

PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) DALAM PENENTUAN FAKTOR DOMINAN YANG MEMPENGARUHI PENGIDAP KANKER SERVIKS (Studi Kasus : *Cervical Cancer Dataset*)

Muhammad Zulfahmi Nasution¹, Adli Abdillah Nababan², Khairul Umam Syaliman³, Muhammad Syahputra Novelan⁴, Miftahul Jannah⁵

^{1,4} Sains dan Teknologi

^{2,5} Rekayasa Perangkat Lunak

³ Teknik Informatika

^{1,4} Universitas Pembangunan Pancabudi, Jl. Jendral Gatot Subroto KM. 4,5 Sei Sikambing 20122 Medan-Sumatera Utara-Indonesia

^{2,5} STMIK Pelita Nusantara, Jl. Iskandar Muda No. 1 Medan 20154, Sumatera Utara- Indonesia

³ Universitas Islam Riau, Jl. Kaharuddin Nst, Simpang Tiga, Kec. Bukit Raya, Kota Pekanbaru, Riau 28284 Indonesia

Email: fhm.nasution@dosen.pancabudi.ac.id

Abstract

Cancer diagnosis is a frightening issue for patients and can affect the patient's psychological condition. Therefore, by giving attention and psychosocial support to cancer patients it is expected to overcome the psychological pressure of the patient. Screening is an early detection effort to identify a disease or disorder that is clinically unclear by using certain tests, examinations or procedures. This effort can be used quickly to distinguish people who seem healthy but actually suffer from an abnormality. The purpose of this study is to simplify and eliminate some less relevant screening without reducing the intent and purpose of the original data using the Principal Component Analysis (PCA). Based on the results of research conducted from the UCI Cervical Cancer dataset repository, shows that there are 9 dominant screening variables that have a large enough correlation to the formation of early detection of cervical cancer with a proportion of 99% of covariance variance, including the 3 highest factors very dominant, namely the age factor with the first highest eigenvalue that is 76.05 with the proportion of variance 58.50%, cigarette addicts with 14.90% variance and Hormonal Contraceptives factor with 9.3% variance. The total variance obtained from the 9 screening variables is 99%.

Keywords: Cervical Cancer, Screening, Principal Component Analysis, Eigenvalue, Factor Analysis

Abstrak

Diagnosis kanker merupakan isu yang menakutkan bagi pasien dan dapat mempengaruhi kondisi psikologis pasien. Oleh karena itu, dengan memberikan perhatian dan dukungan psikososial kepada pasien kanker diharapkan dapat mengatasi tekanan psikologis pasien. Skrining merupakan upaya deteksi dini untuk mengidentifikasi penyakit atau kelainan yang secara klinis belum jelas dengan menggunakan tes, pemeriksaan atau prosedur tertentu. Upaya ini dapat digunakan secara cepat untuk membedakan orang-orang yang kelihatannya sehat tetapi sesungguhnya menderita suatu kelainan. Tujuan penelitian ini adalah untuk menyederhanakan dan menghilangkan beberapa variabel skrining yang kurang relevan tanpa mengurangi maksud dan tujuan dari data aslinya dengan menggunakan metode analisis Principal Component Analysis (PCA). Berdasarkan hasil penelitian yang dilakukan dari UCI repository of machine learning dataset yaitu: Cervical Cancer dataset menunjukkan bahwa terdapat 9 variabel skrining yang dominan memiliki korelasi cukup besar terhadap pembentukan deteksi dini penyakit kanker serviks dengan jumlah proporsi varians kovarians sebesar 99%, diantaranya 3 faktor tertinggi yang sangat dominan yaitu faktor usia dengan nilai eigen paling tinggi pertama yaitu 76.05 dengan proporsi varians sebesar 58.50%, faktor pecandu rokok dengan varians 14,90% dan faktor penggunaan kontrasepsi dengan media suntik (Hormonal Contraceptives) dengan varians 9,3%. Total varians yang diperoleh dari 9 variabel skrining tersebut adalah sebesar 99%.

Kata Kunci: Kanker Serviks, Skrining, Principal Component Analysis, Nilai Eigen, Faktor Analisis

1. Pendahuluan

Kanker leher rahim (*Cervical Cancer*) merupakan sejenis kanker dengan jumlah penderita tertinggi di Indonesia. Dalam rangka

pengendalian kanker tersebut, Kementerian Kesehatan (Direktorat Pengendalian Penyakit Tidak Menular, Subdit Pengendalian Penyakit Kanker). bekerja sama dengan lintas program terkait, pemerintah daerah, Lembaga Swadaya

Masyarakat (LSM), organisasi profesi, Female Cancer Program (FCP), Solidaritas Istri Kabinet Indonesia Bersatu (SIKIB), dan Organisasi Aksi Solidaritas Era Kabinet Kerja (OASE-KK), serta PKK untuk ikutserta mengembangkan program deteksi dini. Sejak Tahun 2007, Program deteksi dini sudah banyak dikembangkan dengan didahului pengembangan pada enam lokasi, kemudian ke daerah lain di seluruh Indonesia [1]

Deteksi penyakit kanker menjadi penting sebab untuk menurunkan kasus baru sehingga dibutuhkan upaya pencegahan dan deteksi dini sehingga mudah dilakukan ketika faktor risiko dan gejala kanker telah dikenali. Pada kuesioner Riset Kesehatan Dasar (Riskesdas) yang dilaksanakan oleh Badan Penelitian dan Pengembangan Kesehatan, Kementerian Kesehatan RI tahun 2013, salah satu pertanyaan adalah apakah penduduk pernah didiagnosis oleh dokter. Berdasarkan diskusi tersebut, didapatkan derajat penderita kanker pada penduduk semua umur di Indonesia sebesar 1,4%. Prevalensi kanker tertinggi berada pada Provinsi DI Yogyakarta, yaitu sebesar 4,1%, jauh lebih tinggi dibandingkan dengan angka nasional. Prevalensi tertinggi lainnya berada pada Provinsi Jawa Tengah dan Bali sebesar 2,1% dan 2,0%. Diagnosis kanker merupakan isu yang menakutkan bagi pasien dan dapat mempengaruhi kondisi psikologis pasien. Dengan demikian, memberikan perhatian dan psikososial kepada para pasien kanker hendaknya dapat mengatasi tekanan psikologis pasien, serta dapat menjaga kualitas hidupnya [2].

Berdasarkan data rutin Subdit Kanker Direktorat Penyakit Tidak Menular, Direktorat Jenderal Pengendalian Penyakit dan Penyehatan Lingkungan, Kementerian Kesehatan RI, Tahun 2013, program deteksi dini kanker serviks akan diselenggarakan pada 717 Puskesmas dari total 9.422 Puskesmas di 32 provinsi. Sehingga Puskesmas yang memiliki program deteksi dini masih sangat sedikit atau sekitar 7,6%. Estimasi pasien kanker serviks di Indonesia pada tahun 2013, dicatat pada Provinsi Jawa Timur, Jawa Tengah dan Jawa Barat memiliki jumlah pasien kanker serviks terbesar, Namun Provinsi Gorontalo dan Papua Barat memiliki estimasi jumlah penderita terkecil dari seluruh Provinsi [3].

Naiknya jumlah pasien kanker serviks di Indonesia sangat dipengaruhi dengan banyaknya jumlah pelaksana program, yang terdiri dari dokter umum dan bidan serta skrining pada Puskesmas. Hingga Tahun 2013, diperoleh sejumlah 1.682 pelaksana program deteksi dini kanker serviks di Indonesia dengan perkiraan jumlah kanker serviks sebesar 98.692 kasus. bahwa pelaksana program deteksi dini tertinggi

berpusat pada provinsi DKI Jakarta, Jawa Tengah dan Bali, sementara di provinsi lainnya seperti Kalimantan Selatan dan Sulawesi Utara masih minim adanya pelaksana program deteksi dini sementara jumlah penderita kanker di provinsi tersebut cukup tinggi [3].

Screening merupakan usaha untuk deteksi dini dalam mengidentifikasi jenis kelainan klinis yang belum bisa ditentukan dengan menggunakan tes, pemeriksaan atau prosedur tertentu. Cara ini dapat digunakan secara tepat guna memilih penderita dalam kondisi kelihatannya prima namun sesungguhnya menderita suatu penyakit klinis. Skrining kanker serviks dilakukan dengan tes IVA (Inspeksi Visual Asam Asetat). Jumlah skrining kanker serviks terbanyak terdapat pada Provinsi Jawa Barat, Jawa Tengah, dan Jawa Timur. Pada Provinsi Aceh, Riau, Jambi, Sumatera Selatan, Gorontalo, Maluku, dan Maluku Utara belum terdapat skrining, sedangkan estimasi jumlah penderita kanker serviks pada provinsi - provinsi tersebut cukup banyak [2].

Penelitian ini dilakukan dengan tujuan untuk menyederhanakan dan menghilangkan faktor atau indikator skrining yang kurang dominan atau relevan tanpa mengurangi maksud dan tujuan dari data aslinya menggunakan *Principal Component Analysis* (PCA).

PCA memberikan hasil yang baik ketika diterapkan pada atribut yang berkorelasi. Dalam penelitian ini, PCA diterapkan dalam pelatihan dan menguji atribut dari set data kanker serviks. PCA akan mengidentifikasi pola dalam kumpulan data, menemukan persamaan dan perbedaannya diantara masing-masing atribut [4]. Karena berfungsi sebagai model yang kuat untuk menganalisis data. Data asli dan rata-rata data asli dipilih. Matriks kovarians dihitung dimana hasilnya digunakan dalam menghitung vektor eigen dan nilai eigen [5] dan vektor eigen dengan nilai eigen tertinggi dipilih sebagai komponen utama dari data kanker serviks yang ditetapkan karena menunjukkan hubungan yang paling signifikan antara atribut kumpulan data. Nilai Eigen diurutkan dalam urutan menaik untuk memilih data yang paling signifikan dan membuang yang paling tidak signifikan. Dengan ini berarti data dengan ruang yang lebih banyak dikurangi ke ruang yang lebih rendah/sedikit.

Dengan demikian, PCA merupakan suatu teknik seleksi data multivariat (*multivariable*) yang mengubah atau mentransformasi suatu matriks data *original* menjadi suatu kumpulan kombinasi homogen yang lebih sedikit namun menyerap sejumlah besar varian dari data awal [6]. Tujuan utamanya ialah mendefinisikan sebanyak mungkin jumlah keragaman data

original dengan seminim mungkin *principal component* [7] [8].

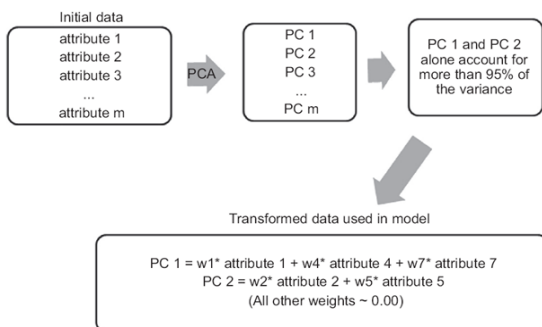
2. Teori

a. Principal Component Analysis (PCA)

PCA merupakan kombinasi linear dari variabel awal yang secara geometris kombinasi linear ini merupakan sistem koordinat baru yang diperoleh dari rotasi sistem semula. Metoda PCA sangat berguna digunakan jika data yang ada memiliki jumlah variabel yang besar dan memiliki korelasi antar variabelnya. Perhitungan dari *principal component analysis* didasarkan pada perhitungan nilai eigen dan vektor eigen yang menyatakan penyebaran data dari suatu dataset [5]. Dengan menggunakan PCA, variabel yang tadinya sebanyak n variabel akan diseleksi menjadi k variabel baru yang disebut *principal component*, dengan jumlah k lebih sedikit dari n . Dengan hanya menggunakan k *principal component* akan menghasilkan nilai yang sama dengan menggunakan n variabel. Variabel hasil dari seleksi disebut *principal component* [11].

PCA digunakan untuk menjelaskan struktur matriks varians-kovarians dari suatu set variabel melalui kombinasi linier dari variabel-variabel tersebut. Secara umum *principal component* (PC) dapat berguna untuk seleksi fitur dan interpretasi variabel-variabel.

Skema konseptual yang diilustrasikan bagaimana PCA dapat membantu untuk menyederhanakan dimensi dari data melalui hipotesa dataset berjumlah m variabel dapat ditunjukkan pada Gambar 2.1 berikut :



Gambar 1. Model konseptual PCA untuk tahap seleksi fitur (Sumber: Kotu & Deshpande, 2015)

Menurut (Jolliffe, 2002), Prosedur pengerjaan *Principal Component Analysis* bertujuan untuk menyederhanakan dan menghilangkan faktor atau indikator skrining yang kurang dominan dan kurang relevan tanpa mengurangi maksud dan tujuan dari data asli dari variabel acak x (matrik berukuran $n \times n$, dimana baris-baris yang berisi observasi sebanyak n dari variabel acak x) adalah

sebagai berikut:

1. Menghitung matrik varians kovarian dari data observasi.

Varians ($\text{Var}(x)$) dihitung untuk menemukan penyebaran data dalam set data kanker serviks untuk menentukan penyimpangan data dalam set data sampel [4]. Matrik **Kovarian** $\text{Cov}(x,y)$ ialah matriks yang nilai-nilai kovariansi pada tiap *cell*-nya diperoleh dari sampel. Misalkan x dan y adalah variabel acak. (Jolliffe, 2002).

$$\text{Var}(x) = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (Z_{ij} - \mu_j)^2$$

$$\text{Cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj})$$

Dengan μ_x dan μ_y merupakan rata – rata (*mean*) sampel dari variabel x dan y , dimana variabel x_i dan y_i merupakan nilai observasi ke- i dari variabel x dan y . Dari data nilai yang digunakan, maka diperoleh matrik kovarian berukuran $n \times n$.

2. Mencari *eigenvalues* dan *eigenvector* dari matrik kovarian yang telah diperoleh [5] yaitu: Nilai eigen dan vektor eigen untuk matriks kovarians dihitung. Nilai eigen yang dikomputasi kemudian ditransformasikan (*rotasi orthogonal varimax*) menggunakan persamaan berikut [5] :

$$\text{Det}(A - \lambda I) = 0$$

dimana :

A = matrik $n \times n$

λ = nilai *eigenvalue*

I = matriks identitas (matriks persegi dengan elemen diagonal utama bernilai 1 sedangkan elemen lain bernilai 0)

3. Menentukan nilai proporsi *Principal Component* (proporsi *Principal Component* (%)) dengan persamaan :

$$\text{PC}(\%) = \frac{\text{NilaiEigen}}{\text{VarianceCovarian}} \times 100\%$$

4. Menghitung bobot factor (*factor loading*) berdasarkan *eigenvector* dengan persamaan (Jhonson & Wichern, 2007) [5]:

$$Ax = \lambda x$$

Sehingga diperoleh kombinasi linear yaitu:

- a. $\lambda_1, \lambda_2, \lambda_3 \dots \lambda_n$ adalah *eigenvalue* matrik A
- b. $x_1, x_2, x_3 \dots x_n$ adalah *eigenvector* sesuai *eigenvalue*-nya (λ_n)

Persamaan *eigenvalue* & *eigenvector* merupakan *Eigen Value Decomposition* (EVD), dengan persamaan sebagai berikut:

$$AX = XD$$

$$A = X D X^{-1}$$

dimana :

A = matrik $n \times n$ yang memiliki n *eigenvalue* (λ_n)

D = *eigenvalue* dari *eigenvector*-nya

X = *eigenvector* dari matrik A

X^{-1} = invers dari *eigenvector* X

3. Metode Penelitian

a. Data Observasi

Data observasi yang digunakan dalam penelitian ini adalah dataset *Cervical Cancer* yang berasal dari [9]. Data kanker serviks ini melibatkan 858 sampel dengan 32 faktor serta empat label kelas meliputi : *Hinselmann*, *Schiller*, *Cytology* dan *Biopsy* yang telah dipublikasikan di [10]. Penelitian ini berfokus pada target *Biopsy* seperti yang direkomendasikan oleh tinjauan literatur.

Table 1. Cervical cancer data set description

Faktor	Type	Faktor	Type
Age	Int	Hormonal	Int
Number of sexual partnes	Int	Contraceptives (years)	Int
Age of first sexual Intercourse	Int	Intra Uterine Device (IUD)	Bool
Number of pregnancies	Int	IUD (years)	Int
Smokes	Bool	Sexually Transmitted Diseases (STDs)	Bool
Smokes (years)	Bool	STDs (number)	Int
Smokes (packs/year)	Bool	STDs: condylomatosis	Bool
Hormonal Contraceptives	Bool	STDs: cervical condylomatosis	Bool
STDs:vulvo-perineal condylomatosis	Bool	STDs: vaginal condylomatosis	Bool
STDs: syphilis	Bool	STDs: HPV	Bool
STDs: pelvic inflammatory disease	Bool	STDs: number of diagnosis	Int
STDs: genital herpes	Bool	STDs: time since first diagnosis	Int
STDs: mollusum contagiosum	Bool	STDs: time since last diagnosis	Int
STDs: AIDS	Bool	Dx: Cancer	Bool
STDs: HIV	Bool	Dx: CIN (cervical intraepithelial neoplasia)	Bool
STDs: Hepatitis B	Bool	Dx: HPV (Human Papillomavirus)	Bool
		Dx (diagnosis diseases)	Bool

b. Tahapan Penelitian

Berikut penjelasan tahapan penelitian yang dapat dijabarkan sebagai berikut :

1) Data Preprocessing

Pada penelitian ini, hanya melakukan dua proses *preprocessing*. Pertama adalah penanganan *missing value*. *Missing value* pada faktor yang bernilai numerik digantikan dengan nilai rata-rata (*mean*) dari faktor pada kolom yang sama. Sedangkan *missing value* pada faktor yang bernilai nominal digantikan dengan nilai kemungkinan terbanyak dari faktor pada kolom yang sama. Selanjutnya adalah proses *cleaning* dilakukan dengan membuang duplikasi data sehingga jumlah data observasi yang semula sebanyak 858 record menjadi 829 record.

Hasil *preprocessing* direpresentasikan kedalam bentuk label ($x_1, x_2, x_3 \dots x_{32}$) yang mewakili urutan data yang sesuai dengan dataset yang diuji. Berikut hasil representasi faktor label:

- X_1 : Age
- X_2 : Number of sexual partners
- X_3 : First sexual intercourse
- X_4 : Number of pregnancies
- X_5 : Smoke
- X_6 : Smoke (years)
- X_7 : Smoke (packs/year)
- X_8 : Hormonal contraceptives
- X_9 : Hormonal contraceptives (years)
- X_{10} : IUD

- X_{11} : IUD (years)
- X_{12} : STDs
- X_{13} : STDs (number)
- X_{14} : STDs: condylomatosis
- X_{15} : STDs: cervical condylomatosis
- X_{16} : STDs: vaginal condylomatosis
- X_{17} : vulvo-perineal condylomatosis
- X_{18} : STDs: syphilis
- X_{19} : pelvic inflammatory disease
- X_{20} : genital herpes
- X_{21} : STDs: mollusum contagiosum
- X_{22} : STDs: AIDS
- X_{23} : STDs: HIV
- X_{24} : STDs: Hepatitis B
- X_{25} : STDs: HPV
- X_{26} : STDs: Number of diagnosis
- X_{27} : STDs: Time since first diagnosis
- X_{28} : STDs: Time since last diagnosis
- X_{29} : Dx: Cancer
- X_{30} : Dx: CIN
- X_{31} : Dx: HPV
- X_{32} : Dx
- X_{33} : Class Biopsy

2) Menghitung korelasi antar data observasi

Kovarian digunakan untuk mengukur besarnya hubungan antara dua faktor. Matrik kovarian (*covariance matrix*) diperoleh dengan menghitung pasangan 2 buah faktor dari total n faktor menggunakan persamaan **Cov (x,y)**
Misalkan: Atribut x, y dan z , dengan jumlah faktor = 3,

$$\begin{matrix} & x & y & z \\ \begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \end{matrix}$$

Maka matrik kovariannya berukuran $3 \times 3 = n \times n$, yaitu:

$$\begin{bmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{bmatrix}$$

Dimana $\text{cov}(x, x)$, $\text{cov}(y, y)$ dan $\text{cov}(z, z)$ nilainya sama dengan perhitungan nilai variansi (*variance*) dari atribut x , y dan z . sehingga matrik kovariannya:

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{var}(y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{var}(z) \end{bmatrix}$$

Jika nilai komponen diagonalnya ($\text{var}(x)$, $\text{var}(y)$ dan $\text{var}(z)$) adalah 1, maka matrik kovariannya sama dengan matrik korelasi sebab hasil korelasi dari data yang tidak distandarasi adalah sama dengan hasil kovarian dari data yang distandarasi. Berarti bahwa matrik kovarian merupakan matrik simetris. Bila didekomposisikan dengan *Eigenvalue*, maka nilai pada diagonal *eigenvalue* ≥ 0 . Dengan demikian nilai:

$$\begin{aligned} \text{cov}(y, x) &= \text{cov}(x, y) \\ \text{cov}(z, x) &= \text{cov}(x, z) \\ \text{cov}(z, y) &= \text{cov}(y, z) \end{aligned}$$

3) Menentukan *eigenvalue* dari *covariance matrix*

Nilai eigen (λ) merupakan bilangan skalar yang mendefinisikan matrik kovarian. Dimana matrik kovarian tersebut adalah matriks bujur sangkar (*square matrix*) berukuran $m \times m$, maka untuk nilai eigen (λ) yang sesuai dengan matrik kovarian tersebut diperoleh melalui persamaan eigenvektor, sehingga setiap skalar ($\lambda_1, \lambda_2, \dots, \lambda_m$) membentuk matrik eigenvektor. Proses rotasi faktor yang

digunakan adalah *varimax rotation* dengan nilai *loading factor* lebih besar dari 0,4.

4) Memilih sejumlah *principal component (PC)*

Komponen utama yang dipilih memiliki jumlah kontribusi maksimum sebesar 99% berdasarkan proporsi varians dari masing-masing *principal component* yang terpilih maka terlebih dahulu dihitung nilai proporsi varians dari masing-masing *principal component* yang diperoleh melalui persamaan:

$$PC (\%) = \frac{\text{NilaiEigen}}{\text{VarianceCovarian}} \times 100\%$$

Dimana, Varians Kovarians diperoleh dari penjumlahan total nilai diagonal matrik kovarian.

5) Menghitung bobot faktor (*factor loading*) menggunakan persamaan eigenvektor

Untuk menentukan atribut mana yang termasuk kedalam *principal component* dengan jumlah kontribusi maksimum sebesar 99%, maka dilakukan perhitungan bobot nilai eigenvektor melalui persamaan eigenvektor, dimana setiap nilai vektor x yang terbentuk bersesuaian dengan satu nilai eigen (λ) untuk masing-masing faktor.

6) Memilih nilai *factor loading* menggunakan perbandingan besar korelasi

Interpretasi hasil dilakukan dengan melihat nilai dari *factor loading* yang terdapat dalam matriks faktor. *Factor loading* adalah angka yang menunjukkan besarnya korelasi antara suatu variabel dengan faktor satu, dua, atau faktor tiga yang terbentuk. Proses penentuan variabel mana yang akan masuk kedalam faktor dilakukan dengan melihat perbandingan besar korelasi pada setiap baris didalam setiap tabel matrik faktor.

4. Hasil Penelitian

Penelitian menggunakan perangkat lunak Rapid Miner® versi 5.3 dan spesifikasi processor Intel Celeron 1.60GHz, kapasitas memory 2 GB. Untuk memudahkan perhitungan nilai-nilai dalam proses perhitungan PCA, maka menggunakan perangkat lunak Matlab versi 2013.

a. Hasil Data Preprocessing

Tabel 2. Hasil Data Preprocessing

No	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	18	4	15	1	0	0	0	0	0	0	...	0
2	15	1	14	1	0	0	0	0	0	0	...	0
3	34	1	17	1	0	0	0	0	0	0	...	0
4	52	5	16	4	1	37	37	1	3	0	...	0
5	46	3	21	4	0	0	0	1	15	0	...	0
6	42	3	23	2	0	0	0	0	0	0	...	0
7	51	3	17	6	1	34	3	0	0	1	...	1
8	26	1	26	3	0	0	0	1	2	1	...	0
9	45	1	20	5	0	0	0	0	0	0	...	0

No	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
10	44	3	15	2	1	1	3	0	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
829	29	2	20	1	0	0	0	1	0	0	...	0

b. Hasil Perhitungan Varians Kovarian

Tabel 3. Hasil Perhitungan Varians Kovarian

No	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	...	X ₃₂
X ₁	71.97	0.97	8.55	6.18	0.13	7.44	...	0.13
X ₂	0.97	2.75	-0.72	0.15	0.13	1.18	...	0.00
X ₃	8.55	-0.72	7.90	-0.27	-0.13	-0.72	...	0.01
X ₄	6.18	0.15	-0.27	1.99	0.03	0.98	...	0.00
X ₅	0.13	0.13	-0.13	0.03	0.12	1.05	...	-0.00
X ₆	7.44	1.18	-0.72	0.98	1.05	17	...	-0.04
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X ₃₂	0.12	0.00	0.01	0.00	-0.00	0.03	...	0.03

c. Hasil Perhitungan Eigenvalue

Tabel 4. Hasil Perhitungan Eigenvalue

PC	Nilai Eigen	Proporsi Varian (%)	Cumulative
1	76.0523	58.50	58.50
2	19.3252	14.90	73.40
3	12.1337	9.30	82.70
4	6.7188	5.20	87.90
5	5.4293	4.20	92.00
6	3.1804	2.40	94.50
7	2.5463	2.0	96.40
8	2.1140	1.60	98.10
9	1.2614	1.00	99.00

Variance Threshold = 99%

d. Hasil Perhitungan Eigenvector

Tabel 5. Hasil Perhitungan Eigenvector

Faktor	PC 1	PC 2	PC 3	PC 4	PC 5	...	PC 32
X ₁	0.969	-0.117	-0.106	0.133	0.032	...	-0.00
X ₂	0.015	0.08	0.009	0.151	0	...	-0.00
X ₃	0.12	-0.162	-0.202	-0.917	-0.046	...	0.00
X ₄	0.085	0.017	0.037	0.174	0.015	...	0.00
X ₅	0.004	0.057	0.004	-0.003	0.002	...	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X ₃₂	0.002	-0.003	-0.002	0.003	-0.006	...	0.00

e. Hasil Akhir PCA

Tabel 6. Hasil Akhir PCA

No.	PC	Faktor	Label	Loading
1.	PC 1	X ₁	Age	0.969
2.	PC 2	X ₆	Smoke (years)	0.887
3.	PC 3	X ₉	Hormonal contraceptives (years)	0.969
4.	PC 4	X ₃	First sexual intercourse	0.917
5.	PC 5	X ₂₇	STDs: Time since first diagnosis	0.715
6.	PC 5	X ₂₈	STDs: Time since last diagnosis	0.695
7.	PC 6	X ₁₁	IUD (years)	0.959
8.	PC 7	X ₂	Number of sexual partners	0.920
9.	PC 8	X ₇	Smoke (packs/year)	0.876
10.	PC 9	X ₄	Number of pregnancies	0.973

5. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan dari *UCI Machine Learning Cervical Cancer dataset repository* bahwa deteksi dini penyakit kanker serviks cukup melewati 9 variabel skrining yang paling dominan dari sejumlah 32 variabel skrining karena hanya 9 variabel skrining tersebut yang memiliki korelasi cukup besar terhadap pembentukan deteksi dini penyakit kanker serviks dengan proporsi varians kovarians sebesar 99%, yaitu diantaranya 3 faktor tertinggi yang paling dominan yaitu faktor usia dengan nilai proporsi varians sebesar 58.50%, faktor pecandu rokok dengan varians 14,90% dan faktor penggunaan kontrasepsi dengan media suntik (*Hormonal Contraceptives*) dengan varians 9,3%. Total varians yang diperoleh dari 9 variabel skrining tersebut adalah sebesar 99%.

Acknowledgement

Penghargaan tinggi sudah semestinya diberikan kepada segenap Lembaga Penelitian Universitas Pembangunan Pancabudi Medan (LPPM UNPAB), Rektor Universitas Pembangunan Pancabudi Medan, Fakultas Sains dan Teknologi, Program Studi Sistem Komputer, Atas dukungan terhadap penyebaran karya penelitian ini serta fasilitas yang telah disediakan.

6. References

- [1] Kementerian Kesehatan RI. 2015. *Buletin Jendela Data dan Informasi Kesehatan*. ISSN 2088 – 270X.
- [2] Wahidin, M. 2015. *Deteksi Dini Kanker Leher Rahim di Indonesia*.
- (Online) <http://www.pusdatin.kemkes.go.id> (Diakses : tanggal 29 Mei 2019).
- [3] <http://www.depkes.go.id/folder/view/01/structure-publikasi-pusdatin-info-datin.html>
- (Online) Diakses : tanggal 29 Mei 2019
- [4] Jolliffe, I.T. 2002. *Principal Component Analysis*. 2nd Edition. Springer-Verlag: New York.
- [5] Johnson, W.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. 6th Edition. Pearson Prentice Hall: New Jersey.
- [6] Hussain, H., Quazilbash, N.Z., Bai, S. & Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis. *International Conference on Computer-Based Medical Systems*, pp. 131-134.
- [7] Maaten, L., Postma, E. & Herik, J. 2009. Dimensionality Reduction: A Comparative Review. (Online) <http://www.uvt.nl/ticc> (Diakses : tanggal 29 Mei 2019)
- [8] Dang, T., Pham, T., Tran, H. & Van, Q. 2016. Using Dimension Reduction with Feature Selection to Enhance Accuracy of Tumor Classification. *International Conference on Biomedical Engineering*, pp.14-17.
- [9] <https://archive.ics.uci.edu/ml>, Center for Machine Learning and Intelligent Systems, University of California, Department of Information and Computer Science. 1998
- [10] Fernandes, K., J.S. Cardoso, and J. Fernandes., 2017, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening" In *Iberian Conference on Pattern Recognition and Image Analysis*, Springer.
- [11] Kotu, V. & Deshpande, B. 2015. *Predictive Analytics and Data Mining*. Morgan Kaufmann Publisher: San Francisco.