficients(train)
        for row in test:
                yhat = b0 + b1 * row[0]
                predictions.append(yhat)
        return predictions

rmse = evaluate_algorithm(cluster3, simple_linear_regression)
print('RMSE: %.3f' % (rmse))
```

Figure 5 Code for Linear Regression

2.4 Anomaly Detection

Anomaly points can be detected with the condition if the residual value is greater than 2 times the RMSE value. The program code to determine the anomaly point is shown in figure 6.

```
# anomaly = y1-y > 2*RMSE
outlier = []
for index, row in table1_df.iterrows():
# iterrows() = fungsi perulangan di pandas dataframe
   if row['y1-y'] > (2*rmse) :
      outlier.append(index)
outlier
```

Figure 6 Code for anomaly detection

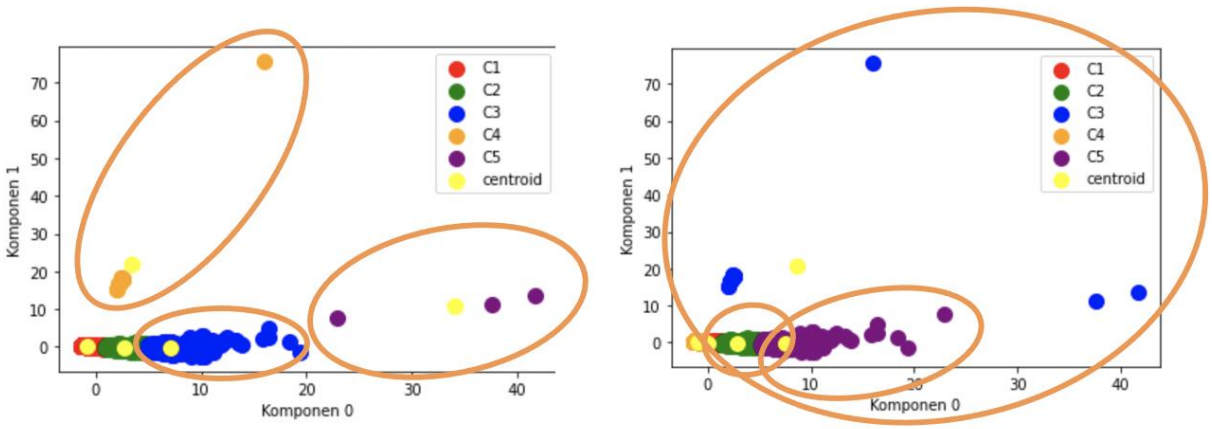
3. RESULTS AND DISCUSSION

The experiment was carried out 4 times to determine the best silhouette index value. By combining 2 methods of determining the centroid point and by using a reduction process or not. The experimental results can be seen in table 2. From 4 experiments, the silhouete index value did not change much, stable at 0.82 but the best number of iterations was obtained in the initial centroid formation process with canopy density and by reducing it using PCA, which was 11 times.

Table 2 The results of the cluster formation by the experiment

Experiment	PCA	Initial centroid	Number of Initial centroid	Number of iteration	Silhoete Index Value
1	Yes	DC	5	11	0.8251657382870
2	No	DC	5	21	0.8208625378520
3	No	Random	5	51	0.8219812055983
4	Yes	Random	5	37	0.6588083219868

Based on table 2, then the results of the cluster in the 1st experiment were chosen to be modeled using linear regression with the equation y = mx + c on x_0 and x_1. x_0 represents the billing cost and x_1 is the long-maintained variable with the variable y being the verification fee. Figure 7 shows the cluster results in experiments 1 and 4.



a) Density canopy K-Means b) Random K-Means

Figure 7 Result of cluster

From figure 7, it can be seen in Figure a that the boundary line of data grouping (in orange) is clearer than Figure B, the result of a random initial centroid. This shows that the results of clustering using a density canopy are better than the random method.

The anomaly point was determined by linear regression modeling. Because the linear regression used is using 1 independent variable, the regression modeling process is carried out 2 times, namely y to x_0 (billed costs) and y to x_1 (length of treatment). The result of the modeling process is the coefficient at x_0 is 0.455 and the coefficient at x_1 is 0.216. Each model is evaluated by calculating its RMSE value.

The RMSE number is used as a reference to determine data anomalies. The data is said to be anomaly if the residual value is more than 2x the RMSE value. Score residue is the value of the modeling results on the regression liner indicated by y_1 minus the factual value indicated by the variable y. Table 3 shows the acquisition of RMSE values in each variable.

Table 3 RMSE and number of anomalies per variables

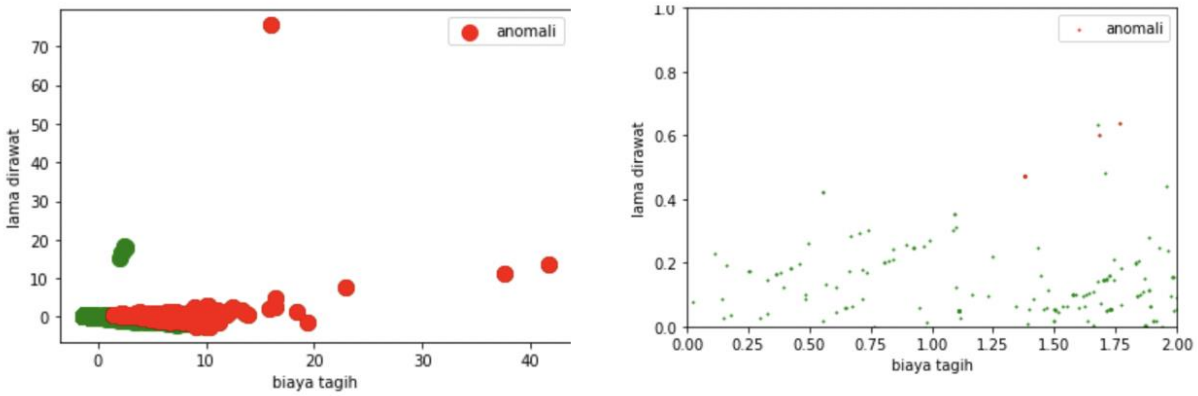
Equation	RMSE	Number of anomaly
$y = b_0 + b_1x_0$	0.4938563666139264	193
$y = b_0 + b_1x_1$	0.9755332739880637	260

Table 3 shows a summary of the RMSE values and the detected anomaly points. From the first and second modeling there are anomalous data slices. So by reducing the data slices, the total number of anomalous data is 435. The data is then cross-checked on the cluster formed in the previous process. The results are as shown in table 4.

Table 4 Number of anomaly in each cluster

Cluster	Number of Cluster Member	Number of Anomaly	Percentage
1	8020	0	0 %
2	1741	231	13.26 %
3	224	200	89.28 %
4	12	1	8.33 %
5	3	3	100 %
Total	10000	435	4.35 %

The following is figure 8 the result of visualization of anomaly points based on linear regression modeling with variable costs. Figure a is viewed at the x-axis and y-axis coordinates minus up to more than 40. Meanwhile, image b is a magnification of image a with a reduced range of coordinate lines again. By looking at the visualization of the cluster results in figures 7, 8 and the data in table 4, it is evident that cluster 5 is an anomaly cluster.



(a) (b)

Figure 8 Anomaly visualization of linear regression modeling results

4. CONCLUSIONS

The combination of Density Canopy K-Means algorithm and Linear regression can be used to detect data anomalies. From the 10,000 sample data used, 435 or 4.35% anomalies were found with the silhouette index value reaching 0.82 and the average RMSE value of 0.73 using billed cost and length of treatment. And than initial centroid using dencity canopy is better than randomly.

In the research that has been done there are several shortcomings, namely as follows: 1) In determining the variables in the variable selection process, its still use the assumptions of the researcher. So that future research is expected to use empirical methods whose validation can be measured. 2) Data utilization is still limited to 10,000 data. This amount of data cannot be categorized as big data, so it is necessary to develop a parallel programming method on the density canopy algorithm so that it can process data on a larger scale. 3) The data used are numerical and categorical data, so that in future research, mixed attribute anomaly detection methods can be used.

REFERENCES

[1] K. K. Firdaus and L. S. Wondabio, 2019, "Analisis Iuran Dan Beban Kesehatan Dalam Rangka Evaluasi Program Jaminan Kesehatan", *Jurnal ASET*, 1, 11, 147-158. [Accessed: 12-Mar-2020]

[2] A. Ramadhan, "79000 Klaim Rumah Sakit Berpotensi Anomali", https://kompas.id/baca/utama/2018/11/13/79-000-klaim-rumah-sakit-berpotensi-anomali, 13 November 2018. [Accessed: 13-Apr-2020]

[3] S. H. Tatik, "Pencegahan Kecurangan *(Fraud)* Dalam Pelaksanaan Program Jaminan Kesehatan Pada Sistem Jaminan Sosial Kesehatan (SJSN) di Rumah Sakit Umum Daerah Menggala Tulang Bawang", *Fiat Justisia*, 4, 10, 715-732, 2016. [Accessed: 12-Mar-2020]

[4] L.F.M. Carvalho, C.H.C. Teixeira, Meira, Wager Jr., Martin Ester, Osvaldo Carvalho and

M. H. Brandao, "*Provider-Consumer Anomaly Detection for Healtcare Systems"*, IEEE International Conference on Healthcare Informatics (ICHI), pp. 229-238, 2017 [online]. Available: https://doi.org/10.1109/ICHI.2017.[Accessed: 23-August-2021]

[5] J. Y. S. Ng, R. V. Ramadani, D. Hendrawan, D. T. Duc, P. H. T. Kiet, "National Health Insurance Databases in Indonesia, Vietnam and the Philippines", PharmacoEconomics-Open, 3, 517-526, 2019 [online]. Available: https://doi.org/10.1007/s41669-019-0127-2 [Accessed: 2-July-2021]

[6] Alguliyev, Rasim & Aliguliyev, Ramiz & Abdullayeva, Fargana, "PSO+K-means Algorithm for Anomaly Detection in Big Data", Statistics, Optimization & Information Computing, Vol. 7, 348-359, 2019 [online]. Available: https://doi.org/10.19139/soic.v7i2.623 [Accessed: 8-July-2020]

[7] S. G. Fashoto, O. Owolabi, O. Adeleye, J. Wandera, "Hybrid Method for Credit Card Froud Detection Using K-Means Clustering with Hidden Markov Model and Multilayer Perceptron Algorithm", British Journal of Applied Science & Technology, 13(5), 1-11, 2016 [online]. Available: https://doi.org/10.9734/BJAST/2016/21603 [Accessed: 12-Mar-2020]

[8] W. Yang, H. Long, L. Ma, H. Sun, "Research on Clustering Method Based on Weighted Distance Density and K-Means", Procedia Computer Science, 166, 507-5011, 2020 [online]. Available: https://doi.org/10.1016/j.procs.2020.02.056 [Accessed: 11-Apr-2020]

[9] N. Rokhman, "A Survey on Mixed-Attribute Outlier Detection Method", CommIT (Communication & Information Technology) Journal 13(1), 39-44, 2019 [online]. Available: https://doi.org/10.21512/commit.v13i1.5558 [Accessed: 8-Apr-2021]

[10] Do Thi Thu Hien, Cu Thi Thu Thuy, Tran Kim Anh, Dao The Son and Cu Nguyen Giap, "Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance", International Journal of Advanced Computer Science and Applications, 11, 10.14569/IJACSA.2020.0111135, 2020 [online]. Available: https://dx.doi.org/10.14569/IJACSA.2020.0111135 [Accessed: 6-July-2021]

[11] Larose, D.T., "Discovering Knowledge in Data Introduction to Data Mining*"*, Wiley Interscience, New Jersey, 2005.

[12] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin and Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction", LK, 3, 9, 192-201, 2018. [Accessed: 12-Aug-2020]

[13] K. Sirait, Tulus, E. B. Nababan, "K-Means Algorithm Performance Analysis With Determining The Value Of Starting Centroid With Random And KD-Tree Method", J. Phys.: Conf. Ser., 930, 2017. [Accessed: 2-July-2020]

[14] G. Zhang, C. Zhang, and H. Zhang, "Improvement of K-Means Clustering Algorithm Based on Density", Knowledge-Based Systems, 289-297, 2018 [online]. Available: https://doi.org/10.1016/j.knosys.2018.01.031. [Accessed: 17-July-2020]

[15] R. Ananda, "Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency", Khazanah Informatika, 2, 5, 158-168, 2019. [Accessed: 15-Jun-2020]

[16] A. Sreenivasulu, "Evaluation of Cluster Based Anomaly Detection", *Thesis*, Univesity of Skovde, Swedia, 2019 [online]. Available: http://www.diva-portal.org/smash/get/diva2:1382324/FULLTEXT01.pdf [Accessed: 5-July-2021]

[17] A. C. Muller & S. Guido, *Introduction to Machine Learning with Python A Guide for Data Scientist*, O'Reilly Media. USA, 2017.

[18] M. A. Mondal and Z. Reehena, "Road Trafic Outlier Detection Technique based on Linear Regression", Procedia Computer Science, 171, 2547-2555, 2020 [online]. Available: https://doi.org/10.1016/j.procs.2020.04.276. [Accessed: 18-May-2021]