Preprocessing Data Faktor-Faktor yang Mempengaruhi Penyakit Diare di Indonesia pada Tahun 2016 dengan menggunakan Analisis *Missing Value* dan *Outlier*

Fiika Arma'atus Syaani, Sekar Krismaya Weni, Bekti Indasari dan Santi Wulan Purnami
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember (ITS)
Jalan Arief Rahman Hakim, Surabaya 60111 Indonesia
email: bektindah77@gmail.com, santiwulan08@gmail.com

***Abstrak*—Indonesia merupakan negara dengan jumlah penduduk terbesar di Asia harus tetap memperhatikan persebaran berbagai jenis penyakit yang semakin berkembang agar dapat meningkatkan derajat kesehatan masyarakat. Salah satu jenis penyakit yang sering terjadi kepada masyarakat adalah diare. Penyakit ini merupakan penyakit endemis yang dapat dikategorikan sebagai penyakit potensial Kejadian Luar Biasa yang dapat berakibat kematian sehingga penyakit ini menjadi objek yang menarik untuk diteliti. Penelitian ini membahas mengenai faktor-faktor yang mempengaruhi timbulnya penyakit diare di Indonesia dengan menggunakan Analisis *Multivariate*. Sebelum menganalisis, dilakukan tahapan *preprocessing data*. Hasil yang diperoleh adalah pada data terdapat beberapa data yang *missing* dan mengalami *outlier*. Selain itu, dilakukan pula pengujian normalitas dan homogenitas varians kovarians yang diperoleh hasil bahwa data tidak mengikuti persebaran distribusi normal serta bersifat heterogen. Berdasarkan hal tersebut, data faktor-faktor yang mempengaruhi penyakit diare di Indonesia pada tahun 2016 diasumsikan telah berdistribusi normal dan bersifat homogen agar dapat dilakukan analisis lebih lanjut.**

***Kata Kunci*—Analisis *Multivariate*, Diare, Kesehatan, *Preprocessing Data*.**

I. PENDAHULUAN

Indonesia sebagai salah satu negara dengan jumlah penduduk terbesar di Asia, tercatat sebagai satu-satunya negara yang tidak mengalami penurunan pada indeks kesehatan pada tahun 2017. Hal tersebut dapat dibuktikan dengan adanya hasil survei yang menyatakan bahwa sebanyak 62% responden dari Indonesia merasa lebih sehat dari kondisi tiga tahun yang lalu. Angka ini merupakan tertinggi di kawasan Asia [1]. Akan tetapi dibalik pencapaian tersebut, Indonesia tetap sangat memerlukan perhatian khusus dalam penangan persebaran jenis penyakit baru maupun yang telah ada agar dapat meningkatkan derajat kesehatan masyarakatnya. Menurut Hendrick L. Blumm, derajat kesehatan dapat dipengaruhi oleh 4 faktor yaitu lingkungan, perilaku, pelayanan kesehatan dan keturunan [2]. Perilaku dapat digambarkan dengan kebiasaan sehari-hari oleh masyarakat seperti pola makan, kebersihan diri, gaya hidup dan sikap dalam menjaga kesehatan. Penduduk yang bermukim didaerah padat seringkali memiliki pola perilaku yang belum sehat. Hal ini disebabkan oleh beberapa faktor seperti lingkungan yang kurang bersih dan layak serta kepadatan penduduk.

Oleh karena terdapat banyak faktor penyebab penyakit diare, maka dilakukan penelitian kembali untuk mengetahui faktor utama yang memiliki hubungan erat dengan penyakit diare di Indonesia pada tahun 2016. Akan tetapi, sebelum melakukan analisis berdasarkan data yang telah diperoleh, terlebih dahulu *preprocessing* data. Menurut penelitian sebelumnya yang menyatakan bahwa *preprocessing* data sangat diperlukan untuk meningkatkan kualitas dari data dengan cara menghilangkan atau menghapus data yang tidak diperlukan dari data yang ada [3]. *Preprocessing* data dapat dilakukan dengan tiga tahap pembersihan data yaitu pembersihan *incomplete data (missing value)* serta *noisy and consistent data (outlier)* [4]. Untuk menentukan metode yang sesuai dalam mengatasi pembersihan data tersebut harus memperhatikan tipe dan distribusi dari data yang ada karena pada tahapan *preprocessing* data sangat menentukan hasil akhir analisis nantinya [5].

Berdasarkan penelitian-penelitian sebelumnya yang telah disebutkan, sebelum melakukan analisis terlebih dahulu dilakukan *preprocessing* terdahap data faktor-faktor penyebab penyakit diare. Tujuan dari proses *preprocessing* data ini adalah agar data yang telah diperoleh telah tersusun rapi tanpa adanya data yang hilang maupun adanya data yang berbeda dengan lainnya sehingga dapat dianalisis dan didapatkan kesimpulan yang akurat sesuai dengan kondisi sebenarnya. Batasan masalah yang diterapkan dalam penelitian ini yaitu hanya sebatas pada tahap *preprocessing* dengan analisis *missing value* dan *outlier*. Manfaat yang ingin diperoleh yaitu agar pada penelitian-penelitian selanjutnya dapat mengetahui pentingnya tahapan *preprocessing* data sebagai proses awal dalam melakukan suatu analisis. Selain itu, manfaat yang dapat diperoleh bagi mahasiswa adalah sebagai pengetahuan tambahan tentang perkembangan kesehatan di Indonesia.

II. TINJAUAN PUSTAKA

A. *Missing Value*

Missing Value adalah informasi yang tidak tersedia untuk sebuah objek (kasus). *Missing value* terjadi karena informasi tentang sesuatu objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak tersedia. *Missing value* pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit, misalkan hanya 1 % dari seluruh data. Dalam kasus analisis multivariat, apabila terdapat jumlah *missing value* yang besar maka variabel tersebut dapat dihilangkan dari penelitian. Terdapat dua bentuk missing value random, yaitu *Missing Completely at Random* (MCAR) dan *Missing at Random* (MAR). MCAR merupakan data hilang yang didistribusikan secara acak di semua pengamatan. Bentuk ini dibagi menjadi dua bagian yaitu satu kumpulan yang berisi nilai yang hilang dan lainnya berisi nilai yang tidak hilang. Apabila data berbentuk MCAR, maka peneliti dapat memilih untuk menggunakan metode *Pair-wise* atau *List wise* untuk mengatasi data yang hilang. Sedangkan MAR merupakan nilai hilang yang tidak didistribusikan secara acak di seluruh pengamatan, namun hanya didistribusikan pada satu atau lebih sub sampel. Bentuk MAR lebih umum dari sebelumnya.

Metode yang digunakan untuk menangani *missing data* antara lain sebagai berikut.

1. Mengabaikan dan membuang *missing data*, contoh metode yang sering digunakan pada kategori ini adalah metode *Listwise delection* dan *Pairwaise delection*
2. Estimasi parameter, contohnya algoritma *Expectation-Maximization (EM Algorithm)* yang digunakan untuk mengestimasi parameter dari *missing data*
3. Imputasi, yaitu proses pengisian atau penggantian nilai-nilai yang hilang *(mising value)* pada sekumpulan data *(dataset)* dengan nilai-nilai yang mungkin berdasarkan informasi yang didapatkan pada *dataset* tersebut [6].

B. *Outlier*

Hawkins mendefinisikan outlier adalah data yang secara signifikan berbeda dari data lainnya yang ada. Sebuah *outlier* merupakan pengamatan yang menyimpang jauh dari pengamatan lain untuk membangkitkan kecurigaan bahwa pengamatan itu dihasilkan oleh mekanisme yang berbeda. Dixon dan Wainer mendefinisaikan outlier sebagai nilai yang "meragukan di mata peneliti" dan pencemar. Barnet dan Lewis menunjukkan bahwa observasi terpencil, atau *outlier*, adalah salah satu yang tampaknya menyimpang nyata dari anggota lain dari sampel, sama halnya Johnson juga mendefinisikan outlier sebagai pengamatan dalam kumpulan data yang tampaknya tidak konsisten dengan sisa data set [6]. Metode deteksi *outlier* dibagi atas *Univariate Detection* dan *Multivariate Detection.*

1. *Univariate detection*

Deteksi *outlier* secara *univariate* dilakukan dengan membandingkan dengan nilai Z.

2. *Multivariate detection*

Deteksi normal *multivariate* dilakukan dengan menggunakan nilai jarak mahalanobis D^2 setiap data ke-i terhadap pusat data tersebut.

C. *Uji Normalitas*

Uji normalitas digunakan untuk mengukur apakah data yang ada telah mengukuti persebaran distribusi normal atau tidak. Terdapat dua pendekatan yang dilakukan dalam pengujian normalitas ini, yaitu dengan pendekatan *univariate* dan *multivariate*

1. Secara *univariate*

Hipotesis yang digunakan dalam pengujian normal *univariate* sebagai berikut.

Hipotesis :

H_0 : Data mengikuti sebaran distribusi normal *univariate*

H_1 : Data tidak mengikuti sebaran distribusi normal *univariate*

Statistik uji:

$$r_q = \frac{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}} \frac{\left(q_j - \bar{q}\right)}{\sqrt{\sum_{j=1}^{n}\left(q_j - \bar{q}\right)^2}} \quad (1)$$

Daerah penolakan:

Tolak H_0 apabila nilai r_q kurang dari nilai r_{tabel}

2. Secara *multivariate*

Untuk mengetahui sebaran distribusi normal suatu data secara *multivariate* dapat dilakukan dengan du acara yaitu dengan menggunakan *Q-Q Plot* yang kemudian dilajutkan dengan pengujian signifikansi koefisian korelasi serta menggunakan proporsi data dengan *mahalanobis*.

Hipotesis yang digunakan dalam pengujian sebaran distribusi normal secara *multivariate* adalah sebagai berikut.

Hipotesis :

H_0 : Data berdistribusi *multivariate normal*

H_1 : Data tidak berdistribusi *multivariate normal*

Statistik uji dari pengujian signifikansi koefisien korelasi adalah seperti di bawah ini.

Statistik uji:

$$r_q = \frac{\sum_{j=1}^{n}\left(qc - \overline{qc}\right)}{\sqrt{\sum_{j=1}^{n}\left(qc - \overline{qc}\right)^2}} \frac{\left(d_j^2 - \overline{d_j^2}\right)}{\sqrt{\sum_{j=1}^{n}\left(d_j^2 - \overline{d_j^2}\right)^2}} \quad (2)$$

Daerah penolakan :

r_0 < *critical points* maka tolak H_0

$r_0 \geq$ *critical points* maka gagal tolak H_0

Sedangkan statistik uji untuk pengujian normalitas secara *multivariate* menggunakan proporsi data adalah berikut ini.

Statistik uji:

$$d_j^2 = \left(x_j - \bar{x}\right)' S^{-1} \left(x_j - \bar{x}\right) \quad (3)$$

Data dapat dikatakan gagal tolak H_0 atau sebaran berdistribusi normal apabila nilai $d_j^2 \leq \chi^2_{(p;0.5)}$ dengan proporsi sekitar 50%.

D. *Uji Homogenitas*

Uji ini digunakan untuk memperlihatkan bahwa dua data atau lebih kelompok data sampel berasal dari populasi yang memiliki variansi yang sama.

Hipotesis :

$H_0 : \sum_1 = \sum_2 = ... = \sum_g \approx \sum$

H_1 : minimal ada satu kelompok yang berbeda, $\sum_i \neq \sum_j ; i, j, ..., g$

Statistik uji *Box's M* :

$$M = \left[\sum_l (n_l - 1)\right] \ln \left|S_{pooled}\right| - \sum_l (n_l - 1) \ln |S_l| \quad (4)$$

$$C = (1 - \mu) M$$

dimana :

g : banyak kelompok

$\left|S_{pooled}\right|$: matrik kovarian gabungan dalam kelompok

S_l : matrik kovarian kelompok ke-l

Daerah penolakan :

Tolak H_0 jika $C \geq \chi^2_{p(p+1)(g-1)/2}(\alpha)$, artinya matrik varian kovarian antar kelompok tidak homogen.

F. Diare

Penyakit diare merupakan penyakit endemis di Indonesia dan juga merupakan penyakit potensial kejadian Luar Biasa (KLB) yang sering disertai dengan kematian. Pada tahun 2016 terjadi 3 kali KLB diare yang tersebar di 3 provinsi, 3 kabupaten, dengan jumlah penderita 198 orang dan kematian 6 orang (CFR 3,04%). Pada umumnya, diare adalah gejala umum dari infeksi yang disebebkan oleh bakteri, virus dan protozoa. Siklus hidup ini berasal dari kotoran manusia/ hewan yang kemudian mengontaminasi lingkungan dan melakukan kontak dengan manusia [7].

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder mengenai beberapa faktor yang berpengaruh terhadap kasus diare di tempat umum yang terjadi pada 34 provinsi di Indonesia. Data yang digunakan berasal dari Kementerian Kesehatan Republik Indonesia pada tahun 2016.

B. Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini terdiri atas sembilan variabel independen dengan satu variabel dependen. Adapun variabel yang digunakan dalam penelitian ini adalah sebagai berikut.

Tabel 1. Variabel Penelitian

Variabel	Keterangan
Y	Kasus Diare di Tempat Umum
X_1	Jumlah Penduduk
X_2	Jumlah Penduduk Miskin
X_3	Jumlah Puskesmas
X_4	Jumlah Rumah Sakit
X_5	Sumber Daya Manusia Kesehatan
X_6	Sarana Air Minum yang Diawasi
X_7	TTU sesuai Standar
X_8	TPM sesuai Standar
X_9	Rumah Sakit dengan Pengolahan Limbah

C. Metode Analisis Data

Langkah-langkah yang digunakan dalam melakukan analisis antara lain sebagai berikut.

1. Identifikasi dan perumusan masalah
2. Mendeskripsikan data
3. Melakukan deteksi *missing data*
4. Imputasi *missing value* dengan metode *Listwise*
5. Mengidentifikasi dan mengatasi adanya *outlier* baik secara *univariate* maupun multivariat
6. Pengujian normal *univariate* dan *multivariate*
7. Pengujian homogenitas

IV. ANALISIS DAN PEMBAHASAN

A. Deteksi Missing Value dan *Cara Mengatasi*

Sebelum melakukan analisis pada suatu data, diperlukan adanya *preprocessing data*. Tujuannya agar data dapat di-*cleaning* sehingga dapat menghasilkan analisis yang sesuai. Salah satu tahapan dari *preprocessing data* adalah mengidentifikasi dan mengatasi *missing value*. Berikut ini diberikan *output* untuk mendeteksi adanya *missing value* pada setiap variabel.

Tabel 2. *Univariate Missing Value*

Provinsi	*Missing*	
	Count	*Percent (%)*
X_1	0	0
X_2	0	0
X_3	0	0
X_4	0	0
X_5	0	0
X_6	1	2,9
X_7	0	0
X_8	0	0
X_9	2	5,9
Y	0	0

Tabel 2 menunjukkan bahwa terdapat *missing value* pada variabel sarana air minum yang diawasi sebanyak satu data serta variabel rumah sakit dengan pengolahan limbah sebanyak dua data. *Missing value* pada kedua variabel tersebut hanya berkisar 2,9% dan 5,9% dari data yang ada sehingga tidak perlu dihilangkan. Maka, dibutuhkan nilai pengganti untuk mengisi kekosongan data tersebut. Untuk menentukan metode yang tepat dalam mengatasi *missing value*, terlebih dahulu dilakukan pengecekan terhadap keacakan data yang hilang.

Menggunakan pengujian *Chi-Square* dengan tingkat signifikansi sebesar 5%, diperoleh *p-value* senilai 0,95. Nilai *p-value* tersebut lebih tinggi daripada tingkat signifikansi yang digunakan (5%). Sehingga dapat disimpulkan gagal tolak H_0 atau *missing value* pada data tidak bersifat random (MCAR).

Tabel 3. *Listwise Means*

X_6	X_9
159,6774	14,5484

Pada data *missing value* MCAR, dapat digunakan metode *listwise* untuk melakukan imputasi nilai yang hilang yaitu dengan menggantikannya dengan angka pada Tabel 3 sesuai dengan masing-masing variabel.

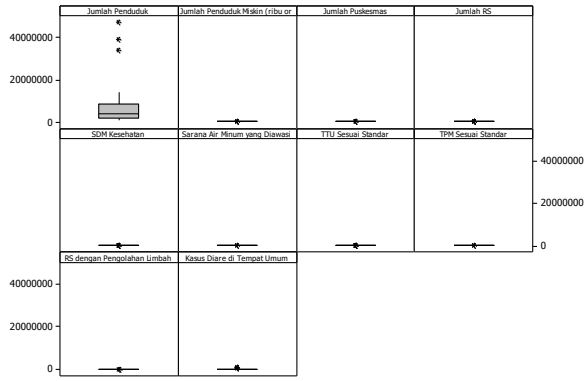
B. Deteksi Outlier dan Cara Mengatasi

Deteksi *outlier* dilakukan untuk mengetahui adanya data yang berbeda dengan data lainnya. Deteksi ini

dilakukan baik secara *univariate* maupun *multivariate* sebagai berikut.

1. Mendeteksi *outlier* secara *univariate*

Langkah pertama dapat dilakukan dengan mendeteksi secara visual menggunakan *Box-plot* terhadap masing-masing variabel. Berikut merupakan hasil *Box-plot* yang diperoleh.



Gambar 1. *Box-Plot* Masing-Masing Variabel

Berdasarkan *Box-plot* di atas maka dapat diketahui bahwa terdapat *outlier* pada variabel Jumlah Penduduk yang ditandai dengan adanya tanda bintang yang jauh dari pengamatan. Tanda bintang tersebut menunjukkan bahwa data jumlah penduduk pada Provinsi Jawa Barat, Jawa Timur dan Jawa Tengah lebih besar dibandingkan dengan provinsi lainnya di Indonesia. Sedangkan untuk variabel lainnya tidak terdapat *outlier* karena tidak ada tanda bintang yang masih disekitar garis *quartile*.

Langkah selanjutnya, mendeteksi *outlier* dengan menggunakan *Standardized*, deteksi ini dilakukan dengan membandingkan nilai z sebagai normal standard pada masing-masing obsevasi dengan ambang batas tertentu. Data akan dikatakan *outlier* apabila nilai z melebihi 3.00. Pada lampiran 1 disajikan hasil *Studentized Residual* yang didapatkan dengan menggunakan bantuan *software* SPSS. Berdasarkan hasil tersebut maka data pada Provinsi Jawa Barat merupakan data *outlier* dengan nilai *Studentized Residual* melebihi 3 yaitu sebesar 3,514.

2. Mendeteksi *outlier* secara *multivariate*

Deteksi *outlier* dilakukan dengan menggunakan nilai jarak Mahalanobis D^2 pada setiap data ke-i terhadap pusat data tersebut. Taraf signifikan yang digunakan dalam pengujian *outlier* pada umumnya menggunakan 0.001. Pada lampiran 2, disajikan hasil jarak mahalanobis beserta dengan hasil *Outlier* yang didapatkan dengan menggunakan bantuan *software* SPSS. Hasil tersebut menunjukkan bahwa terdapat data *outlier* pada data Provinsi Jawa Barat dan Sumatera Barat.

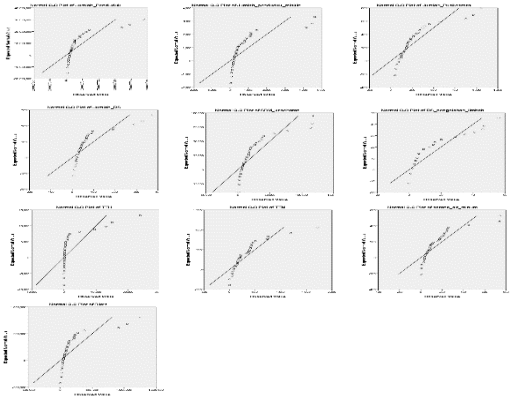
Setelah dideteksi dengan menggunakan pendekatan *outlier* secara *univariate* dan *multivariate* maka didapatkan beberapa data *outlier* sehingga perlu diatasi dengan menghilangkan data-data tersebut. Akan tetapi, data yang digunakan merupakan data seluruh provinsi di Indonesia, maka tidak mungkin untuk menghilangkan salah satu provinsi. Sehingga perlu dilakukan metode lain untuk mengatasi masalah tersebut yaitu dengan mentransformasikan data.

C. Pengujian Normalitas

Setelah *outlier* dapat diatasi maka dilakukan pengujian normal terhadap data baik secara *univariate* maupun *multivariate*.

1. Uji normalitas secara *univariate*

Identifikasi normal *univariate* dilakukan dengan menggunakan *Q-Q plot* yang merupakan analisis plot grafik probabilitas yang digunakan untuk menetapkan apakah distribusi suatu variabel tertentu sesuai dengan variabel yang telah ditetapkan. Gambar 3.2 merupakan *Q-Q plot* yang dihasilkan terhadap masing-masing variabel.



Gambar 2. *Q-Q Plot* untuk Masing-Masing Variabel Penelitian

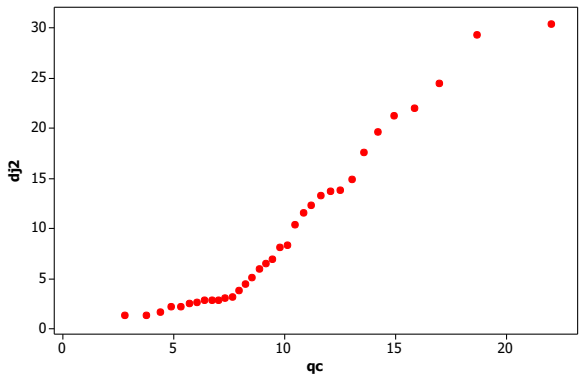
Berdasarkan gambar diatas maka secara visual dapat ditarik kesimpulan bahwa pada masing-masing variabel penelitian tidak mengikuti sebaran distribusi normal *univariate*. Hal ini terlihat dari hasil *Q-Q Plot* dimana pada setiap variabel penelitian, tanda titik-titik tidak mengikuti garis normal serta persebaran data berupa titk-titik membentuk pota seperti bentuk U.

2. Uji normalitas secara *multivariate*

Salah satu asumsi yang perlu dipenuhi oleh sebuah data adalah data tersebut harus berdistribusi normal. Untuk melihat apakah data telah berdistribusi normal atau belum, dilakukan pengujian secara *multivariate*. Pengujian normalitas secara *multivariate* dapat dibagi menjadi dua pengujian yaitu dengan menggunakan uji signifikansi koefisien korelasi dan proporsi dengan *Mahalanobis*.

a) Pengujian signifikansi koefisien korelasi

Pengujian ini diawali dengan membuat *plot chi-square* yang digunakan untuk mengetahui secara visual pola data faktor-faktor yang mempengaruhi penyakit diare di Indonesia pada tahun 2016 apakah telah memenuhi asumsi distribusi normal maupun tidak. Berikut merupakan *output* yang dihasilkan.



Gambar 3. *Plot Chi-square*

Gambar 3 menunjukkan secara visual bahwa pola data yang tidak lurus serta membentuk pola tertentu, sehingga dapat dikatakan bahwa data tidak berdistibusi normal dengan menggunakan *plot chi-square*.

Kemudian dilanjutkan dengan pengujian signifikansi koefisien korelasi yang digunakan untuk melihat apakah data memenuhi asumsi distribusi normal *multivariate* atau tidak. Berdasarkan hasil pengolahan maka didapatkan koefisien korelasi sebesar 0,975 dimana nilai ini lebih kecil daripada Tabel R yang memiliki nilai 0,9957. Apabila nilai koefisien korelasi lebih kecil daripada nilai Tabel R maka dapat dikatakan bahwa pengujian mengalami tolak H_0 yang artinya data yang digunakan tidak mengikuti persebaran distribusi normal secara *multivariate*.

b) Proporsi

Data dapat dikatakan berdistribusi normal secara *multivariate* jika nilai $d_j^2 < \chi^2_{(p;0,5)}$ memiliki proporsi sekitar 50% ± 4%. Dari perhitungan yang dilakukan, didapati proporsi adalah sebanyak 58,823%. Sesuai dengan syarat yang telah diberikan, maka dapat disimpulkan bahwa data tidak mengikuti sebaran distribusi normal.

Berdasarkan kedua pendekatan baik secara *univariate* maupun *multivariate*, menunjukkan bahwa data faktor-faktor yang mempengaruhi penyakit diare di Indonesia pada tahun 2016 tidak berdistribusi normal. Agar dapat dilakukan analisis lebih lanjut, maka data yang tersebut diasumsikan telah mengikuti sebaran distribusi normal.

D. Uji Homogenitas Varians Kovarians

Untuk mengetahui apakan data yang digunakan memiliki matriks kovarian yang sama, maka perlu dilakukan pengujian homogenitas varians kovarians dengan *Box's Test*.

Tabel 4. *Box's Test*

Box's M	
p-value	0,000

Hasil yang diperoleh dari *Box's Test* memberikan hasil *p-value* sebesar 0,000. Pada pengujian, tingkat signifikansi yang telah ditentukan adalah sebesar 0,05. Hal ini berarti bahwa pengujian mengalami tolak H_0 atau terdapat salah satu variabel memiliki persebaran data yang tidak homogen. Persebaran data tidak homogen artinya *range* data memiliki nilai cukup tinggi hingga menyebabkan antara data satu dengan data lainnya memiliki jarak yang cukup jauh sehingga menyebabkan data tersebut heterogen. Oleh karena itu, data perlu untuk diasumsikan bersifat homogen agar dapat dilakukan analisis lebih lanjut.

V. KESIMPULAN

Sebelum melakukan analisis lebih lanjut, dilakukan terlebih dahulu tahapan *preprocessing data* seperti mendeteksi *missing value* dan mendeteksi *outlier*. Setelah dilakukan *preprocessing* data, dilanjutkan dengan uji asumsi menggunakan pengujian normalitas maupun pengujian homogenitas varians kovarians. Hasil yang diperoleh adalah terdapat *missing value* pada faktor sarana air minum yang diawasi dan rumah sakit dengan pengelolaan limbah. Terdapat pula *outlier* pada Provinsi Jawa Barat, Banten, Sumatera Barat, Jawa Timur, dan Jawa Tengah akan tetapi data *outlier* tersebut tidak dapat dihilangkan karena data provinsi merupakan satu kesatuan dari wilayah Indonesia. Kemudian dilakukan pengujian normalitas dan homogenitas varians kovarians dengan menggunakan pendekatan *univariate* maupun *multivariate*. Hasilnya data tidak mengikuti persebaran distribusi normal dan data tidak memiliki variansi yang sama (heterogen) sehingga diasumsikan data telah berdistribusi normal dan bersifat homogen agar dapat dilakukan analisis dengan metode lainnya.

Saran bagi peneliti yaitu agar mengecek terlebih dahulu data yang akan digunakan sehingga hasil analisis nantinya akan menghasilkan *output* yang bermanfaat bagi orang lain. Untuk masyarakat, agar selalu menjaga kebersian diri maupun lingkungan tempat tinggal.

DAFTAR PUSTAKA

[1] Varabi, Hatim. 2017. *Tingkat Kesehatan Masyarakat Indonesia Meningkat*. Jakarta: Koran Sindo.

[2] Dinas Kesehatan Provinsi DKI Jakarta. 2016. *Profil Kesehatan Provinsi DKI Jakarta Tahun 2016*. Jakarta: Dinkes Provinsi DKI Jakarta.

[3] Rani, Sanggeta dan Sonika. 2014. *Effectiveness of Data Preprocessing for Data Mining*. Bahadurgarh: *BLS Institute of Technology Management*.

[4] Meilina, Popy. 2014. *Penerapan Data Mining dengan Metode Kalsifikasi menggunakan Decision Tree dan Regresi*. Jakarta: Universitas Muhammadiyah Jakarta.

[5] Patel, Kinnari dkk. 2015. *Incremental Missing Value Replacement Techniques for Stream Data*. India: *University Jaipur*.

[6] Iran, B.-G. 2005. *Outlier Detection*. Retrieved from Outlier Detection: http://www.eng.tau.ac.il/~bengal/outlier.pdf

[7] Kementerian Kesehatan Republik Indonesia. 2016. *Profil Kesehatan Indonesia Tahun 2016*. Jakarta: Kemenkes RI.

[8] Kusumawati, A, dkk. Academia. Retrieved from Pengujian Normal Multivariate: https://www.academia.edu/11712313/Pengujian_Normal_Multivariat

Lampira-lampiran

Lampiran 1. Hasil *Output Standardized*

Provinsi	SRE
Aceh	-0.1731
Sumatera Utara	-0.0972
Sumatera Barat	-0.4538
Riau	-0.0642
Jambi	-0.5014
Sumatera Selatan	-0.1894
Bengkulu	0.04025
Lampung	0.10826
Kepulauan Bangka B	0.8993
Kepulauan Riau	0.34137
DKI Jakarta	1.19022
Jawa Barat	3.51409
Jawa Tengah	-0.6517
DI Yogyakarta	1.00427
Jawa Timur	-1.1994
Banten	-4.6338
Bali	0.69167
Nusa Tenggara Bara	1.44422
Nusa Tenggara Timu	-0.5236
Kalimantan Barat	-0.3773
Kalimantan Tengah	-0.1805
Kalimantan Selatan	-2.121
Kalimantan Timur	0.09396
Kalimantan Utara	-0.4072
Sulawesi Utara	0.02908
Sulawesi Tengah	0.24022
Sulawesi Selatan	1.06195
Sulawesi Tenggara	-0.3731
Gorontalo	0.31047
Sulawesi Barat	0.1714
Maluku	1.92526
Maluku Utara	-0.0381
Papua Barat	0.14166
Papua	-0.3963

Lampiran 2. Hasil *Output* Jarak Mahalanobis beserta dengan hasil *outlier*

Provinsi	Mahalanobis	*Outlier*
Aceh	6.3858	0.7
Sumatera Utara	11.4852	0.24
Sumatera Barat	24.31283	0
Riau	2.79951	0.97
Jambi	2.09642	0.99
Sumatera Selatan	6.84629	0.65
Bengkulu	2.54603	0.98
Lampung	17.57591	0.04
Kepulauan Bangka B	4.12753	0.9
Kepulauan Riau	2.83353	0.97
DKI Jakarta	21.2668	0.01
Jawa Barat	26.38485	0
Jawa Tengah	19.33722	0.02
DI Yogyakarta	12.4215	0.19
Jawa Timur	20.5281	0.01
Banten	15.9833	0.07
Bali	7.78676	0.56
Nusa Tenggara Barat	10.34333	0.32
Nusa Tenggara Timur	10.06243	0.35
Kalimantan Barat	2.58367	0.98
Kalimantan Tengah	4.30708	0.89
Kalimantan Selatan	10.85842	0.29
Kalimantan Timur	2.70519	0.97
Kalimantan Utara	5.71308	0.77
Sulawesi Utara	1.30114	1
Sulawesi Tengah	3.61909	0.93
Sulawesi Selatan	12.69441	0.18
Sulawesi Tenggara	2.87934	0.97
Gorontalo	1.94874	0.99
Sulawesi Barat	2.05883	0.99
Maluku	10.45808	0.31
Maluku Utara	1.24576	1
Papua Barat	1.57504	1
Papua	7.92879	0.54

Lampiran 3. *Syntax R* dan *Output* untuk Menentukan Proporsi dari Pengujian Normalitas secara *Multivariate*

```
multinorm.test<-function(x)
{
  x<-as.data.frame(x)
  mu<-colMeans(x)
  S<-cov(x)
  invS<-solve(S)
  d<-matrix(rep(0,nrow(x)),nrow(x),1)
```

```
  eval<-matrix(rep(0,nrow(x)),nrow(x),1)
  q<-qchisq(0.5,ncol(x))
  for (i in 1:nrow(x))
  {
    d[i]<-as.numeric(x[i,]-mu) %*% (invS) %*% as.numeric(t(x[i,]-mu))
  ifelse (d[i]<=q, eval[i]<-1, eval[i]<-0)
  }
prop <- sum(eval)/nrow(x)
result<- list(distance=d, chisquared=q, proportion=prop)
return (result)
}
library(readxl)
data_diare <- read_excel("E:/MATERI/SEMESTER 6/Multivariate/data_diare.xlsx",
+     col_types = c("blank", "blank", "numeric",
+         "numeric", "numeric", "numeric",
+         "numeric", "numeric", "numeric",
+         "numeric", "numeric", "numeric"))
`col_type = "blank"` deprecated. Use "skip" instead.
View(data_diare)
multinorm.test(data_diare)
$distance
          [,1]
 [1,]  6.417818
 [2,] 11.493290
 [3,] 24.379034
 [4,]  2.804533
 [5,]  2.409916
 [6,]  6.883940
 [7,]  2.548022
 [8,] 17.582972
 [9,]  5.067740
[10,]  2.975288
[11,] 21.902072
[12,] 29.289167
[13,] 19.561793
[14,] 13.245486
[15,] 21.217496
[16,] 30.339093
[17,]  8.270015
[18,] 12.227996
[19,] 10.313397
[20,]  2.758335
[21,]  4.344728
[22,] 14.826704
[23,]  2.715975
[24,]  5.894875
[25,]  1.302221
[26,]  3.687397
[27,] 13.602939
[28,]  3.048441
[29,]  2.069560
[30,]  2.095519
[31,] 13.789608
[32,]  1.247623
[33,]  1.600505
[34,]  8.086499

$chisquared
[1] 9.341818

$proportion
[1] 0.5882353
```

Lampiran Data

No	Provinsi	Jumlah Penduduk	Jumlah Penduduk Miskin (ribu orang)	Jumlah Puskesmas	Jumlah RS	SDM Kesehatan	Sarana Air Minum yang Diawasi	TTU Sesuai Standar	TPM Sesuai Standar	RS dengan Pengolahan Limbah	Kasus Diare di Tempat Umum
1	Aceh	5096248	841.31	340	68	40702	83	364	171	6	135054
2	Sumatera Utara	14102911	1452.55	571	195	56329	71	354	85	9	376321
3	Sumatera Barat	5259528	376.51	264	67	26752	740	13088	1201	35	1403
4	Riau	6500971	501.59	213	72	28515	213	115	425	20	171299
5	Jambi	3458926	290.81	183	34	19388	169	107	432	9	91857
6	Sumatera Selatan	8160901	1096.5	322	65	38256	27	212	69	2	217412
7	Bengkulu	1904793	325.6	180	21	12288	222	1842	507	-	50622
8	Lampung	8205141	1139.78	292	64	17053	61	116	147	56	219167
9	Kepulauan Bangka B	1401827	71.07	62	17	9101	246	1181	399	2	37066
10	Kepulauan Riau	2028169	119.14	73	28	8645	85	127	208	3	53271
11	DKI Jakarta	10277628	385.84	340	190	76905	31	523	54	41	274803
12	Jawa Barat	47379389	4168.11	1050	328	117674	738	23724	1719	56	1261159
13	Jawa Tengah	34019095	4493.75	875	290	113872	450	919	754	20	911901
14	DI Yogyakarta	3720912	488.83	121	74	19863	41	519	162	47	99338
15	Jawa Timur	39075152	4638.53	960	377	116303	185	527	407	5	1048885
16	Banten	12203148	657.74	233	95	33666	257	14981	371	46	32279
17	Bali	4200069	174.94	120	57	2405	8	1152	90	29	112126
18	Nusa Tenggara Bara	4896162	786.58	158	28	17632	73	4847	213	6	130561
19	Nusa Tenggara Timu	5203514	1150.08	371	45	2213	46	181	130	-	138243
20	Kalimantan Barat	4861738	390.32	238	45	19051	89	180	225	1	129319
21	Kalimantan Tengah	2550192	137.46	195	21	12363	176	200	323	5	67365
22	Kalimantan Selatan	4055479	184.16	230	39	17007	216	121	735	6	107725
23	Kalimantan Timur	3501232	211.24	175	48	22052	244	290	454	13	92518
24	Kalimantan Utara	666333	47.03	49	7	3148	98	102	391	4	17331
25	Sulawesi Utara	2436921	200.35	188	43	13447	90	468	221	1	65127
26	Sulawesi Tengah	2921715	413.15	189	33	1515	111	137	180	0	77671

27	Sulawesi Selatan	8606375	796.81	448	90	37725	117	9561	415	16	230048
28	Sulawesi Tenggara	2551008	327.29	269	31	14236	38	181	112	4	67487
29	Gorontalo	1150765	203.69	93	13	6193	192	62	359	6	30596
30	Sulawesi Barat	1306478	146.9	94	11	5202	59	56	158	0	34619
31	Maluku	1715548	331.79	199	28	7806	-	1425	3	2	45536
32	Maluku Utara	1185912	76.4	128	20	7038	36	177	155	3	31382
33	Papua Barat	893362	223.6	151	16	4693	6	84	47	0	23531
34	Papua	3207444	914.87	393	41	16545	0	302	2	0	85034