ANALISIS SUPPORT VECTOR MACHINE (SVM) DAN PRINCIPAL COMPONENT ANALYSIS (PCA) ALTERNATIF UNTUK DETEKSI KANKER SERVIKS

Octaviani Intan Riadi,
Teknik Informatika Universitas Ma Chung, Villa Puncak Tidar N-1 Malang Email : 311610014@student.machung.ac.id

Abstrak

Penyakit kanker adalah salah satu penyakit yang banyak memakan korban jiwa. Pada kaum wanita, kanker payudara dan kanker serviks adalah salah satu jenis kanker yang paling banyak dijumpai. Tingginya angka korban jiwa pada kanker payudara dan kanker serviks salah satunya disebabkan oleh keterlambatannya penanganan kanker. Banyak penderita kanker yang tidak menganggap serius terhadap gejala yang ditemui dan akhirnya penanganan yang diberikan menjadi terlambat. Oleh sebab itu penelitian ini bertujuan untuk menganalisa metode pendeteksian kanker dengan cara mereduksi jumlah atribut menggunakan PCA dan SVM yang paling optimal. Penelitian ini menggunakan library python untuk membantu proses analisis menggunakan PCA dan SVM. Dataset akan diolah dengan menggunakan metode normalisasi *MinMaxScaling* dan standarisasi. Yang selanjutnya akan digunakan perhitungan sensitivitas dan spesifisitas untuk mencari seberapa akurat model yang dihasilkan SVM. Dengan menggunakan 3 kernel *yaitu radial basis function* (RBF), *sigmoid* dan *polynomial* didapatkan bahwa menggunakan PCA dapat mereduksi dengan baik variabelvariabel yang sudah ada. Hasil pada kernel RBF didapatkan akurasi tertinggi 97.76% di model jumlah PC 30 hingga 31, pada kernel *polynomial* didapatkan 99.25% di model jumlah PC 21 dan pada kernel *sigmoid* didapatkan 99,25% di model jumlah PC 13 hingga 14 dan beberapa model PC lainnya. Dari hasil yang didapatkan penggunaan PCA dapat membantu meningkatkan akurasi pada beberapa kernel dalam studi kasus kanker serviks.

Kata kunci: *MinMaxScaling, Principal Component Analysis,* Standarisasi, *Support Vector Machine*.

Abstract

Cancer is a disease that takes many lives. In women, breast cancer and cervical cancer are one of the most common types of cancer. The high number of fatalities in breast cancer and cervical cancer is one of them caused by the delay in handling cancer. Many cancer sufferers do not take seriously the symptoms that are found and finally the treatment given is too late. Therefore this study aims to analyze cancer detection methods by reducing the number of attributes using the most optimal PCA and SVM. This research uses the python library to help the analysis process using PCA and SVM. The dataset will be processed using the MinMaxScaling normalization method and standardization. The next calculation will be used sensitivity and specificity to find how accurate the model produced by SVM. By using 3 kernels, namely radial basis function (RBF), sigmoid and polynomial, it is found that using PCA can reduce well existing variables. The results in the RBF kernel obtained the highest accuracy of 97.76% in the PC number model 30 to 31, the polynomial kernel obtained 99.25% in the PC number 21 model and the sigmoid kernel obtained 99.25% in the PC number 13 to 14 model and several other PC models. From the results obtained using PCA can help improve accuracy in some kernels in cervical cancer case studies.

Keywords*: MinMaxScaling, Principal Component Analysis, Standardization, Support Vector Machine*

Pendahuluan

Teknologi pada abad 21 saat ini sudah sangat berkembang baik dalam hal perindustrian, komunikasi, militer maupun keasehatan. Di berbagai bidang, teknologi sudah sangat membantu mempercepat dan mengefisienkan pekerjaan yang ada. Dalam bidang kesehatan hampir semua pemeriksaan kesehatan menggunakan teknologi-teknologi yang ada untuk membantu para dokter melihat penyakit yang diderita pasien. Kanker termasuk salah satu penyakit yang sudah menelan banyak korban jiwa. Menurut kompas.com dengan tajuk "Kanker Penyakit Kronik yang Merusak Tubuh, Apa Saja Penyebabnya?" yang terbit pada tanggal 30 Januari 2020 disebutkan bahwa pada tahun 2019 kanker telah memakan korban jiwa sebanyak 103.000 pria dan 92.200 wanita oleh WHO.

Sudah ada beberapa penelitian tentang deteksi kanker menggunakan kecerdasan buatan untuk memprediksi kanker dari beberapa kondisi yang ada, salah satunya adalah deteksi kanker berdasarkan klasifikasi Microarray Data menggunakan Back Propagation Neural Network dan PCA (Adwijaya 2018). Hasil akurasi yang didapatkan sudah baik dengan akurasi mencapai lebih dari 80% dengan waktu *training* 0-4 detik. Ada juga sebuah jurnal yang membahas "Implementasi Support Vector Machine dan Random Forest pada Diagnosis Kanker Payudara" (Aliady 2018) dengan hasil metode random forest menghasilkan akurasi hingga 94.5%. Penelitian-penelitian yang dilakukan sudah mencapai akurasi diatas 80%. Pada penelitian oleh aliady menggunakan PCA untuk membantu mengolah dataset sebelum dilakukan pengklasifikasian. Sementara pada penelitian aliady menggunakan SVM dan random forest untuk langsung dilakukan pengklasifikasian data. Oleh sebab itu penelitian ini akan membuat sistem pendeteksian kanker khususnya kanker serviks menggunakan metode SVM dan PCA untuk metode pendeteksiannya berdasarkan beberapa penelitian terdahulu yang sudah ada. Tujuan penelitian ini untuk melihat hasil akurasi PCA dengan SVM yang paling optimal dengan mereduksi atribut-atribut yang ada menggunakan PCA.

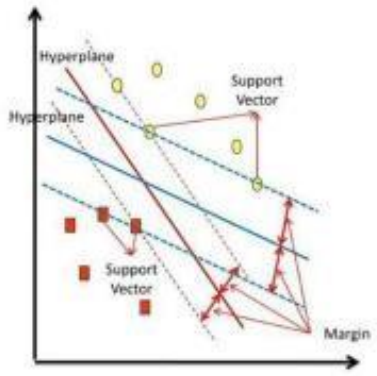
Tinjauan Pustaka

Kanker adalah salah satu penyakit paling ganas di seluruh dunia. Salah satu jenis kanker yang sering dijumpai pada kaum hawa ialah kanker serviks. Kanker serviks memiliki angka tertinggi setelah kanker payudara yang menyerang kaum hawa. Menurut WHO ada sebanyak 18,1 juta kasus kanker baru dan 9,6 juta kematian yang terjadi pada tahun 2018 (Juniman 2018). Penyebab utama tingginya angka

kematian pada penyakit kanker ialah keterlambatan penanganannya. Lebih dari 70% penderita kanker serviks yang datang berobat ke sebuah rumah sakit sudah pada stadium tinggi, yaitu stadium II dan III (Nurrana dan Purwoto 2000). Kanker serviks biasanya dikaitkan dengan Human Papilloma Virus (HPV). Ada sekitar 100 jenis virus HPV, tapi hanya jenis tertentu yang bisa menjadi penyebab kanker serviks. Ada dua jenis virus HPV paling umum yang menjadi penyebab kanker serviks yaitu HPV-16 dan HPV-18 (Setiaputri 2020).

Kecerdasan buatan adalah salah satu kemajuan teknologi saat ini, dengan kecerdasan buatan proses pengerjaan suatu masalah dapat terbantu terselesaikan dengan baik. Menurut Rich dan Knight pada tahun 1991, Kecerdasan Buatan (AI) merupakan sebuah studi tentang bagaimana membuat komputer melakukan hal-hal yang pada saat ini dapat dilakukan lebih baik oleh manusia. Salah satu percabangan dari kecerdasan buatan ini ialah *machine learning. Machine learning* adalah salah satu cabang ilmu pengetahuan yang menggabungkan gagasan dari beberapa cabang ilmu pengetahuan lainnya seperti kecerdasan buatan, statistik, teori informasi, matematika dan lain-lain (He, et al. 2014). Di *machine learning* sendiri ada 3 penjurusan ilmu besar secara umum yang sampai sekarang terus berkembang, yaitu *Neural Network* (NN), *Fuzzy Logic System* dan *Support Vector Machine* (SVM).

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) yang berdimensi tinggi (Joachims 1998). SVM disini membantu sebuah program menyelesaikan masalah untuk menentukan pengklasifikasian data. Ada beberapa parameter yang digunakan dalam SVM ini yaitu *constant*, kernel, *degree*, *gamma*, *coefisien*, *cache size* dan *max iteration*. Parameter ini akan digunakan untuk membantu menghitung hasil prediksi nantinya. Cara kerja SVM ialah dengan mencari *hyperplane* terbaik dari 2 kelas yang berbeda. *Hyperplane* adalah fungsi yang digunakan untuk memisahkan kedua kelas. Kelas yang dimaksud ialah hasil output yang nantinya akan dihasilkan. Dengan mendapat *hyperplane* maka data yang ada bisa dikelompokan sesuai dengan bagiannya. Gambaran *hyperplane* ini dapat dilihat pada gambar 1.



Gambar 1. Cara kerja Hyperplane (Susanto 2018)

Dalam pengklasifikasian data belum tentu posisi setiap data menyebar sama rata, bisa saja penyebaran tidak seimbang sehingga pemisahan kelas ini tidak bisa dilakukan secara 2 dimensi. Pemisahan 2 kelas ini bisa dilakukan dalam bentuk 3 dimensi dimana garis hyperplane ditarik secara 3 dimensi untuk pengelompokan data yang ada. Oleh sebab itu untuk data yang memiliki posisi yang tidak konstan perlu menggunakan bantuan kernel untuk mecari posisi tiap data dengan pasti yang akhirnya bisa diolah nantinya. Ada beberapa rumus kernel yang biasanya digunakan untuk membantu perhitungan SVM, rumus tersebut bisa dilihat pada tabel 1.

Tabel 1. Rumus Beberapa Kernel

Kernel	Persamaan
Polynomial	
Gaussian	
Sigmoid	
Gaussian Radial Basis Function	

Selanjutnya ialah penggunaan Principal Component Analysis (PCA) yang dapat membantu mengolah sebuah data. PCA adalah suatu teknik handal untuk mengekstrasi stuktur dari suatu set data dengan dimensi yang cukup banyak (Santosa 2007). Tujuan utama dari proses PCA disini ialah mengekstraksi informasi yang paling penting dalam suatu data yang ada. Nilai dari proposi *principal component* (PC) dapat dihitung dengan rumus sebagai berikut pada 1-1:

(1-1)

Untuk dapat menghitung *principal component*, data *varian covariance* dan nilai eigen harus ditemukan terlebih dahulu. Nilai *varian covarian*ce dapat dicari dengan perhitungan sebagai berikut pada 1-2 dan 1-3:

(1-2)

(1-3)

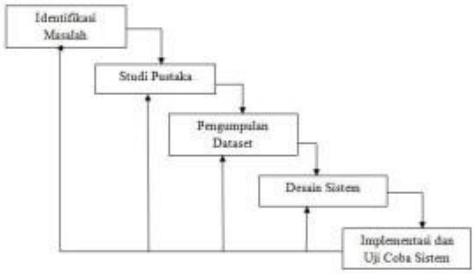
Dimana adalah rata-rata sampel dari x dan y, sementara nilai x dan y adalah nilai observasi yang ke i. Output dari data nilai ini adalah matriks kovarian dengan ukuran n x n. Selanjutnya untuk perhitungan nilai eigen dapat dilihat sebagai berikut pada 1-4:

(1-4)

Dimana adalah matriks covarian, λ adalah nilai eigen dan adalah matriks identitas.

Analisis dan Perancangan Sistem

Desain perancangan sistem ini menggunakan metode *waterfall* atau yang biasa sering disebut *classic life cycle*. Metode *waterfall* merupakan model pengembangan sistem informasi yang sistematik dan sekuensial (Roger S. Pressman 2012). Ada 5 tahapan pengerjaan yang akan dilakukan, tahapan ini dapat dilihat lebih jelas pada gambar 2.

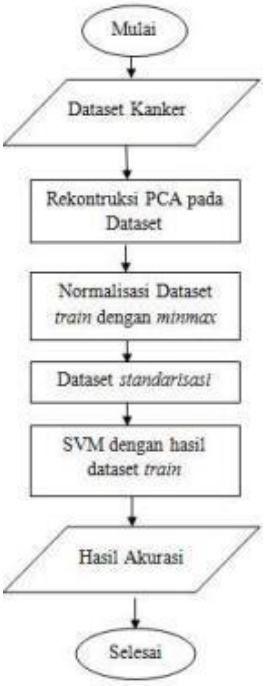


Gambar 2. Waterfall Penelitian

Pada identifikasi masalah ini akan dilakukan beberapa pengujian dari referensi beberapa penelitian terdahulu tentang analisis prediksi deteksi kanker serviks. Sebagian besar penelitian terdahulu melakukan analisis menggunakan *support vector machine* dengan beberapa metode tambahan lainnya. Akurasi yang didapatkan beragam dengan beberapa metode yang ada, walaupun data yang diidentifikasi tidak sama antara beberapa peneliti akan tetapi akurasi yang dihasilkan bisa mencapai ratarata di atas 70% pada setiap pengujian.

Pada tahap studi pustaka ini, semua kebutuhan yang nantinya akan dibutuhkan sistem ditelusuri lebih lanjut lagi. Dalam penelitian ini akan digunakan metode PCA dan SVM sebagai metode untuk menganalisa prediksi kanker serviks. Lalu pada bagian dataset akan dilakukan metode normalisasi data untuk memudahkan data-data yang ada diolah lebih lanjut. Inputan dalam proses sistem ini adalah dataset kanker serviks. Dataset yang digunakan berasal dari website "UCI Machine Learning Repository" yang memiliki dataset *Cervical Cancer*. Atribut yang ada pada dataset ini berjumlah 36 variabel yang nantinya hasil atau *output* dari dataset ini ialah terkenanya atau tidak terkenanya sesorang dengan penyakit kanker serviks dalam bentuk *type* data boolean 0 dan 1, dimana 0 bernilai *false* dan 1 bernilai *true*. Jumlah data yang ada sebanyak 858 data.

Desain sistem yang nantinya akan digunakan ialah suatu sistem yang akan menerima inputan berupa beberapa hasil data dari atribut-atribut yang sudah ada di dalam dataset yang berjumlah 36 atribut dengan banyak data 858 data. Sebelum dataset ini akan digunakan, akan dilakukan seleksi terlebih dahulu untuk data-data yang memiliki nilai data kosong. Dalam penelitian ini data kosong akan dipisahkan dan tidak akan digunakan dalam proses *training* maupun proses *test* nantinya. Selanjutnya akan dilakukan proses PCA untuk mencari atribut-atribut apa saja yang sangat berpengaruh dalam penentuan prediksi ini. Hasil dari dataset ini akan dinormalisasi dan distandarisasi untuk membuat rentangan data tidak terlalu jauh antara satu dengan yang lain, proses ini akan membantu perhitungan SVM menjadi lebih baik. Setelah dataset sudah diolah, selanjutnya akan dilakukan proses *training* dengan metode SVM. Data *train* yang telah diolah dengan PCA akan diolah lagi dengan SVM sehingga model dari *training* ini bisa didapatkan. Jika sudah mendapat model yang dicari, data *test* akan digunakan untuk analisis seberapa besar akurasi yang didapatkan dengan hasil model yang sudah ada. Proses sistem yang akan digunakan dapat dilihat pada *flowchart* gambar 3.



Gambar 3. *Flowchat* Sistem

Selanjutnya pengujian sistem ini dilakukan dengan menginputkan dataset *test* yang ada, hasil dari pengujian ini akan dibandingkan dengan hasil akhir sebenarnya. Dataset yang ada akan dibagi menjadi 2 data yaitu dataset *train* dan dataset *test* dengan jumlah perbandingan data 80%:20% data. Faktor-faktor yang akan di uji coba ialah:

1. Jumlah komponen pada PCA yang paling optimal untuk ke 36 atribut yang ada di dataset kanker serviks. 2. Hasil akurasi yang didapatkan dengan beberapa variasi PCA dan SVM akan dibandingkan untuk mendapat nilai paling optimal.

Untuk melihat seberapa besar klasifikasi sistem ini berhasil melakukan pengklasifikasian, perhitungan sensitivitas dan spesifisitas diperlukan untuk analisis ini yang dapat dilihat pada tabel 2.

Tabel 2.Daftar Kondisi

Hasil Prediksi	Kondisi Sesungguhnya	
	Sakit Kanker	Tidak Sakit Kanker
Sakit Kanker	*True Positive*	*False Positive*
Tidak Sakit Kanker	*False Negative*	*True Negative*

Hasil prediksi adalah hasil dari *output* sistem analisis ini dengan data *test* yang akan dibandingkan dengan *output* sesungguhnya. Dimana *True Positive* berarti data sakit kanker berhasil dideteksi sakit kanker oleh sistem dan *False Negative* jika data sakit kanker gagal dideteksi sebagai tidak sakit kanker oleh sistem. Begitu juga dengan *False Positive* dimana data tidak sakit kanker berhasil dideteksi sebagai tidak sakit kanker dan *True Positive* jika data tidak sakit kanker gagal dideteksi sebagai data yang sakit kanker oleh sistem. Selain menggunakan sensitivitas dan spesifisitas, akan dilakukan uji korelasi antara jumlah PC dan akurasi SVM yang dihasilkan. Rumus korelasi yang digunakan dapat dilihat pada 1-5 dibawah ini.

(1-5)

Dimana adalah jumlah PC, adalah akurasi SVM yang didapatkan dan adalah jumlah banyaknya pasangan PC.

Hasil dan Pembahasan

Dataset yang berjumlah 858 data ini akan di periksa terlebih dahulu untuk beberapa variabel yang memiliki nilai kosong. Di antara 36 atribut ini, ada 2 atribut yang hampir semua data tidak memiliki nilai. Kedua atribut ini adalah atribut STDs: times since first diagnosis dan STDs: times since last diagnosis. Oleh karena itu, kedua atribut ini tidak digunakan dalam proses ini dan hanya 34 atributyang akan digunakan. Lalu dari jumlah 858 data yang ada, hanya ada 668 data yang tidak memiliki nilai kosong di dalam setiap atributnya. Dataset ini akan dibagi sesuai yang disebutkan sebelumnya menjadi 80%:20% yaitu 534 data *training* dan 134 data *test*.

Selanjutnya akan dilakukan proses PCA pada dataset yang sudah dibersihkan. Hasil dari proses ini dapat dilihat pada tabel 3.

Tabel 3. Hasil PCA

No	PC	Kontribusi (%)	Kumulatif PPV (%)
1	1	62.99	62.99
2	2	14.18	77.17
3	3	10.03	87.20
4	4	4.94	92.14
5	5	2.62	94.76
6	6	2.02	96.79
7	7	1.47	98.26
8	8	0.85	99
9	9	0.36	99
10	10	0.0013	99
11	11	0.0011	99
12	12	0.00044	99
13	13	0.00039	99
14	14	0.00033	99
15	15	0.00030	99
16	16	0.00029	99
17	17	0.00018	99
18	18	0.00015	99
19	19	0.00012	99
20	20	0.00007	99
21	21	0.00004	99
22	22	0.00003	99
23	23	0.00002	99
24	24	0.00001	99
25	25	0.00001	99
26	26	0.00001	99
27	27	0.00001	99
28	28	0.00001	99
29	29	0.000005	99
30	30	0.000002	99
31	31	0.0000001	99
32	32	0.0000001	100
33	33	0.0000001	100

Pada tabel 3 terlihat dari tiap atribut pertama hingga atribut ke 33 memberikan kontribusi dalam penyimpanan data yang berbeda-beda. Pada atribut pertama yaitu usia berhasil menyimpan sebanyak 62,99% keseluruhan informasi yang ada. Dilanjutkan dengan atribut lainnya, yang memiliki peran masing-masing sesuai dengan jumlah kontribusi yang ada.

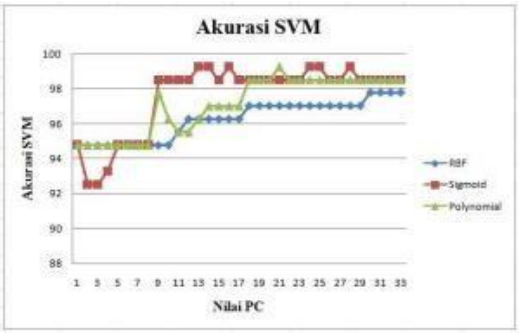
Pada proses selanjutnya, hasil nilai PCA akan ditambahkan dengan rata-rata tiap atribut yang ada untuk melihat hasil rekonstruksi pada setiap atribut dataset. Dapat dilihat pada tabel 4 adalah hasil dari setiap rata-rata atribut dan hasil dari *principal component* 1.

Tabel 4. Hasil PCA pada PC 1

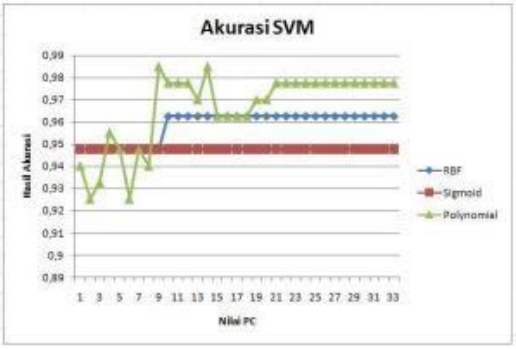
No	Atribut	Nilai Rata-rata	Hasil PCA PC 1
1	1	27.220	6.04
2	2	2.589	0.111
3	3	17.041	0.68
4	4	2.376	0.62
5	5	0.147	0.02
6	6	1.231	0.79
7	7	0.421	0.26
8	8	0.657	0.04
9	9	2.412	0.85
10	10	0.112	0.05
11	11	0.512	0.25
12	12	0.106	0.003
13	13	0,176	-0.002
14	14	0.058	-0.002
15	15	0	0
16	16	0.003	0.0008
17	17	0.056	-0.001
18	18	0.262	0.0009
19	19	0.001	0.0006
20	20	0.001	-0.001
21	21	0.001	-0.00009
22	22	0	0
23	23	0.022	-0.0002
24	24	0.001	-0.0007
25	25	0.001	0.0008
26	26	0.024	0.01
27	27	0.003	-0.0009
28	28	0.022	0.01
29	29	0.020	0.003
30	30	0.046	0.009
31	31	0.099	0.02
32	32	0.054	-0.002
33	33	0.069	0.01

Hasil rekontruksi dataset didapatkan dari menjumlahkan hasil rata-rata atribut dengan hasil *principal component* yang sudah didapatkan. Seperti pada atribut pertama yaitu usia, pada hasil PCA di PC 1 usia mendapat hasil sebesar 6.04 dan untuk rata-rata pada atribut usia didapatkan nilai sebesar 27.22. Kedua nilai ini akan dijumlahkan menjadi 33.26 untuk hasil dari rekonstruksi data atribut pertama. Begitu juga dengan hasil rekonstruksi pada atribut lainnya.

Setelah itu hasil dari rekontruksi dataset ini akan dilakukan proses normalisasi dan standarisasi sebelum masuk ke dalam model SVM. Dalam proses SVM, kernel yang digunakan adalah kernel *Radial Basis Function* (RBF), *sigmoid* dan *polynomial* dengan *coefisien* sebesar 0.5, *cache size* bernilai 200 dan *maximum iteration* sebesar -1. Nilai pada komponen SVM yaitu *coefisien, cache size* dan *maximum iteration* ini diambil dari nilai default yang ada. Akan dilakukan 2 jenis inputan dalam penelitian ini, yaitu dataset yang sudah dinormalisasi dan dataset yang tidak dinormalisasi. Data hasil akurasi dapat dilihat pada gambar 4 untuk dataset yang dinormalisasi dan gambar 5 untuk dataset yang tidak dinormalisasi berikut ini.



Gambar 4. Hasil Akurasi Dataset dengan Normalisasi



Gambar 5. Hasil Akurasi Dataset tanpa Normalisasi

Dapat dilihat hasil akurasi setiap kernel yang ada, setiap kernel memiliki nilai akurasi tertinggi pada model PCA yang berbeda-beda. Ini disebabkan oleh pengaruh kernel yang ada pada dataset tiap model PCA.

Tabel 5. Hasil Korelasi

Kernel	RBF	Sigmoid	Polynomial
Korelasi dataset normalisasi	0.337	0.494	0.470
Korelasi datsaet tanpa normalisasi	0.1689	0	0.4003

Selain itu dapat dilihat pada tabel 5 nilai korelasi yang didapatkan pada setiap kernel hampir mendekati 0.5 dari keseluruhan kernel yang ada. Nilai dari korelasi ini positif yang berarti hubungan kedua variabel saling searah dan juga lemah yaitu tidak terlalu saling mempengaruhi satu sama lain. Jika nilai PC tinggi, nilai akurasi akan searah meningkat juga namun pada beberapa titik nilai ini bisa tidak searah karena hubungannya yang lemah untuk mempengaruhi satu sama lain.

Selain itu dapat dilihat hasil perhitungan sensitivitas pada tabel 6 untuk dataset yang sudah dinormalisasi.

Tabel 6. Hasil Sensitivitas

Nilai PC	Sensitivitas RBF	Sensitivitas *Sigmoid*	Sensitivitas *Polynomial*
1	1	1	1
2	1	0.9763	1
3	1	0.9763	1
4	1	0.9842	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	0.9921	0.9921	0.9921
10	0.9921	0.9921	0.9921
11	0.9921	0.9921	0.9842
12	0.9921	0.9921	0.9842
13	0.9921	0.9921	0.9921
14	0.9921	0.9921	0.9921
15	0.9921	0.9921	0.9921
16	0.9921	0.9921	0.9921
17	0.9921	0.9921	0.9921
18	0.9921	0.9921	0.9921
19	0.9921	0.9921	0.9921
20	0.9921	0.9921	0.9921
21	0.9921	0.9921	0.9921
22	0.9921	0.9921	0.9921
23	0.9921	0.9921	0.9921
24	0.9921	0.9921	0.9921
25	0.9921	0.9921	0.9921
26	0.9921	0.9921	0.9921
27	0.9921	0.9921	0.9921
28	0.9921	0.9921	0.9921
29	0.9921	0.9921	0.9921
30	0.9921	0.9921	0.9921
31	0.9921	0.9921	0.9921
32	0.9921	0.9921	0.9921
33	0.9921	0.9921	0.9921

Pada hasil perhitungan sensitivitas untuk semua kernel hasil yang didapatkan sangat baik pada beberapa nilai PC yaitu mencapai nilai 1. Dimana nilai ini berarti sistem berhasil secara sempurna memprediksi hasil data yang terkena kanker benar terkena kanker. Perhitungan spesifisitas dapat dilihat pada tabel 7.

Tabel 7. Hasil Spesifisitas

Nilai PC	Spesifisitas RBF	Spesifisitas *Sigmoid*	Spesifisitas *Polynomial*
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0.1428	0.8571	0.7142
10	0.1428	0.8571	0.6
11	0.2857	0.8571	0.6
12	0.4285	0.8571	0.6
13	0.4285	1	0.6
14	0.4285	1	0.8
15	0.4285	0.8571	0.8
16	0.4285	1	0.8
17	0.4285	0.8571	0.8
18	0.5714	0.8571	0.8571
19	0.5714	0.8571	0.8571
20	0.5714	0.8571	0.8571
21	0.5714	0.8571	1
22	0.5714	0.8571	0.8571
23	0.5714	0.8571	0.8571
24	0.5714	1	0.8571
25	0.5714	1	0.8571
26	0.5714	0.8571	0.8571
27	0.5714	0.8571	0.8571
28	0.5714	1	0.8571
29	0.5714	0.8571	0.8571
30	0.7142	0.8571	0.8571
31	0.7142	0.8571	0.8571
32	0.7142	0.8571	0.8571
33	0.7142	0.8571	0.8571

Sementara untuk perhitungan spesifisitas dapat dilihat bahwa nilai spesifisitas yang dihasilkan bervariasi mulai dari 0 hingga 1. Pada perhitungan ini sistem masih ada yang tidak bisa memprediksi data yang tidak terkena kanker dengan baik.

Selain sensitivitas dan spesifisitas, nilai standar deviasi serta rata-rata tiap akurasi juga diperlukan pada tiap kernel. Dari hasil nilai standar deviasi ini bisa dilihat bahwa hasil akurasi tiap kernel ada seberapa banyak nilai keragaman varian atribut pada data. Hasil nilai ini dapat dilihat pada tabel 8 berikut ini.

Tabel 8. Hasil Standar deviasi dan Rata-rata Akurasi

No	Hasil	Kernel RBF	Kernel *Sigmoid*	Kernel *Polynomial*
1	Standar Deviasi	0.010	0.021	0.016
	Rata-rata Akurasi	96.24%	97.55%	97.09%
2	Standar Deviasi	0.0067	4.6×10^{-16}	0.0172
	Rata-rata Akurasi	95.8%	94.77%	96.62%

Dapat dilihat hasil nilai dari standar deviasi pada dataset yang sudah dinormalisasi (pada nomor 1) maupun yang tidak dinormalisasi (pada nomor 2) tiap kernel tidaklah tinggi. Setiap kernel tidak ada yang mencapai 0.1 pada standar deviasi yang ditemukan. Karena nilai standar deviasi ini sangat rendah, rentang penyebaran data tidaklah besar dan akurasi yang di dapatkan pun bisa optimal. Pada rata-rata akurasi yang didapatkan pun lebih tinggi pada dataset yang dinormalisasi walaupun perbedaan dengan dataset yang tidak dinormalisasi tidak begitu besar.

Tabel 9. Hasil Akurasi Dataset Normalisasi

Hasil	Kernel RBF	Kernel *Sigmoid*	Kernel *Polynomial*
SVM + PCA	97.76%	99.25%	99.25%
SVM	97.76%	98.5%	98.5%

Pada tabel 9 dapat dilihat bahwa perbandingan nilai akurasi dengan PCA tertinggi dan SVM tanpa PCA pada tiap kernel berbeda. Pada dataset yang dinormalisasi didapatkan pada kernel *polynomial* dan *sigmoid* didapatkan nilai akurasi yang sama antara akurasi menggunakan PCA yaitu 99.25% dan tanpa PCA sebesar 98.5%. Sementara pada kernel RBF didapatkan nilai akurasi dengan PCA sama dengan nilai akurasi tanpa PCA yaitu 97.76%. Hasil akurasi dengan dataset yang tidak dinormalisasi dapat dilihat pada tabel 10.

Tabel 10. Hasil Akurasi Dataset tanpa Normalisasi

Hasil	Kernel RBF	Kernel *Sigmoid*	Kernel *Polynomial*
SVM + PCA	96.26%	94.77%	98.5%
SVM	96.26%	94.77%	97.76%

Pada kernel RBF dan sigmoid didapatkan nilai akurasi yang didapatkan sama antara akurasi menggunakan PCA dan tanpa menggunakan PCA. Sementara pada kernel polynomial didapatkan nilai akurasi dengan PCA lebih baik dari pada akurasi tanpa PCA.

Perbedaan akurasi yang didapatkan pada 2 jenis inputan ini dapat dilihat bahwa dataset yang dinormalisasi memiliki akurasi yang lebih tinggi walaupun perbedaannya tidak lah terlalu jauh dengan dataset yang tidak dinormalisasi.

Pada dataset yang dinormalisasi dapat dilihat jika menggunakan kernel *polynomial* model PCA yang paling tinggi ada pada jumlah PC ke 21. Begitu juga dengan kernel RBF dimana mdel PCA tertinggi ada pada model jumlah PC ke 30 hingga ke 33 dan pada kernel sigmoid ada pada jumlah PC 13, 14 dan beberapa PC lainnya.

Kesimpulan

Penggunaan PCA pada proses ini berhasil mendapatkan nilai kontribusi pada setiap komponen yang ada untuk menghasilkan nilai akurasi SVM. Komponen pertama yaitu usia mendapatkan nilai kontribusi paling besar diantara komponen lainnya yaitu 62.99% dari semua informasi atribut yang ada. Selain itu, PCA ini juga sudah berhasil mereduksi nilai atribut pada dataset sesuai dengan nilai kontribusi yang ada.

Hasil akurasi data yang sudah dinormalisasi ini juga lebih baik dari pada hasil akurasi dataset yang tidak di normalisasi walaupun perbedaan akurasi yang didapatkan tidak terlalu besar. Hasil akurasi dengan dataset yang dinormalisasi pada kernel RBF didapatkan sebesar 97.76% pada model jumlah PC 30 hingga 33, untuk kernel *polynomial* didapatkan akurasi sebesar 99.25% pada model jumlah PC 21, dan terakhir pada kernel *sigmoid* didapatkan akurasi sebesar 99.25% juga pada beberapa PC yang ada. Dapat dilihat dengan menggunakan PCA disini hasil akurasi tertinggi pada kernel RBF dapat dicari hanya dengan mereduksi atribut menggunakan jumlah 30 hingga 33 PC. Begitu juga dengan kernel *polynomial* yang hanya perlu mereduksi data menggunakan PCA sebanyak 21 PC saja. Dapat dilihat bahwa penggunaan PCA dapat memberikan hasil akurasi yang lebih baik dalam metode SVM.

Daftar Pustaka

Adwijaya. (2018). Deteksi Kanker Berdasarkan Klasifikasi Microarray Data Menggunakan Back Propagation Neural Network dan PCA.

Aliady, H. (2018). Implementasi Support Vector Machine dan Random Forest pada Diagnosis Kanker Payudara.

He, Q., N.Li, Luo, W.-J., & Shi, Z. (2014). A survey of machine learning algorithms for big data.

Joachims, T. (1998). TextCotegorization with Support Vector Machines : Learning with Many Relevant Features.

Juniman, P. T. (2018). WHO: Kanker Membunuh Hampir 10 Juta Orang di Dunia Tahun Ini. CCN Indonesia .

Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D. P., Saikhom, R., et al. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA)

Nurrana, L., & Purwoto, G. (2000). Motode Skrining Alternatif Pada Kanker Serviks.

Roger S. Pressman, P. (2012). Dalam Rekayasa Perangkat Lunak pendekatan praktisi. Penerbit Andi.

Santosa, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu.

Setiaputri, K. A. (2020, January 8). Penyebab dan Berbagai Faktor Risiko Kanker Serviks Pada Wanita.