Feature Selection pada Dataset Faktor Kesiapan Bencana pada Provinsi di Indonesia Menggunakan Metode PCA (*Principal Component Analysis*)

Septa Firmansyah Putra, Renny Pradina, dan Irmasari Hafidz
Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: renny.pradina@gmail.com, ir.hafidz@gmail.com, septa.poetra@gmail.com

***Abstrak*. Penelitian ini bertujuan untuk mengetahui atribut-atribut apa yang akan digunakan untuk klasterisasi provinsi di Indonesia berdasarkan faktor kesiapan dalam menghadapi bencana. Data yang digunakan terdiri dari tiga kelompok data yaitu data jumlah kejadian bencana yang terdiri dari 19 sub-atribut, data jumlah fasilitas kesehatan yang terdiri dari 14 sub-atribut dan data jumlah tenaga kesehatan yang terdiri dari 11 sub atribut. Penelitian ini dapat menjadi gambaran tentang bagaimana melakukan pembersihan dan pemilihan data sebelum digunakan dalam proses klasterisasi. Data-data ini akan dibersihkan dan dipilih sebelum nantinya digunakan pada proses klasterisasi. Proses pembersihan dan pemilihan data dilakukan dengan bantuan PCA (*Principal Component Analysis*) namun sebelumnya dibersihkan telebih dahulu dengan cara manual. Penelitian dibagi menjadi 3 percobaan. Pada percobaan pertama didapatkan 31 sub-atribut yang siap digunakan, percobaan kedua didapatkan 29 sub-atribut yang siap digunakan dan pada percobaan ketiga didapatkan 24 sub-atribut yang siap digunakan.**

***Kata Kunci: Data mining*, Klasterisasi, K-*Means*, PCA, R-Studio.**

I. PENDAHULUAN

HAMPIR setiap saat Indonesia mengalami bencana alam. Hal ini bisa terlihat dari seringnya terjadi bencana alam secara beruntun seperti gempa bumi, tanah longsor, banjir, letusan gunung berapi dan juga tsunami. Menurut *United Nations International Stategy for Disaster Reduction* (UNISDR; Badan PBB untuk strategi internasional Pengurangan Risiko Bencana) Indonesia merupakan salah satu negara yang paling rawan bencana alam [1].

Menurut data dari BNPB terdapat 2105 bencana yang terjadi di Indonesia pada tahun 2012. Hal ini berarti setiap hari terjadi 6 bencana di wilayah Indonesia [2]. Dengan adanya hal-hal semacam ini, pernahkah terpikirkan bagaimana caranya untuk melakukan pencegahan sejak awal agar dampak kerugian material dan korban yang ditimbulkan dapat lebih diminimalisir.

Badan Nasional Penanggulangan Bencana melakukan pendataan bencana setiap harinya. Data tersebut berisikan jenis bencana, tempat kejadian bencana (kota kejadian) beserta dampak dari bencana baik berupa korban jiwa maupun harta benda. Data-data yang ada belum dilakukan pengolahan secara baik menjadi sebuah informasi yang berharga dan mudah dipahami. Selain melakukan pendataan terhadap kejadian bencana, BNPB juga menentukan tingkat kerawanan bencana suatu provinsi namun dalam aplikasinya pembangunan infrastruktur di Indonesia tidak terlalu memperdulikan faktor bencana sehingga pembangunan infrastruktur bersifat terpusat dan tidak tersebar ke seluruh provinsi.

Jumlah penduduk Indonesia pada tahun 2004 mencapai 220 juta jiwa yang terdiri dari beragam etnis, kelompok, agama dan adat-istiadat. Keragaman tersebut merupakan kekayaan bangsa Indonesia yang tidak dimiliki bangsa lain. Namun karena pertumbuhan penduduk yang tinggi tidak diimbangi dengan kebijakan dan pembangunan ekonomi, sosial dan infrastruktur yang merata dan memadai [3] sehingga, terjadi kesenjangan pada beberapa aspek dan terkadang muncul kecemburuan sosial. Kondisi ini potensial menyebabkan terjadinya konflik dalam masyarakat yang dapat berkembang menjadi bencana nasional.

Algoritma clustering dapat diaplikasikan terhadap data bencana di Indonesia dan menambahkan beberapa faktor lain yang berkaitan dengan bencana seperti jumlah fasilitas kesehatan, jumlah penduduk, jumlah tenaga medis dan luas wilayah untuk melihat tingkat kesiapan provinsi dalam menghadapi bencana. Faktor-faktor tersebut tidak bisa langsung diaplikasikan ke dalam algoritma k-Means, perlu dilakukan sebuah proses pengolahan data sehingga faktor-faktor tersebut siap untuk diolah.

Salah satu metode yang dapat digunakan adalah PCA (*Principal Component Analysis*). Metode ini digunakan untuk mereduksi dimensi dari data atau dengan kata lain dapat digunakan untuk memilih atribut atau faktor-faktor yang memiliki hubungan yang kuat

II. TINJAUAN PUSTAKA

A. Data mining

Data mining atau dalam bahasa Indonesia dsebut penggalian data merupakan suatu proses pencarian korelasi, pola dan tren baru yang berguna dalam media penyimpanan data berukuran besar menggunakan teknologi pengenalan pola seperti teknik-teknik statistik dan matematis [4] .Penggalian data erat kaitannya dengan pencarian pola. Urutan dari pencarian pola dalam proses *data mining* adalah:

1. Pembersihan Data (proses penghapusan data yang dianggap menganggu penelitian (*noise*)).

2. Integrasi Data (proses untuk menyatukan berbagai sumber data menjadi sebuah data gabungan)
3. Pemilihan Data (memperoleh data yang relevan dengan cara memilih data yang dianggap relevan dan menghapus data yang dianggap tidak relevan)
4. Transformasi Data (proses untuk mengubah data ke dalam format untuk diproses dalam penggalian data)
5. Penggalian Data (menerapkan metode cerdas untuk ekstraksi pola)
6. Evaluasi pola (mengenali pola-pola yang menarik saja yang sebelumnya telah diidentifikasi pada proses penggalian data)
7. Penyajian pola (melakukan visualisasi pola kepada masyarakat umum)

Proses *data mining* sendiri bisa dibedakan menjadi dua tujuan utama [5] yaitu:

1. *Descriptive Data mining*
   Sebuah proses *data mining* yang bertujuan untuk menampilkan data dalam bentuk yang ringkas, informatif dan diskriminatif.
2. *Predictive Data mining*
   Sebuah proses *data mining* yang bertujuan untuk merubah analisis data menjadi sebuah model yang nantinya akan digunakan sebagai alat untuk memprediksi trend dari data yang tidak diketahui nilainya.

B. *Aplikasi R*

R-Studio merupakan aplikasi *open source* yang digunakan untuk menerjemahkan bahasa R menjadi lebih *user friendly*. Aplikasi ini dibuat karena pada zaman sekarang, orang-orang sudah mulai beralih ke bahasa R sebagai sebuah bahasa untuk proses statistik [6].

R-Studio bisa berjalan hampir di seluruh sistem operasi seperti windows, Mac OS dan juga Linux. R-Studio sendiri biasa digunakan oleh para ahli statistik maupun ahli *data mining* untuk membuat sebuah aplikasi statistik maupun analisis data.

Di dalam R-Studio terdapat berbagai macam *package*. *Package* ini berguna dalam melakukan pengolahan data. *Package* ini bisa di*download* dan di *install* secara manual atau bisa diinstall langsung melalui *software* R-Studio.

C. *PCA (Principal Component Analysis)*

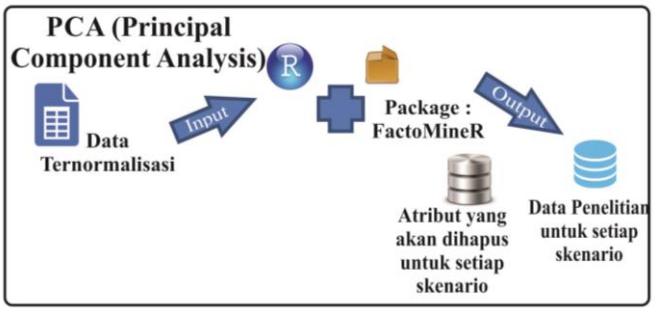
Principal Component Analysis atau analisis komponen utama (AKU) adalah teknik yang digunakan untuk menyederhanakan suatu data, dengan cara mentransformasi data secara linier sehingga terbentuk sistem koordinat baru dengan varians maksimum. Analisis komponen utama dapat digunakan untuk mereduksi dimensi suatu data tanpa mengurangi karakteristik data tersebut secara signifikan.

Tujuan utama analisis komponen utama ialah untuk mengurangi dimensi peubah-peubah yang saling berhubungan dan cukup banyak variabelnya sehingga lebih mudah untuk menginterpretasikan data-data tersebut [7]. Metode yang digunakan yaitu menentukan komponen utama dengan melakukan alih ragam *orthogonal* atau membentuk kombinasi linier Y = A' X (Sumarga, 1996). Dari sini akan dipilih beberapa komponen utama yang dapat memberikan sebagian besar keragaman total data semula.

III. METODE PENELITIAN

Metode penelitian ini terbagi menjadi dua yakni metode pemilihan atribut dengan cara manual dan metode pemilihan atribut dengan PCA. Pada proses penyeleksian atribut secara manual dilakukan dengan menghapus sub-sub atribut yang memiliki nilai 0 diseluruh kolom dan juga baris.

Metode berikutnya yakni proses pemilihan atribut dengan PCA. Proses ini bertujuan untuk melakukan penyeleksian atribut berdasarkan nilai korelasinya dengan keseluruhan atribut. Proses ini dilakukan sebelum proses validasi dan klasterisasi pada data. Seleksi atribut dengan PCA dilakukan dengan menggunakan *software* R-Studio dan menggunakan *package* **FactoMineR** [8]. Langkah-langkah melakukan pemilihan atribut dengan PCA dengan menggunakan *software* R-Studio bisa dilihat pada Gambar 1.



Gambar 1. Langkah-Langkah PCA

Proses penyeleksian atribut dilakukan dengan melihat nilai korelasi dan juga P-*Value*. Dimana saat nilai korelasi kurang dari 0.5 maka atribut tersebut akan dihapus dan saat nilai P-*Value* lebih dari 0.05 maka atribut tersebut juga akan dihapus [9].

Langkah-langkah untuk melakukan PCA dengan menggunakan *software* RStudio sebagai berikut:

1. Memilih data yang akan digunakan.
2. Mengaktifkan *package* FactoMineR

```
> library("FactoMineR", lib.loc="~/R/win-library/3.2")
```

3. Memasukkan perintah untuk melakukan PCA

```
> res.pca = PCA(Data_Bencana_2, scale.unit=TRUE, ncp=10, graph=F)
```

4. Memasukkan perintah untuk melihat dimensi (nilai korelasi) untuk setiap atribut.

```
> dimdesc(res.pca, proba = 0.5)
```

5. Sehingga akan muncul hasil dari PCA.

```
> dimdesc(res.pca, axes=c(1,2))
$Dim.1
$Dim.1$quanti
                                correlation      p.value
Jumlah.RS.Kelas.C                 0.9757201 5.222218e-22
Perawat.Bidan                     0.9630350 3.230231e-19
Posyandu                          0.9599420 1.098634e-18
Jumlah.Kamar.RS.Kelas.C           0.9554320 5.564638e-18
Jumlah.Kamar.RS.Kelas.B           0.9544246 7.813678e-18
Puskesmas                         0.9523981 1.512142e-17
Dokter.Umum                       0.9506833 2.586030e-17
Tenaga.Medis                      0.9439866 1.776086e-16
Tenaga.Gizi                       0.9409250 3.966484e-16
Jumlah.RS.Kelas.B                 0.9299991 5.098975e-15
Tenaga.Farmasi                    0.8946528 2.239980e-12
Tenaga.Sanitasi                   0.8933740 2.676049e-12
Dokter.Gigi                       0.8927819 2.903582e-12
```

Gambar 2. Hasil PCA dengan R-Studio

IV. HASIL DAN PEMBAHASAN

A. *Percobaan*

Pada penelitian ini digunakan beberapa percobaan untuk sub-atribut yang digunakan. Terdapat tiga percobaan yang digunakan yaitu percobaan 1, 2 dan 3. **Percobaan 1** berisikan data penelitian awal dengan sub-atribut yang sama tanpa diubah sedikitpun. **Percobaan 2** berisikan sub-sub atribut yang telah diproporsikan dengan Luas Wilayah. Sub-atribut yang diproporsikan adalah sub-atribut pada atribut Fasilitas Kesehatan dan Tenaga Kesehatan, kecuali untuk jumlah kamar pada Rumah Sakit. Berikut adalah rumus proporsi dari sub-atribut tersebut.

$$proporsi\ Fasilkes\ s2 = \frac{Sub\ Atribut\ FasilKes}{Luas\ Wilayah} \quad (1)$$

$$proporsi\ Tekes\ s2 = \frac{Sub\ Atribut\ Tekes}{Luas\ Wilayah} \quad (2)$$

Alasan dilakukan proporsi terhadap luas wilayah adalah untuk mengetahui jangkauan setiap fasilitas kesehatan dan tenaga kesehatan dalam sebuah wilayah, atau bisa disebutkan sebuah fasilitas kesehatan dan tenaga kesehatan dapat menjangkau wilayah seluas apa.

Percobaan 3 berisikan sub-sub atribut yang telah diproporsikan dengan Jumlah Penduduk. Sub-atribut yang diproporsikan adalah sub-atribut pada atribut Fasilitas Kesehatan dan Tenaga Kesehatan berikut adalah rumus proporsi dari sub-atribut tersebut.

$$proporsi\ Fasilkes\ s3 = \frac{Sub\ Atribut\ Fasilkes}{Jumlah\ Penduduk} \quad (3)$$

$$proporsi\ Tekes\ s3 = \frac{Sub\ Atribut\ Tekes}{Jumlah\ Penduduk} \quad (4)$$

Alasan dilakukan proporsi terhadap jumlah penduduk adalah untuk mengetahui cakupan fasilitas kesehatan dan tenaga kesehatan tehadap jumlah penduduk [10], atau bisa juga disebutkan seberapa banyak penduduk yang harus ditangani oleh fasilitas kesehatan dan tenaga kesehatan di wilayah tersebut.

B. *Hasil Seleksi Sub-Atribut dengan Proses Manual*

Berikut adalah hasil seleksi atribut dengan cara manual yakni menghapus sub-sub atribut yang memiliki nilai 0 diseluruh kolom dan juga baris. Dengan proses ini terdapat 7 sub-atribut yang dihapus dan dapat dilihat pada Tabel 1.

Tabel 1.
Hasil Pemilihan Atribut dengan Proses Manual

Atribut	Sub Atribut yang dihapus
Jumlah Kejadian Bencana Alam	• Perubahan Iklim • Tsunami
Jumlah Kejadian Bencana Non Alam	• Hama Tanaman • KLB
Jumlah Kejadian Bencanan Sosial	• Kecelakaan Industri • Aksi Teror / Sabotase
Jumlah Fasilitas Kesehatan	• Polindes

C. *Hasil Pemilihan Sub-Atribut dengan Proses PCA*

Tabel 2.
Hasil PCA Percobaan 1

Sub-Atribut	Korelasi	P-*Value*
Perawat Bidan	0.963468	1.88E-20
Jumlah RS Kelas C	0.963229	2.09E-20
Jumlah Kamar RS Kelas C	0.953713	8.67E-19
Posyandu	0.953153	1.05E-18
Puskesmas	0.951385	1.92E-18
Tenaga Medis	0.951186	2.05E-18
Tenaga Gizi	0.945226	1.31E-17
Dokter.Umum	0.943406	2.21E-17
Jumlah Kamar RS Kelas B	0.941487	3.78E-17
Tenaga Farmasi	0.925214	1.92E-15
Tenaga Keteknisan Medis	0.925161	1.94E-15
Jumlah RS Kelas B	0.922502	3.38E-15
Dokter Gigi	0.914925	1.49E-14
Putting Beliung	0.908972	4.34E-14
Jumlah RS Kelas D	0.908746	4.52E-14
Kekeringan	0.905148	8.32E-14
Tenaga Sanitasi	0.904076	9.93E-14
Jumlah Kamar RS Kelas D	0.898799	2.31E-13
Banjir	0.897223	2.94E-13
Dokter Spesialis	0.880906	2.95E-12
Jumlah Kamar RS Kelas A	0.873570	7.48E-12
Tanah Longsor	0.869456	1.23E-11
Pustu	0.865218	2.01E-11
Tenaga Kesma	0.860323	3.49E-11
Jumlah RS Kelas A	0.847171	1.39E-10
Jumlah RS Kelas Tidak Diketahui	0.828954	7.73E-10
Jumlah Kamar RS Kelas Tidak Diketahui	0.771872	5.66E-08
Kebakaran	0.755058	1.60E-07
Keterapian Fisik	0.704061	2.37E-06
Banjir dan Tanah Longsor	0.639233	3.57E-05
Gempa Bumi	0.528205	1.11E-03
Jumlah Penduduk	0.495588	2.02E-03
Gempa dan Tsunami	0.495476	2.47E-03
Gelombang Pasang	0.458337	5.62E-03
Letusan Gunung Api	0.383340	2.30E-02

Kecelakaan Transportasi	0.378669	2.49E-02
Kebakaran Hutan dan Lahan	0.351097	3.86E-02
Konflik Sosial	0.268861	1.18e-01
Luas Wilayah	0.224634	1.94e-01

Tabel 2, kolom dengan warna kuning menyatakan sub-atribut yang akan dihapus yakni sub-atribut yang memiliki nilai korelasi kurang dari 0.5 dan nilai P-*Value* lebih dari 0.05. Sub-atribut yang tidak dihapus berjumlah 31 dan sub-atribut yang dihapus berjumlah 8.

Tabel 3.
Hasil PCA Percobaan 2

Sub-Atribut	Korelasi	p.*value*
Tenaga Gizi	0.993605	6.17E-31
Jumlah RS Kelas C	0.992374	9.38E-30
Perawat Bidan	0.992361	9.62E-30
Tenaga Farmasi	0.990735	1.89E-28
Sanitarian	0.987656	1.58E-26
Puskesmas	0.986997	3.53E-26
Tenaga Keteknisan Medis	0.986168	9.15E-26
Tenaga Kesma	0.986084	1.00E-25
Jumlah Kamar RS Kelas C	0.982069	4.97E-24
Tenaga Medis	0.968961	2.24E-20
Dokter Gigi	0.962312	4.34E-19
Jumlah Kamar RS Kelas B	0.929256	5.97E-15
Jumlah RS Kelas B	0.923335	1.99E-14
Jumlah RS Kelas Tidak Diketahui	0.917686	5.76E-14
Dokter Umum	0.915053	9.21E-14
Dokter Spesialis	0.886986	6.30E-12
Keterapian Fisik	0.885614	7.52E-12
Jumlah Kamar RS Kelas A	0.88239	7.10E-11
Jumlah Kamar RS Kelas Tidak Diketahui	0.866615	8.63E-11
Jumlah RS Kelas D	0.864813	1.95E-10
Jumlah Kamar RS Kelas D	0.856998	2.46E-10
Posyandu	0.854653	4.82E-10
Puting Beliung	0.850187	2.59E-08
Jumlah RS Kelas A	0.847718	5.28E-04
Banjir	0.838707	4.36E-03
Tanah Longsor	0.822051	4.14E-02
Tenaga Sanitasi	0.798389	3.81E-10
Kekeringan	0.796699	1.10E-09
Pustu	0.570446	4.47E-09
Kebakaran	0.542308	2.92E-08
Gelombang Pasang	0.466198	1.11E-03
Jumlah Penduduk	0.357074	6.25E-03
Banjir dan Tanah Longsor	-0.028477	8.75E-01
Kecelakaan Transportasi	-0.029335	8.71E-01
Kebakaran Hutan dan Lahan	-0.036212	8.41E-01
Gempa dan Tsunami	-0.047338	7.93E-01
Letusan Gunung Api	-0.058475	7.46E-01
Gempa Bumi	-0.075433	6.77E-01
Konflik Sosial	-0.158769	3.77E-01

Tabel 3, pada kolom dengan warna kuning menyatakan sub-atribut yang akan dihapus yakni sub-atribut yang memiliki nilai korelasi kurang dari 0.5 dan nilai P-*Value* lebih dari 0.05. Sub-atribut yang tidak dihapus berjumlah 28 dan sub-atribut yang dihapus berjumlah 9.

Tabel 4.
Hasil PCA Percobaan 3

Sub-Atribut	Korelasi	P-*value*
Tenaga Medis	0.897952	1.40E-12
Tenaga Farmasi	0.831849	1.57E-08
Dokter Umum	0.827137	3.49E-08
Posyandu	0.805439	9.35E-08
Jumlah Kamar RS Kelas C	0.794063	1.23E-07
Dokter Gigi	0.778921	1.57E-07
Jumlah RS Kelas C	0.774466	2.79E-07
Tenaga Keteknisan Medis	0.770477	2.54E-06
Perawat Bidan	0.760659	1.28E-05
Jumlah RS Kelas Tidak Diketahui	0.718055	2.48E-04
Jumlah Kamar RS Kelas B	0.681147	5.91E-04
Banjir dan Tanah Longsor	0.662151	7.04E-04
Puting Beliung	0.654288	1.01E-03
Banjir	0.636063	1.07E-03
Tanah Longsor	0.603011	2.40E-03
Tenaga Gizi	0.596614	2.72E-03
Kekeringan	0.588231	4.10E-03
Jumlah RS Kelas D	0.566318	4.32E-03
Dokter Spesialis	0.55987	8.35E-03
Kebakaran	0.545997	1.93E-02
Tenaga Kesma	0.545941	2.53E-02
Puskesmas	0.543687	3.27E-02
Jumlah RS Kelas B	0.510535	4.16E-02
Tenaga Sanitasi	0.505018	4.36E-02
Keterapian Fisik	0.486467	4.75E-02
Pustu	0.483938	3.63E-05
Jumlah RS Kelas A	0.458475	2.04E-04
Jumlah Kamar RS Kelas A	0.451515	1.01E-03
Gelombang Pasang	0.406219	5.31E-04
Jumlah Kamar RS Kelas Tidak Diketahui	0.405398	9.51E-04
Gempa Bumi	0.400935	1.94E-03
Jumlah Kamar RS Kelas D	0.356606	2.08E-02
Gempa dan Tsunami	0.249464	1.61E-01

Kebakaran Hutan dan Lahan	0.181242	3.13E-01
Kecelakaan Transportasi	0.115737	5.21E-01
Letusan Gunung Api	0.080828	6.27E-01
Luas Wilayah	0.056683	7.54E-01
Konflik Sosial	-0.266276	1.34E-01

Tabel 4 pada kolom dengan warna kuning menyatakan sub-atribut yang akan dihapus yakni sub-atribut yang memiliki nilai korelasi kurang dari 0.5 dan nilai P-*Value* lebih dari 0.05. Sub-atribut yang tidak dihapus berjumlah 23 dan sub-atribut yang dihapus berjumlah 14.

V. KESIMPULAN

PCA dapat digunakan untuk memilih atribut mana yang akan digunakan pada proses klasterisasi. Dengan menggunakan PCA, dari ketiga percobaan rata-rata sub-atribut berkurang sampai 30%, sehingga atribut lebih sedikit dan waktu pengolahan menjadi lebih cepat dan efisien. Pada percobaan 1 dari 39 sub-atribut, 31 sub-atribut tidak dihapus dan 8 sub-atribut lainnya dihapus. Pada percobaan 2 dari 37 sub-atribut, 28 sub-atribut tidak dihapus dan 9 sub-atribut lainnya dihapus. Pada percobaan 3 dari 37 sub-atribut 23 sub-atribut tidak dihapus dan 14 sub-atribut lainnya dihapus.

DAFTAR PUSTAKA

[1] S. Briceno, Perkataan menjadi Tidakan: Panduan untuk implementasi kerangka kerja Hyogo, 2014.

[2] BNPB, "Data Bencana di Indonesia," BNPB, 2014.

[3] K. N. P. P. Nasional, Rencana Aksi Nasional Pengurangan Risiko Bencana 2006-2009, Jakarta: Perum Percetakan Negara RI, 2006.

[4] L. D. Discovering Knowledge in Data: An Introduction to *Data mining*, New Jersey: John Wiley & Sons, Inc., 2005.

[5] K. M. *Data mining* : Cencepts, Models, Methods and Algorithm., New Jersey: John Wiley & Sons Inc., 2003.

[6] R-Studio, "About R-Studio," R-Studio, 2014. [Online]. Available: www.rstudio.com/about. [Accessed 13 05 2015].

[7] J. Richard and D. Wichern, Applied Multivariate Statistical Analysis 6th Edition, New Jersey: Pretince Hall, 2007.

[8] S. Le, J. Josse and F. Husson, "FactoMineR : An R *Package* for Multivariate Analysis," *Jurnal of Statictical Software*, vol. 25, 2008.

[9] Muhidin, S. Ali and M. Abdurahman, Analisis korelasi, regresi, dan jalur dalam penelitian., Bandung: Pustaka Setia, 2007.

[10] K. K. R. Indonesia, Profil Kesehatan Indonesia, Jakarta: Kementrian Kesehatan Republik Indonesia, 2012.