ANALISIS PENGARUH METODE REDUKSI DIMENSI *INDEPENDENT PRINCIPAL COMPONENT ANALYSIS* (IPCA) PADA KLASIFIKASI DATA EKSPRESI GEN MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM) DAN *NAIVE BAYES CLASSIFIER* (NBC)

(Studi Kasus: Data Ekspresi Gen Kanker Payudara (GSE43837), Kanker Pankreas (GSE16515) dan *Acute Myocardial Infarction* (GSE48060))

TUGAS AKHIR



Asmawi Gunawan

14 611 129

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

ANALISIS PENGARUH METODE REDUKSI DIMENSI *INDEPENDENT PRINCIPAL COMPONENT ANALYSIS* (IPCA) PADA KLASIFIKASI DATA EKSPRESI GEN MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM) DAN *NAIVE BAYES CLASSIFIER* (NBC)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Jurusan Statistika

Asmawi Gunawan

14 611 129

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Analisis Pengaruh Metode Reduksi Dimensi *Independent Principal Component Analysis* (IPCA) pada Klasifikasi Data Ekspresi Gen Menggunakan *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC) (Studi Kasus: Data Ekspresi Gen Kanker Payudara (GSE43837), Kanker Pankreas (GSE16515) dan *Acute Myocardial Infarction* (GSE48060))
Nama Mahasiswa	:	Asmawi Gunawan
Nomor Mahasiswa	:	14 611 129

TUGAS AKHIR INI DIPERIKSA DAN DISETUJUI UNTUK DIAJUKAN

Yogyakarta, 20 November 2020

Pembimbing,

Dr.techn. Rohmatul Fajriyah, S.Si., M.Si.

HALAMAN PENGESAHAN

TUGAS AKHIR

ANALISIS PENGARUH METODE REDUKSI DIMENSI *INDEPENDENT PRINCIPAL COMPONENT ANALYSIS* (IPCA) PADA KLASIFIKASI DATA EKSPRESI GEN MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM) DAN *NAIVE BAYES CLASSIFIER* (NBC)

(Studi Kasus: Data Ekspresi Gen Kanker Payudara (GSE43837), Kanker Pankreas (GSE16515) dan *Acute Myocardial Infarction* (GSE48060))

Nama Mahasiswa : Asmawi Gunawan

Nomor Mahasiswa : 14 611 129

TUGAS AKHIR INI TELAH DIAJUKAN

PADA TANGGAL: 1 Desember 2020

Nama Penguji	**Tanda Tangan**
1. Saepudin, M.Si., Ph.D., Apt.	: .....
2. Rahmadi Yotenka, S.Si., M.Sc.	: .....
3. Dr.techn. Rohmatul Fajriyah, S.Si., M.Si.	: .....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

Prof. Riyanto, S.Pd., M.Si., Ph.D.

KATA PENGANTAR

Assalamu'alaikum wr. wb.

Alhamdulillahirabbil 'alamin, segala puji bagi Allah SWT. atas segala nikmat, berkat, rahmat, hidayah dan karunia-Nya sehingga penulis dapat menyusun dan menyelesaikan Tugas Akhir yang berjudul "Analisis Pengaruh Metode Reduksi Dimensi *Independent Principal Component Analysis* (IPCA) pada Klasifikasi Data Ekspresi Gen Menggunakan *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC)" dengan Studi Kasus: Data Ekspresi Gen Kanker Payudara (GSE43837), Kanker Pankreas (GSE16515) dan *Acute Myocardial Infarction* (GSE48060). Selawat serta salam semoga selalu tercurah kepada junjungan Nabi Agung, Nabi Muhammad SAW. beserta keluarga, sahabat dan pengikutnya sampai akhir zaman.

Tugas Akhir ini disusun dengan sebaik-baiknya sebagai salah satu persyaratan dalam menyelesaikan jenjang strata satu di Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.

Penulis memahami tanpa bantuan, dukungan, doa, dan bimbingan dari berbagai pihak akan sangat sulit untuk menyelesaikan Tugas Akhir ini dengan baik. Maka dari itu penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Orang tua tercinta, Bapak Wardihan dan Ibu Tahira yang selalu memberikan doa restu dan kasih sayang serta adik-adik terkasih, Arham Efendi, Asrina Rani, Asni Rahmadani dan Akbar Faizun yang selalu menjadi penyemangat terbesar bagi penulis;
2. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia berserta jajarannya;
3. Bapak Dr. Edy Widodo, S.Si., M.Si. selaku ketua Program Studi Statistika beserta jajarannya yang telah banyak memberikan doa, dukungan, dan masukan;

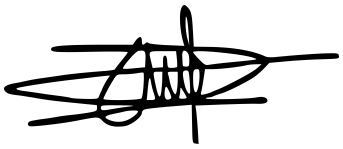
4. Ibu Kariyam, S.Si., M.Si. selaku Dosen Pembimbing Akademik yang sangat membantu mengarahkan selama masa perkuliahan hingga mencapai tahap ini;
5. Ibu Dr.techn. Rohmatul Fajriyah, S.Si., M.Si. selaku Dosen Pembimbing I atas waktu, tenaga, ilmu, kritikan, motivasi dan kesabaran dalam membimbing sehingga Tugas Akhir ini dapat terselesaikan dengan baik;
6. Bapak Ibu Dosen Program Studi Statistika Universitas Islam Indonesia atas ilmu yang telah diberikan, semoga berkah dan bernilai amal jariah di sisi Allah SWT.;
7. Teman-teman seperjuangan kelompok bimbingan Kerja Praktik dan Tugas Akhir Ibu Ema yang telah bekerja sama dengan baik dalam menyelesaikan penulisan Tugas Akhir;
8. Seluruh teman-teman Program Studi Statistika pada umumnya dan angkatan 2014 pada khususnya yang telah banyak berbagi informasi terkait akademik serta banyak bertukar ilmu selama menempuh pendidikan hingga penulisan Tugas Akhir ini selesai;
9. Teman-teman Forum Komunikasi Mahasiswa Bone Yogyakarta beserta warga Asrama Arung Palakka atas segala pengertian dan dukungannya yang selalu saling memotivasi dan berjuang bersama di daerah rantau;
10. Teman-teman Ikatan Keluarga Alumni Pondok Pesantren Al-Ikhlas Ujung yang selalu memberikan doa, semangat dan dukungan kepada penulis selama ini;
11. Semua pihak yang tidak dapat disebutkan satu persatu atas segala doa dan dukungannya baik dalam bentuk moral maupun materi, baik secara motivasi maupun aksi, semoga Tuhan membalas seluruh kebaikannya.

Meskipun tugas akhir ini telah disusun dengan sebaik-baiknya, namun penulis menyadari bahwa tugas akhir ini masih banyak membutuhkan bantuan koreksi dari berbagai pihak, baik berupa saran, kritik, pembaharuan, dan penyempurnaan. Penulis mengucapkan terima kasih dan mohon maaf atas segala kekurangan yang ada, karena kesempurnaan hanyalah milik Allah SWT. Semoga tugas akhir ini dapat bermanfaat tidak untuk penulis saja, namun juga dapat bermanfaat bagi publik dan akademisi sebagai bahan penelitian yang

dapat diaplikasikan dimasa mendatang yang bermanfaat bagi banyak orang. Akhir kata, *wabillahi taufik walhidayah*

Wassalamu'alaikum wr. wb.

Yogyakarta, 15 November 2020



Asmawi Gunawan

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN PERSETUJUAN PEMBIMBING ii
HALAMAN PENGESAHAN iii
KATA PENGANTAR iv
DAFTAR ISI vii
DAFTAR TABEL ix
DAFTAR GAMBAR x
DAFTAR LAMPIRAN xi
PERNYATAAN xii
INTISARI xiii
ABSTRACT xiv
BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 3
1.3. Batasan Masalah 3
1.4. Tujuan Penelitian 4
1.5. Manfaat Penelitian 4
BAB II TINJAUAN PUSTAKA 6
2.1. Penelitian terkait IPCA 6
2.2. Penelitian terkait Reduksi Dimensi dan Klasifikasi 7
BAB III LANDASAN TEORI 13
3.1. Bio-informatika 13
3.2. *Microarray* 13
3.3. Ekspresi Gen 14
3.4. Kanker Payudara 15
3.5. Kanker Pankreas 16
3.6. *Acute Myocardial Infarction* 17
3.7. Statistika Deskriptif 17
3.8. *Machine Learning* 18
3.9. *Pre-Processing* 19
3.10. Reduksi Dimensi 20
3.11. *Independent Principal Component Analysis* 20

3.11.1. *Principal Component Analysis* 20
3.11.2. *Independent Component Analysis* 22
3.11.3. *Independent Principal Component Analysis* 24
3.12. Klasifikasi 26
3.13. *Support Vector Machine* 26
3.13.1. *Soft Margin* 28
3.13.2. *Kernel* SVM 29
3.14. *Naive Bayes Classifier* 30
3.15. Evaluasi Kinerja Klasifikasi 32
3.15.1. *Confusion Matrix* 32
3.15.2. Kurva *Receiver Operating Characteristic* 35
BAB IV METODOLOGI PENELITIAN 36
4.1. Tempat dan Waktu Penelitian 36
4.2. Data dan Sumber Data 36
4.3. Variabel Penelitian 37
4.4. Metode Analisis Data 37
4.5. Langkah Penelitian 38
BAB V HASIL DAN PEMBAHASAN 41
5.1. Analisis Deskriptif 41
5.1.1. Kanker Payudara 41
5.1.2. Kanker Pankreas 43
5.1.3. Penyakit AMI 44
5.2. *Pre-Processing* 45
5.3. Reduksi Dimensi IPCA 46
5.4. Pembagian Data *Training* dan *Testing* 48
5.5. Klasifikasi 48
5.5.1. Hasil Klasifikasi *Support Vector Machine* 49
5.5.2. Hasil Klasifikasi *Naive Bayes Classifier* 50
5.6. Hasil Evaluasi Klasifikasi 51
BAB VI PENUTUP 54
6.1. Kesimpulan 54
6.2. Saran 55
DAFTAR PUSTAKA 56
LAMPIRAN 65

DAFTAR TABEL

Tabel 2.1 Perbandingan penelitian sebelumnya dengan penelitian saat ini ... 10
Tabel 3.1 *Confusion matrix* untuk dua kelas klasifikasi 32
Tabel 3.2 Nilai AUC 34
Tabel 3.3 *Confusion matrix* untuk klasifikasi *multi class* 34
Tabel 4.1 Informasi data..... 36
Tabel 4.2 Variabel dan definisi operasional variabel 37
Tabel 5.1 Variabel *pheno* data..... 41
Tabel 5.2 Nilai *explained variance* tiap fitur IPC 47
Tabel 5.3 Pembagian data *training* dan data *testing* 48
Tabel 5.4 Perbandingan akurasi tiap *kernel* SVM..... 49
Tabel 5.5 Nilai *Prior Probabilities* 50
Tabel 5.6 Nilai akurasi klasifikasi metode NBC..... 50
Tabel 5.7 Hasil akurasi IPC10..... 51
Tabel 5.8 *Confusion matrix* SVM *kernel* RBF dan NBC terhadap IPC10 kanker pankreas..... 52
Tabel 5.9 Hasil perhitungan ukuran kinerja klasifikasi..... 52

DAFTAR GAMBAR

Gambar 3.1 Ilustrasi fungsi $\Phi(X)$ 29
Gambar 4.1 *Flow chart* penelitian..... 38
Gambar 5.1 Histogram *diseases status*..... 42
Gambar 5.2 Histogram variabel *pheno* umur pasien..... 42
Gambar 5.3 Histogram kategori jaringan kanker pankreas..... 43
Gambar 5.4 Kategori jaringan kanker berdasarkan jenis kelamin 44
Gambar 5.5 Histogram usia pasien kanker pankreas 44
Gambar 5.6 *Bar Plot Status Disease*..... 45
Gambar 5.7 *Boxplot* data sebelum *pre-processing*..... 46
Gambar 5.8 *Boxplot* data setelah *pre-processing*..... 46
Gambar 5.9 ROC *Curve* Klasifikasi SVM *kernel* RBF (a) dan NBC (b) berdasarkan data IPC10 kanker pankreas 53

DAFTAR LAMPIRAN

Lampiran 1 Script pemrograman software R 65
Lampiran 2 Nilai explained variance IPCA berdasarkan jumlah komponen IPC 73
Lampiran 3 Output hasil reduksi dimensi IPCA data kanker payudara dengan 10 komponen IPC 74
Lampiran 4 Output hasil reduksi dimensi IPCA data kanker pankreas dengan 10 komponen IPC 75
Lampiran 5 Output hasil reduksi dimensi IPCA data penyakit AMI dengan 10 komponen IPC 76
Lampiran 6 Output hasil confusion matrix klasifikasi data kanker payudara (IPC10)..... 77
Lampiran 7 Output hasil confusion matrix klasifikasi data kanker pankreas (IPC10)..... 77
Lampiran 8 Output hasil confusion matrix klasifikasi data penyakit AMI (IPC10)..... 78

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah atau diterbitkan oleh orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 15 November 2020



Asmawi Gunawan

ANALISIS PENGARUH METODE REDUKSI DIMENSI *INDEPENDENT PRINCIPAL COMPONENT ANALYSIS* (IPCA) PADA KLASIFIKASI DATA EKSPRESI GEN MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM) DAN *NAIVE BAYES CLASSIFIER* (NBC)

(Studi Kasus: Data Ekspresi Gen Kanker Payudara (GSE43837), Kanker Pankreas (GSE16515) dan *Acute Myocardial Infarction* (GSE48060))

Oleh: Asmawi Gunawan

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Teknologi *microarray* memungkinkan untuk mengamati beribu-ribu ekspresi gen dalam waktu bersamaan. Namun data *microarray* memiliki dimensi yang sangat besar yang akan berpengaruh terhadap performansi klasifikasi. Untuk menyelesaikan masalah tersebut, maka digunakan *Independent Principal Component Analysis* (IPCA) sebagai metode reduksi dimensi dan *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC) sebagai metode klasifikasi. Data yang digunakan adalah data ekspresi gen dari kanker payudara, kanker pankreas dan *Acute Myocardial Infarction* (AMI). Tujuan yang ingin dicapai adalah untuk mengetahui pengaruh metode reduksi dimensi IPCA terhadap klasifikasi data ekspresi gen menggunakan SVM dan NBC. Berdasarkan hasil reduksi dimensi terhadap data ekspresi gen, dapat disimpulkan bahwa 10 IPC pertama dapat menjelaskan lebih dari 60% dari total variansi yang dapat mewakili sebagian besar set data asal. Hasil klasifikasi SVM dan NBC berdasarkan jumlah fitur IPC dari ketiga set data memperoleh rata-rata nilai akurasi sebesar 87,78% dan nilai AUC rata-rata sebesar 84,62% yang menunjukkan bahwa model klasifikasi yang diperoleh merupakan *classifier* yang baik (*good claissifier*).

Kata Kunci: *Microarray*, Kanker, Reduksi Dimensi, IPCA, Klasifikasi, SVM, NBC

ANALYSIS OF THE INFLUENCE OF INDEPENDENT PRINCIPAL COMPONENT ANALYSIS (IPCA) AS DIMENSIONALITY REDUCTION METHOD ON GENE EXPRESSION DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE (SVM) AND NAIVE BAYES CLASSIFIER (NBC)

(*Case Study: Gene Expression Data of Breast Cancer (GSE43837), Pancreatic Cancer (GSE16515) and Acute Myocardial Infarction (GSE48060)*)

By: Asmawi Gunawan

Departement of Statistics, Faculty of Mathematics and Natural Sciences

Islamic University of Indonesia

ABSTRACT

Microarray technology makes it possible to observe thousands of gene expressions at the same time. However, microarray data has very large dimensions which will affect classification performance. To solve this problem, Independent Principal Component Analysis (IPCA) is used as a dimensional reduction method and Support Vector Machine (SVM) and Naive Bayes Classifier (NBC) as classifier method. The data used in this study were data on gene expression from breast cancer, pancreatic cancer and Acute Myocardial Infarction (AMI). The purpose of this study was to determine the effect of the IPCA dimension reduction method on the classification of gene expression data using SVM and NBC. Based on the result of dimensional reduction of gene expression data, it can be concluded that the first 10 IPCs can explain more than 60% of the total variance which can represent most of the original datasets. The SVM and NBC classification results based on the number of IPC features from the three data sets obtained an average accuracy value of 87.78% and an average AUC value of 84.62% which indicates that the classification model obtained is a good classifier.

Key Words: *Microarray, Cancer, Dimension Reduction, IPCA, Classification, SVM, NBC*

BAB I
PENDAHULUAN

1.1. Latar Belakang

World Health Organization (WHO) menjelaskan bahwa kanker adalah sekelompok besar penyakit yang dapat bermula pada hampir semua organ atau jaringan tubuh ketika sel abnormal tumbuh tak terkendali atau melampaui batas normal dan menyerang bagian tubuh yang berdekatan lalu kemudian menyebar ke organ lain (WHO, 2018). Proses tersebut disebut metastasis dan merupakan penyebab utama kematian akibat kanker. Kanker adalah penyebab kematian kedua secara global, terhitung sekitar 9,6 juta kematian, atau satu dari enam kematian, pada tahun 2018 (Kemenkes, 2019). Kanker paru-paru, prostat kolorektal, perut dan hati adalah jenis kanker yang paling umum pada pria, sedangkan kanker payudara, kanker kolorektal, paru-paru, serviks dan tiroid adalah yang paling umum di antara wanita. Beban kanker terus meningkat secara global, memberikan tekanan fisik, emosional dan finansial yang luar biasa pada individu, keluarga, komunitas dan sistem kesehatan (WHO, 2018).

Penelitian yang berkaitan dengan kanker menjadi topik utama dalam bidang kesehatan dan bio-informatika dan terus berkembang hingga saat ini. Hal ini juga didukung dengan adanya pengembangan teknologi yang bernama *microarray*. Analisis *microarray* berperan penting dalam mendiagnosis suatu penyakit karena dapat digunakan untuk melihat sampai ke tingkat ekspresi gen pada sampel sel tertentu dan dapat memeriksa ribuan gen secara bersamaan (Nurfalah, Adiwijaya, & Suryani, 2016). *Microarray* digunakan pada sejumlah besar aplikasi untuk menyelidiki ratusan sampel hingga jutaan molekul yang berbeda, dan juga telah menjadi salah satu teknik yang paling cepat berkembang sejak diperkenalkan. *Microarray* membantu dalam menganalisis sejumlah besar sampel yang telah direkam sebelumnya atau sampel yang terbaru, bahkan membantu untuk menguji adanya penanda penyakit tertentu yang ada pada suatu tumor (Govindarajan, Duraiyan, Kaliyappan, & Palanisamy, 2012).

Teknologi *microarray* dapat digunakan pada genotipe skala besar, profil ekspresi gen, hibridisasi genomik dan lain sebagainya (Raza, 2010). Teknologi ini

merupakan kombinasi dari berbagai bidang disiplin ilmu pengetahuan dan teknologi yang dikembangkan dari berbagai penelitian seperti mekanik, pembuatan mikro, kimia, perilaku DNA, mikrofluida, enzim, optik dan bio-informatika (Dufva, 2009). *Microarray* kerap digunakan untuk mempelajari ekspresi dari banyak gen yang dikhususkan untuk menangani masalah kanker, walaupun pada perkembangannya juga diterapkan pada berbagai macam penyakit lainnya (Sanz & Sánchez, 2018). Teknologi ini banyak digunakan dalam bidang kedokteran untuk proses pendekatan statistik terhadap kanker yang cepat dan lebih efisien.

Seiring dengan perkembangan teknologi yang sangat pesat, seperti halnya teknologi *microarray* dan *Next-Generation Sequencing* (NGS), eksplorasi terhadap data menjadi kebutuhan utama demi mengungkap informasi yang relevan yang terkandung di dalam data. Banyak pendekatan statistik yang gagal atau berkinerja buruk saat diterapkan pada data *microarray* karena dua alasan utama yaitu jumlah sampel pengamatan atau observasi jauh lebih kecil daripada jumlah variabel atau entitas biologis yang diukur serta data yang terkandung sangat bising (Yao, Coquery, & Cao, 2012). Oleh karena itu diperlukan alat eksplorasi yang efisien yang tidak hanya untuk menilai kualitas data, tetapi juga untuk memberikan gambaran komprehensif dari data, mengekstrak informasi yang signifikan dan mengatasi dimensi yang tinggi.

Teknik *microarray* mampu menghasilkan data yang dibutuhkan untuk proses prediksi dan klasifikasi gen yang diambil dari beberapa jaringan tertentu pada manusia untuk digolongkan ke dalam jenis kanker atau bukan. Namun terdapat kendala dalam penggunaan *microarray*, yaitu besarnya dimensi yang dimiliki sehingga perlu dilakukan proses reduksi dimensi untuk menghindari masalah *overfitting* pada klasifikasi gen (Yip, Amin, & Li, 2011). Dimensi yang dimaksud adalah ukuran data berupa atribut dan sampel. Besarnya dimensi inilah yang menjadi masalah dalam suatu data *microarray* karena informasi penting dan berguna sering kali terhalang pada data tersebut, yang mengakibatkan beban komputasi yang tidak stabil dan tingkat performansi yang rendah. Untuk mengatasi masalah tersebut, dalam penelitian ini, peneliti menggunakan metode *Independent Principal Component Analysis* (IPCA) sebagai alat untuk mereduksi dimensi data ekspresi gen dari data *microarray*. Dalam hal ini, set data yang berdimensi tinggi dipartisi ke dalam suatu himpunan bagian (*subset*).

Metode IPCA menggunakan *Principal Component Analysis* (PCA) sebagai langkah *pre-processing* untuk mengurangi dimensi data dan menghasilkan vektor *loading*. Kemudian menerapkan algoritma *FastICA* pada vektor *loading* tersebut yang kemudian akan menghasilkan *subset Independent Principal Component* (IPC) (Hyvärinen & Oja, 2000). Vektor *loading* juga akan digunakan untuk menyusun IPC. Variabel-variabel baru yang terbentuk berdasarkan IPC pertama dari *subset* akan digunakan untuk melakukan proses klasifikasi.

Proses klasifikasi yang bertujuan untuk mengklasifikasikan data kanker atau bukan kanker dilakukan dengan menggunakan metode klasifikasi *Support Vector Machine* (SVM) yang telah dibuktikan pada buku yang ditulis oleh Lu, Schölkopf dan Zhao (2011) yang mengatakan bahwa "*Support Vector Machine* (SVM) dapat mengatasi data yang memiliki dimensi besar dan berhasil diterapkan pada penelitian *microarray*". Kemudian dalam penelitian ini juga diterapkan metode *Naive Bayes Classifier* (NBC) sebagai bahan perbandingan untuk hasil akurasi klasifikasi. Dengan demikian akan diketahui apakah reduksi dimensi *Independent Principal Component Analysis* (IPCA) berpengaruh pada hasil klasifikasi metode SVM dan NBC.

1.2. Rumusan Masalah

Berdasarkan pada latar belakang yang telah dipaparkan, maka penulis dapat merumuskan masalah yang akan diselesaikan dalam penelitian ini, yaitu:

1. Bagaimana gambaran dari data pasien kanker payudara, pasien kanker pankreas dan pasien *Acute Myocardial Infarction* (AMI)?
2. Bagaimana implementasi hasil dari metode IPCA dalam proses reduksi dimensi data ekspresi gen?
3. Bagaimana hasil klasifikasi dari metode SVM dan NBC berdasarkan hasil reduksi dimensi IPCA?
4. Bagaimana perbandingan hasil implementasi dari metode klasifikasi SVM dan NBC?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan pada penelitian ini adalah data ekspresi gen kanker payudara, data ekspresi gen kanker pankreas dan data ekspresi gen pasien *Acute Myocardial Infarction* (AMI).
2. Metode yang digunakan untuk melakukan proses reduksi dimensi adalah metode *Independent Principal Component Analysis* (IPCA).
3. Metode yang digunakan untuk melakukan klasifikasi adalah metode *Support Vector Machine* (SVM) dan metode *Naive Bayes Classifier* (NBC).
4. Pengolahan data menggunakan *software Rstudio versi* 3.6.1.

1.4. Tujuan Penelitian

Tujuan yang ingin dicapai sebagai jawaban dari rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Mengetahui gambaran dari data pasien kanker payudara, pasien kanker pankreas dan pasien AMI.
2. Mengetahui hasil implementasi dari metode IPCA pada proses reduksi dimensi data ekspresi gen.
3. Mengetahui hasil klasifikasi dari metode SVM dan NBC berdasarkan hasil reduksi dimensi IPCA.
4. Mengetahui perbandingan hasil implementasi dari metode klasifikasi SVM dan NBC.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat di antaranya adalah sebagai berikut:

1. Bagi Penulis

   Dapat menerapkan ilmu yang diperoleh dari bangku perkuliahan dan mengembangkan pengetahuan baru dalam bidang bio-informatika.
2. Bagi Pembaca
   - Memberikan informasi tentang pengklasifikasian data ekspresi gen pasien kanker.
   - Memberikan penjelasan tentang penggunaan metode IPCA dalam melakukan proses reduksi dimensi terhadap data ekspresi gen.
   - Memberikan penjelasan tentang penggunaan metode SVM dan NBC dalam klasifikasi data ekspresi gen kanker.

3. Bagi Penelitian Selanjutnya

   Hasil dari penelitian ini dapat dijadikan acuan dalam penelitian terkait atau dapat dikembangkan serta dianalisis lebih mendalam dengan menambahkan atau membandingkan beberapa metode reduksi dimensi dan metode klasifikasi lainnya yang akan mendapatkan hasil kinerja yang lebih baik.

BAB II
TINJAUAN PUSTAKA

Dalam suatu penelitian diperlukan dukungan dari hasil penelitian yang telah ada sebelumnya yang berkaitan dengan penelitian yang dilakukan saat ini. Penelitian terdahulu sebagai kajian bagi penulis bertujuan untuk mengetahui hubungan antara penelitian yang dilakukan sebelumnya dengan penelitian yang dilakukan saat ini serta dapat menghindari adanya duplikasi. Hal ini bermanfaat untuk menunjukkan bahwa penelitian yang dilakukan mempunyai arti penting sehingga dapat diketahui kontribusinya terhadap ilmu pengetahuan.

2.1. Penelitian terkait IPCA

Yao et al. (2012) dalam penelitiannya yang berjudul "*Independent Principal Component Analysis for Biologically Meaningful Dimension Reduction of Large Biological Data Sets*" telah mengembangkan varian PCA yang disebut *Independent Principal Component Analysis* (IPCA) yang menggabungkan keunggulan dari *Principal Component Analysis* (PCA) dan *Independent Component Analysis* (ICA). IPCA mengasumsikan bahwa komponen yang bermakna secara biologis dapat diperoleh jika sebagian besar kebisingan telah dihapus dari vektor *loading* terkait. Dengan mengidentifikasi vektor *loading non-Gaussian* dari data biologis, maka IPCA dapat lebih baik dalam mencerminkan struktur internal data dibandingkan dengan PCA dan ICA. Pada set data simulasi, penelitian tersebut menunjukkan bahwa IPCA mengungguli PCA dan ICA dalam kasus *super-Gaussian*, dan bahwa nilai kurtosis vektor *loading* dapat digunakan untuk memilih jumlah komponen utama yang independen. Pada set data *real*, telah dihasilkan validitas *cluster* menggunakan indeks Davies Bouldin dan menunjukkan bahwa dalam kasus dimensi tinggi, IPCA dapat meringkas informasi data dengan lebih baik atau dengan jumlah komponen yang lebih kecil daripada PCA atau ICA. Dalam penelitian ini telah dibuktikan bahwa pendekatan IPCA dapat bermanfaat, tidak hanya untuk meningkatkan visualisasi data dan mengungkapkan karakteristik eksperimental, tetapi juga untuk mengidentifikasi variabel yang relevan secara biologis.

Selain itu, Boluwade dan Madramootoo (2016) juga telah melakukan penelitian yang berjudul "*Independent Principal Component Analysis for Simulation of Soil Water Content and Bulk Density in a Canadian Watershed*". Penelitian tersebut menerapkan metode IPCA yang telah dikembangkan oleh Yao et al. (2012) yang menggabungkan kekuatan PCA dan ICA untuk simulasi spasial *soil water content* (SWC) dan *bulk density* (BD) di daerah aliran sungai Kanada. Penelitian ini bertujuan untuk menemukan teknik yang efisien dan kuat untuk menstimulasikan variabel yang berkorelasi secara spasial. Hal ini penting dalam hal pemahaman dan karakterisasi variabilitas spasial terhadap data. Penelitian ini dapat berguna dalam hal keakuratan estimasi terhadap permukaan SWC yang kemudian dapat diterapkan dalam memandu keputusan pengelolaan pertanian, mengurangi beban berat pada pengambilan sampel dan pemantauan latihan serta mencegah dampak lingkungan yang negatif seperti erosi tanah, banjir, dan sebagainya.

Penelitian ini juga telah menunjukkan bahwa IPCA dapat menghilangkan korelasi antar *lag distances* yang merupakan kekurangan dari metode *principal component analysis* (PCA). Selain itu, metode ini juga tidak memerlukan transformasi data ke dalam bentuk distribusi *Gaussian*. Dengan demikian metode ini dapat mengatasi kekurangan metode *maximum autocorrelation factors* (MAF). IPCA memiliki keunggulan ini yang menunjukkan kepentingannya dalam bidang geo-statistik.

Hasil penelitian Boluwade dan Madramootoo (2016) menunjukkan bagaimana atribut tanah dapat disimulasikan secara bersama-sama tanpa biaya komputasi yang tinggi. Hasil ini penting dalam karakterisasi yang akurat dari variabilitas kadar air dan variasi densitas curah terhadap aplikasi pertanian yang lebih presisi. Atribut kelembaban tanah merupakan faktor penting dalam pertanian irigasi dalam hal penerapan air yang tepat waktu. Pengetahuan yang memadai tentang variabilitas properti tanah akan membantu dalam menentukan pembuatan, pengelompokan spasial, dan partisi berbagai zona pengelolaan untuk sistem irigasi yang optimal.

2.2. Penelitian terkait Reduksi Dimensi dan Klasifikasi

Adiwijaya (2018) melakukan penelitian yang berjudul "Deteksi Kanker Berdasarkan Klasifikasi *Microarray* Data". Penelitian yang dilakukan bertujuan

untuk mengaplikasikan suatu metode reduksi dimensi dan klasifikasi terhadap beberapa kelompok data ekspresi gen kanker seperti *Breast, Colon, Ovarian, Prostate, Central Nervous, Lung,* dan *Leukimia* untuk mendapatkan hasil dan akurasi klasifikasi yang baik. Beberapa metode reduksi dimensi yang digunakan dalam penelitian ini seperti *Minimum Redundancy Maximum Relevance* (MRMR), *Mutual Information*, PCA, *Relief Method*, dan lain sebagainya. Sedangkan metode klasifikasi yang digunakan seperti *Neural Network*, *Bayesian*, *Decision Tree* dan lain-lain. Hasil penelitian yang didapatkan bahwa beberapa metode klasifikasi teruji dapat melakukan klasifikasi ekspresi gen untuk mendeteksi kanker dengan nilai performansi maksimal 98,71%. Proses reduksi dimensi dapat dilakukan setelah melakukan *pre-processing cleansing dataset* dan *cross validation*. Tiga metode terbukti dapat digunakan untuk melakukan reduksi dimensi yaitu PCA, *Mutual Information* dan *Relief Method*. Atribut yang terbaik digunakan berbeda-beda pada setiap data. Metode klasifikasi dan metode reduksi dimensi mempunyai karakteristik masing-masing jika dikombinasikan. Hal ini ditunjukkan saat sistem dibangun dengan suatu metode klasifikasi yang sama dapat dihasilkan akurasi berbeda ketika reduksi dimensi berbeda.

Penelitian yang berjudul "Analisis Perbandingan Reduksi Dimensi *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS) untuk Deteksi Kanker menggunakan Data *Microarray*" yang pernah dilakukan oleh Sirait, Adiwijaya dan Astuti (2019). Pada penelitian ini, sistem yang dibangun menggunakan dua metode fitur ekstraksi yaitu *Principal Component Analysis* (PCA) dan *Partial Least Squar*e (PLS) dengan *Support Vector Machine* (SVM) sebagai *classifier*. Hal ini berguna untuk mengurangi atribut yang terlalu banyak. Sistem yang dibangun mampu mengklasifikasi kanker dan memperoleh nilai rata-rata 82% untuk PCA-SVM dan 55,17% untuk PLS-SVM.

Pada tahun 2018, sebuah penelitian yang berjudul "Analisis Reduksi Dimensi pada Klasifikasi *Microarray* Menggunakan MBP *Powell Beale*" dilakukan oleh Munzir, Adiwijaya dan Aditsania (2018). Penelitian ini bertujuan untuk membandingkan performansi sistem klasifikasi data *microarray* menggunakan metode MBP *Powell Beale* yang telah melalui proses reduksi dimensi, baik seleksi fitur maupun ekstraksi fitur. PCA dipilih sebagai metode ekstraksi fitur, sedangkan GA dipilih sebagai metode seleksi fitur. Hasil yang

didapatkan adalah skema klasifikasi MBP tanpa reduksi dimensi menghasilkan akurasi 70,59% − 100%, skema klasifikasi MBP+PCA menghasilkan akurasi 76,47% - 100%, dan skema klasifikasi MBP+GA menghasilkan akurasi 35,26% - 84,62%. Hal ini disebabkan karena pengaruh pengambilan vektor *eigen* (PPV) pada MBP+PCA serta pengaruh ukuran populasi dan *max gen* pada MBP+GA yang mempengaruhi klasifikasi pelatihan dan pengujian.

Ma'ruf, Adiwijaya dan Wisesty (2018) melakukan penelitian yang berjudul "Analisis Pengaruh Metode Reduksi Dimensi *Minimum Redundancy Maximum Relevance* pada Klasifikasi Kanker Berdasarkan Data *Microarray* Menggunakan *Classifier Support Vector Machine*". Penelitian ini bertujuan untuk melakukan reduksi dimensi dengan menggunakan metode *Minimum Redundancy Maximum Relevance* (MRMR) dan ekstraksi fitur menggunakan metode *Principal Component Analysis* (PCA). Kesimpulan yang dihasilkan dari penelitian ini antara lain: Pemilihan fitur oleh metode MRMR, jumlah vektor *eigen* pada proses reduksi dimensi oleh metode PCA, serta pemilihan fungsi *kernel* pada SVM, amat sangat berpengaruh terhadap *f-score* yang dihasilkan dan lamanya waktu komputasi pada proses klasifikasi data kanker pada setiap metode (SVM, MRMR + SVM, PCA + SVM). Dari hasil akhir klasifikasi (*f-score*) dengan metode reduksi dimensi MRMR + SVM, nilainya lebih bagus dari hasil akhir klasifikasi data kanker tanpa metode reduksi dimensi. Salah satu buktinya adalah pada klasifikasi kanker usus besar, di mana MRMR FCQ + SVM *kernel linear* menghasilkan *f-score* sebesar 0,81667 dengan pemilihan fitur oleh MRMR FCQ sebanyak 50% dari total 2000 fitur, di mana hasil klasifikasi tanpa metode reduksi dimensi dengan *classifier* SVM *kernel linear* yang perolehan *f-score*-nya adalah 0,72. Hal ini membuktikan bahwa fitur-fitur yang terpilih ini meskipun sedikit tapi memiliki sifat generalisasi yang lebih baik dan secara efektif mencakup karakteristik dari kelas.

Penelitian berjudul "*Principal Component Analysis* sebagai Ekstraksi Fitur Data *Microarray* untuk Deteksi Kanker Berbasis *Linear Discriminant Analysis*" dilakukan oleh Astuti dan Adiwijaya (2019) yang bertujuan untuk melakukan klasifikasi data *microarray* menggunakan metode *Linear Discriminant Analysis* (LDA) yang dikombinasikan dengan reduksi dimensi menggunakan metode *Principal Component Analysis* (PCA) untuk deteksi kanker usus besar. Adapun kesimpulan yang diperoleh adalah bahwa kombinasi metode PCA dan LDA mampu

mengklasifikasikan data *microarray*. PCA mampu mengurangi jumlah fitur yang diperlukan dari 2000 fitur menjadi 61 fitur. Kinerja terbaik dihasilkan dengan penggunaan PCA dengan jumlah fitur 8 yang mampu meningkatkan *f-score* klasifikasi data *microarray* sebesar 64,28% dan akurasi sebesar 29,04% dibandingkan saat tanpa menggunakan PCA. Lebih jauh lagi, hasil penelitian yang diperoleh mampu memperbaiki kinerja beberapa penelitian sebelumnya. Hal ini menunjukkan bahwa PCA juga mampu menghilangkan *noise* yang terdapat pada data yang digunakan.

Ma'wa, Adiwijaya dan Rohmawati (2019) melakukan penelitian yang berjudul "Klasifikasi *K-Nearest Neighbor* untuk Data *Microarray* dengan Seleksi *Genetic Algorithm*". Penelitian ini bertujuan untuk mengetahui bagaimana mengimplementasikan proses seleksi atribut dengan *Genetic Algorithm* (GA) pada data *microarray* dan bagaimana hasil akurasi klasifikasi *microarray* dengan *K-Nearest Neighbor* (KNN). Dalam hal ini, proses reduksi dimensi bertujuan untuk meringankan beban komputasi serta menghindari *overfitting* pada klasifikasi. Proses reduksi dimensi yang digunakan pada penelitian ini yaitu seleksi fitur *Genetic Algorithm* (GA). Kemudian, proses klasifikasi yang bertujuan untuk mendeteksi penyakit kanker atau bukan kanker dilakukan dengan menggunakan metode klasifikasi *K-Nearest Neighbor* (KNN). Adapun akurasi dari metode GA-KNN pada data tumor usus sebesar 93,02%, data kanker paru-paru sebesar 96,03% dan kanker darah sebesar 96%, dengan rata-rata akurasi data keseluruhan adalah 95,01%.

Tabel 2.1 Perbandingan penelitian sebelumnya dengan penelitian saat ini

Penelitian Sebelumnya			
No.	**Judul Penelitian**	**Peneliti, Tahun**	**Tujuan**
1	Deteksi Kanker Berdasarkan Klasifikasi *Microarray* Data	(Adiwijaya, 2018)	Menentukan skema yang tepat dalam melakukan reduksi dimensi dan klasifikasi pada beberapa data kanker.

No.	Judul Penelitian	Peneliti, Tahun	Tujuan
2	Analisis Perbandingan Reduksi Dimensi *Principal Component Analysis* (PCA) dan *Partial Least Square* (PLS) untuk Deteksi Kanker Menggunakan Data *Microarray*	(Sirait et al., 2019)	Penelitian ini melakukan analisis terhadap performansi klasifikasi menggunakan SVM yang telah direduksi dimensi terlebih dahulu menggunakan algoritma PCA dan PLS.
3	Analisis Reduksi Dimensi pada Klasifikasi *Microarray* Menggunakan MBP *Powell Beale*	(Munzir et al., 2018)	Membandingkan performansi sistem klasifikasi data *microarray* menggunakan metode MBP *Powell Beale* yang telah melalui proses reduksi dimensi PCA sebagai metode ekstraksi fitur dan GA sebagai metode seleksi fitur.
4	Analisis Pengaruh Metode Reduksi Dimensi *Minimum Redundancy Maximum Relevance* pada Klasifikasi Kanker Berdasarkan Data *Microarray* Menggunakan *Classifier Support Vector Machine*	(Ma'ruf et al., 2018)	Melakukan reduksi dimensi dengan menggunakan metode *Minimum Redundancy Maximum Relevance* (MRMR) dan klasifikasi menggunakan metode *Support Vector Machine* (SVM).

No.	Judul Penelitian	Peneliti, Tahun	Tujuan
5	*Principal Component Analysis* Sebagai Ekstraksi Fitur Data *Microarray* untuk Deteksi Kanker Berbasis *Linear Discriminant Analysis*	(Astuti & Adiwijaya, 2019)	Melakukan klasifikasi data *microarray* menggunakan metode *Linear Discriminant Analysis* (LDA) yang dikombinasikan dengan reduksi dimensi menggunakan metode *Principal Component Analysis* (PCA) untuk deteksi kanker usus besar.
6	Klasifikasi *K-Nearest Neighbor* untuk Data *Microarray* dengan Seleksi *Genetic Algorithm*	(Ma'wa et al., 2019)	Mengetahui bagaimana mengimplementasikan proses seleksi atribut dengan *Genetic Algorithm* (GA) pada data *microarray* dan bagaimana hasil akurasi klasifikasi *microarray* dengan *K-Nearest Neighbor* (KNN).
Penelitian Saat Ini			
No.	**Judul Penelitian**		**Tujuan**
1	Analisis Pengaruh Metode Reduksi Dimensi *Independent Principal Component Analysis* (IPCA) pada Klasifikasi Data Ekspresi Gen Menggunakan *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC)		Penelitian ini bertujuan untuk mengetahui apakah terdapat pengaruh reduksi dimensi menggunakan metode IPCA terhadap hasil klasifikasi data ekspresi gen menggunakan metode SVM dan NBC.

BAB III
LANDASAN TEORI

3.1. Bio-informatika

Menurut Fatchiyah (2008), bio-informatika dikemukakan pada pertengahan era 1980-an yang mengacu pada penerapan komputer dalam biologi. Pembuatan basis data dalam bio-informatika sendiri telah dilakukan sejak tahun 1960-an. Sejak era tersebut bio-informatika telah diaplikasikan dalam bidang klinis baik untuk mengidentifikasi agen penyakit atau diagnosa penyakit baru maupun penemuan obat-obatan untuk penyakit tersebut.

Bio-informatika merupakan penggabungan beberapa disiplin ilmu, seperti matematika, statistika, komputer, biokimia, genetika, dan biologi molekuler (Rana & Vaisla, 2012). Bio-informatika adalah cabang Bioteknologi yang menggunakan komputer untuk menganalisis dan mengelola data DNA maupun protein (Palladino & Thieman, 2013). Menurut Bayat (2002) bio-informatika adalah suatu aplikasi dan alat komputasi yang digunakan untuk menginterpretasikan data-data biologis. Terdapat 3 tujuan utama penggunaan bio-informatika, yaitu: (a) mengorganisasikan data, baik mengakses data yang sudah ada atau menambahkan data baru; (b) mengembangkan alat dan sumber daya untuk membantu menganalisis data; dan (c) menggunakan alat untuk menganalisis data dan menginterpretasikan hasil sesuai dengan istilah dalam biologi (Luscombe, Greenbaum, & Gerstein, 2001).

Salah satu analisis pada bio-informatika adalah analisis ekspresi gen, karena kebanyakan ekspresi dalam gen dapat diukur pada level mRNA dengan berbagai teknik seperti pembacaan mikro, pengurutan cDNA, dan analisis serial pengurutan ekspresi gen. Hal tersebut dapat dilakukan dengan berbagai teknik, salah satunya adalah dengan menggunakan teknik *microarray* (Raza, 2010).

3.2. *Microarray*

Sanz dan Sánchez (2018) menjelaskan bahwa *microarray*, atau secara khusus *microarray* RNA, adalah mesin yang dikembangkan pada akhir 1990-an untuk mengukur ekspresi gen. Saat ini *microarray* telah berkembang pesat menjadi teknologi dengan kapasitas yang matang dan sangat meningkat dari yang sebelumnya hanya dapat menampung beberapa ratus atau ribuan EST dalam *array*

pertama sampai dapat menampung jutaan *probe* yang meliputi *exson*, *intron*, dan bagian lain dari gen. Analisis data *microarray* telah menjadi bidang yang sangat aktif. Bidang ini telah diterapkan pada berbagai macam masalah seperti memilih gen yang diekspresikan secara berbeda, membangun prognosis atau prediksi diagnostik ataupun menemukan kelompok dalam data. Analisis biasanya berlangsung melalui serangkaian langkah: eksplorasi data, kontrol kualitas, normalisasi, analisis statistik, dan signifikansi biologis atau analisis jalur. Ada banyak alat yang tersedia untuk analisis data *microarray*, salah satunya adalah *library Bioconductor* (http://bioconductor.org) yang menggunakan bahasa Statistik R.

Microarray sendiri merupakan hasil kombinasi dari berbagai bidang disiplin ilmu pengetahuan dan teknologi yang dikembangkan dari berbagai penelitian seperti mekanik, pembuatan mikro, kimia, perilaku DNA, mikrofluida, enzim, optik dan bio-informatika (Dufva, 2009). Data *microarray* dapat digunakan dalam mendeteksi dan mengklasifikasikan suatu jaringan penyakit pada manusia. *Microarray* menghasilkan ekspresi gen yang berisi informasi gen yang kemudian dicocokkan dengan suatu penyakit (Trevino, Falciani, & Barrera-Saldaña, 2007).

3.3. Ekspresi Gen

Ekspresi gen adalah proses transkripsi DNA di dalam sel menjadi RNA (Madigan, Martinko, Dunlap, & Clark, 2008). Ekspresi gen merupakan suatu rangkaian proses penerjemahan informasi genetik, di dalam bentuk urutan basa pada DNA atau RNA, menjadi protein. Ekspresi genetik adalah suatu rangkaian proses kompleks yang melibatkan banyak faktor. Salah satu ciri penting pada jasad hidup adalah keteraturan sistem. (Lehninger, 1994).

Ekspresi gen di dalam sel memerlukan dua proses yaitu transkripsi dan translasi. Transkripsi adalah proses pembuatan RNA terutama mRNA dengan menyalin sebagian berkas DNA oleh enzim RNA *polimerse*. Proses ini mengubah basa nitrogen *timina* di DNA yang pada RNA digantikan oleh *urasil* (Miller, Griffiths, Lewontin, Suzuki, & Gelbart, 2002). Sedangkan Translasi adalah proses penerjemahan urutan nukleotida yang ada pada molekul mRNA menjadi rangkaian asam amino yang menyusun suatu polipeptida atau protein (Yuwono, 2005). Transkripsi dan Translasi merupakan dua proses utama yang menghubungkan gen

ke protein (Campbell, Reece, & Mitchell, 2002). Translasi adalah proses yang hanya terjadi pada molekul mRNA, sedangkan rRNA dan tRNA tidak melalui proses translasi (Yuwono, 2005). Molekul mRNA yang merupakan salinan urutan DNA menyusun suatu gen dalam bentuk kerangka baca terbuka dan membawa informasi urutan asam amino (Dale & Park, 2004).

3.4. Kanker Payudara

Menurut Humaera dan Mustofa (2017), Kanker payudara (*Carcinoma mammaee*) dalam bahasa Inggris disebut *breast cancer* yang merupakan kanker pada jaringan payudara. *Ca mammaee* adalah sekelompok sel yang tidak normal pada payudara yang terus tumbuh berlipat ganda. Pada akhirnya sel-sel ini menjadi bentuk benjolan di payudara. Dalam tubuh terdapat berjuta-juta sel. Salah satunya, sel *abnormal* atau sel *metaplasia*, yaitu sel yang dapat berubah-ubah tetapi masih dalam batas normal. *Ca Mammae* paling umum menyerang wanita, walaupun laki-laki mempunya potensi terkena akan tetapi kemungkinannya sangat kecil dengan perbandingan 1 : 1000. *Ca Mammae* terjadi karena kondisi sel telah kehilangan pengendalian dan mekanisme normalnya, sehingga mengalami pertumbuhan yang tidak normal, cepat dan tidak terkendali. *Ca Mammae* sering didefinisikan sebagai suatu penyakit *neoplasma* yang ganas yang berasal dari *parenchyma*.

Berdasarkan data penelitian Harianto, Mutiara, dan Surachmat (2005). di Rumah Sakit Dr. Cipto Mangunkusumo pada tahun 2005, faktor risiko kanker payudara di antaranya adalah riwayat keluarga dengan penderita kanker payudara (15,79%), *menarche* dini (8,77%), *nullipara* (7,02%) dan pemakaian pil yang mengandung estrogen jangka panjang (42,11%). Selain itu, juga terdapat faktor risiko lain yang diduga berpengaruh terhadap kejadian kanker payudara yaitu menopause terlambat, riwayat pemberian ASI, dan obesitas (Harianto et al., 2005).

Faktor Risiko yang erat kaitannya dengan peningkatan kanker payudara antara lain jenis kelamin wanita, usia >50 tahun, riwayat keluarga dan genetik (Pembawa mutasi gen BRCA1, BRCA2, ATM atau TP53 (p53)), riwayat penyakit payudara sebelumnya (DCIS pada payudara yang sama, LCIS, densitas tinggi pada mamografi), riwayat menstruasi dini (<12 tahun) atau *menarche* lambat (>55 tahun), riwayat reproduksi (tidak memiliki anak dan tidak menyusui), faktor

hormonal, obesitas, konsumsi alkohol, riwayat radiasi dinding dada, dan faktor lingkungan (Harianto et al., 2005).

Kanker payudara merupakan penyebab ke dua terbanyak tumor otak metastasis dengan risiko 10-16% pada orang hidup dan meningkat menjadi 30% setelah dilakukan autopsi (Rostami, Mittal, Rostami, Tavassoli, & Jabbari, 2016). Otak merupakan organ pertama metastasis pada 12% penderita kanker payudara. Faktor yang diduga mempengaruhi metastasis otak pada kanker payudara adalah usia muda, ukuran kanker payudara besar, *grading* sel kanker yang tinggi, kanker payudara tanpa ekspresi *reseptor estrogen*, dengan ekspresi HER2 dan metastasis *limfonodi* (Rostami et al., 2016).

Silalahi, Rahmi dan Sutarni (2018) telah melakukan penelitian yang menyimpulkan bahwa tumor otak metastasis dari kanker payudara merupakan kondisi yang secara epidemiologi cukup tinggi. Teknik pencitraan otak dan terapi yang efektif pada kanker payudara telah meningkatkan prevalensi kondisi tersebut. Pemilihan terapi yang tepat serta identifikasi histologi kanker payudara dapat berperan penting dalam meminimalkan kemungkinan metastasis otak dan memperpanjang angka harapan hidup penderita.

3.5. Kanker Pankreas

Kanker pankreas merupakan salah satu keganasan yang paling membahayakan jiwa dan menjadi penyebab kematian nomor 4 di dunia, dengan angka ketahanan hidup 5 tahun pasien kurang dari 7%. Dilaporkan sekitar 80-85% pasien datang terdiagnosis tumor yang sudah meluas atau metastasis dikarenakan keganasan ini bersifat tenang dan tanpa gejala pada saat stadium awal (Eveline, Purwanto, & Lestari, 2017).

Kanker pankreas merupakan penyebab keempat kematian akibat kanker di Amerika Serikat setelah kanker paru, kolon, dan payudara. Kurang dari 5% pasien kanker pankreas yang bisa bertahan hidup lebih dari 5 tahun (Rosai, 2011). Kasus keganasan pankreas lebih sering terjadi pada laki-laki dengan usia 60-80 tahun. Delapan puluh lima persen kasus keganasan pankreas adalah *adenokarsinoma duktal* dari eksokrin pankreas (Widityasari & Dewi, 2014). Kasus *adenokarsinoma duktal* pankreas jarang terjadi pada usia di bawah 40 tahun. Insiden *adenokarsinoma duktal* pankreas pada usia kurang dari 40 tahun sebesar 0,3%.

Pengobatan konvensional pada kanker pankreas adalah kemoterapi dan radioterapi. Terdapat beberapa faktor yang menyebabkan kanker pankreas, yaitu: merokok, peradangan pankreas, diabetes, keturunan genetik, obesitas dan umur (CMScripst, 2017).

3.6. *Acute Myocardial Infarction*

Penyakit *acute myocardial infarction* (AMI) merupakan gangguan aliran darah ke jantung yang menyebabkan sel otot jantung mati. Hal ini menyebabkan penurunan aliran darah di koroner. Pasokan oksigen yang dibutuhkan tidak memenuhi sehingga mengakibatkan iskemia jantung (Scheen, 2018). Plak *aterosklerotik* klasik pecah dan menyebabkan trombosis berkontribusi terhadap penurunan aliran darah akut di koroner (Massberg & Polzin, 2018). Aliran darah di pembuluh darah terhenti setelah terjadi sumbatan koroner akut, kecuali sejumlah kecil aliran kolateral dari pembuluh darah di sekitarnya. Daerah otot di sekitarnya yang sama sekali tidak mendapat aliran darah atau alirannya sangat sedikit sehingga tidak dapat mempertahankan fungsi otot jantung, dikatakan mengalami *infarction* (Hall & Guyton, 2006). AMI dibagi menjadi dua kategori yaitu *Non ST-Segment MI* (NSTEMI) dan *ST-Segment MI* (STEMI) (Nascimento, Brant, Marino, Passaglia, & Ribeiro, 2019).

Di antara pasien yang menderita AMI, 70% kejadian fatal disebabkan oleh oklusi plak *aterosklerotik*. Karena *aterosklerosis* merupakan penyebab utama AMI, faktor risiko penyakit *aterosklerotik* dikurangi dengan pencegahan penyakit. Faktor-faktor risiko yang dapat dimodifikasi pada pria sebesar 90% dan wanita sebesar 94%. Faktor risiko *myocardial infarction* yang dapat dimodifikasi termasuk merokok, olahraga, hipertensi, obesitas, kolesterol, LDL, dan kadar trigliserida (Santoso & Setiawan, 2005). Sebaliknya, usia, jenis kelamin, dan riwayat keluarga adalah faktor risiko biologis yang tidak dapat dimodifikasi untuk *aterosklerosis* (Berg, et al., 2018).

3.7. Statistika Deskriptif

Statistika deskriptif adalah metode yang berkaitan dengan pengumpulan dan penyajian data yang bertujuan untuk memberikan informasi yang berguna berupa bentuk distribusi data. Ukuran dalam statistika deskriptif di antaranya adalah ukuran pemusatan dan penyebaran data.

Ukuran pemusatan data bertujuan untuk memberikan informasi dimana data terkumpul dengan ukuran atau jumlah tertentu. Contoh : *Mean* (rataan) dengan rumus berdasarkan persamaan (3.1).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

di mana:

$\bar{x}$ adalah nilai rata-rata

x_i adalah data ke-i

n adalah jumlah data

Ukuran penyebaran data bertujuan untuk memberikan informasi bagaimana data menyebar di sekitar pusat data. Contoh : *Standar deviasi* (simpangan baku) dengan rumus berdasarkan persamaan (3.2).

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{3.2}$$

di mana:

s adalah nilai standar deviasi

$\bar{x}$ adalah nilai rata-rata

x_i adalah data ke-i

n adalah jumlah data

3.8. *Machine Learning*

Awal mula *machine learning* muncul diperkenalkan oleh Arthur Samuel pada tahun 1959 melalui jurnal yang berjudul "*Some Studies in Machine Learning Using the Game of Checkers*" yang dipublikasikan oleh IBM *Journal of Research and Development* pada Juli 1959 (Samuel, 1959).

Machine Learning atau pembelajaran mesin banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Proses pembelajaran *machine learning* sama dengan *data mining* di mana sistem membutuhkan data untuk menemukan pola, sehingga dapat meningkatkan pemahaman program (Primartha, 2018). *Machine Learning* mencoba untuk menirukan bagaimana proses manusia untuk belajar dan menggeneralisasi. Ciri khas dari *machine learning* adalah adanya proses pelatihan dan pembelajaran, oleh karena itu membutuhkan data untuk dipelajari atau yang

biasa disebut data *training* dan data untuk di uji atau biasa disebut data *testing* (Hania, 2017).

3.9. *Pre-Processing*

Pre-processing adalah proses yang dilakukan untuk menghilangkan atau tidak mengikutsertakan efek non biologis yang terdapat pada data sehingga dapat memberikan hasil yang lebih baik dalam analisis (Serin, 2011).

Menurut (Rong Yang, 2006) proses *pre-processing* pada data *microarray* terbagi menjadi tiga tahapan yaitu *background correction*, *normalization*, dan *summarization*.

1) *Background correction* merupakan tahapan yang digunakan untuk menghilangkan *background noise* (data *noise* merupakan data yang berisi nilai-nilai yang salah atau anomali yang biasanya disebut juga dengan *outlier*). Istilah *background correction* juga berarti penyesuaian *signal* atau menyesuaikan *cross hybridization* yang merupakan pengikat dari DNA non spesifik yang melekat pada *array*.
2) *Normalization* merupakan tahapan untuk menghilangkan variansi non biologis yang tidak diinginkan dan dimungkinkan terdapat pada *array* sehingga data menjadi homogen. (Harteminka, Gifford, Jaakkola, & Young, 2001) menjelaskan bahwa sumber variasi dapat diklasifikasikan menjadi dua yaitu *interesting variation* dan *obscuring variation* (variansi yang kabur/tidak jelas). Sumber yang tidak jelas dari variasi dapat memiliki banyak efek yang berbeda pada data, kecuali *array* tersebut melalui proses normalisasi yang baik.
3) *Summarization* adalah proses yang digunakan untuk mengukur gen yang terdapat pada *chip*, sehingga akan menghasilkan nilai ekspresi gen. Dengan kata lain setiap set *probe microarray Affymetrix GeneChip* terdiri dari beberapa *probe*. Dalam satu set *probe*, masing-masing *probe* mengandung bagian berbeda dari urutan gen tertentu. *Summarization* adalah proses yang menggabungkan beberapa intensitas *probe* untuk setiap set *probe* dan menghasilkan nilai ekspresi gen.

Proses *pre-processing* dapat dilakukan dengan menggunakan bantuan *packages* `AffyPLM` dan dengan menggunakan fungsi perintah `Threestep`.

Perintah `Threestep` pada tahap *pre-processing* merupakan alternatif yang dapat digunakan pada perhitungan ukuran ekspresi gen, yang dilakukan dengan menggunakan RMA.2 pada tahapan *background correction*, *quantile* pada tahapan *normalization* dan median pada tahapan *summarization* (Bolstad, 2004).

3.10. Reduksi Dimensi

Reduksi dimensi adalah suatu proses yang bertujuan untuk meringankan beban komputasi serta menghindari *overfitting* pada klasifikasi (Ma'wa et al., 2019). Beberapa penelitian telah menerapkan dua pendekatan reduksi dimensi untuk menemukan variabel yang relevan dan menghapus redundansi dalam kumpulan data, yaitu dengan menggunakan *feature extraction* dan *feature selection* (Hira & Gillies, 2015). *Feature selection* berfungsi untuk menghapus fitur dan redundansi yang tidak relevan. Tujuan *feature selection* adalah untuk menghilangkan gen yang tidak relevan dan bising dari kumpulan data masukan (*input*) untuk mempercepat waktu pengolahan data dengan mengurangi dimensi data dan untuk menghindari *overfitting* pada proses klasifikasi (Ammu & Preeja, 2013). Sementara itu, *feature extraction* mengubah data asli menjadi representasi baru. *Feature extraction* memiliki tujuan yang sama dengan *feature selection* karena menghilangkan fitur yang tidak relevan atau bising dalam data dan menghilangkan redundansi dalam data yang bertujuan untuk meningkatkan nilai akurasi klasifikasi (Tan, et al., 2014).

3.11. *Independent Principal Component Analysis*

Yao et al. (2012) telah mengembangkan variasi dari metode *Principal Component Analysis* (PCA) yang disebut *Independent Principal Component Analysis* (IPCA). Metode ini menggabungkan keunggulan antara metode PCA dan metode *Independent Component Analysis* (ICA). IPCA mengasumsikan bahwa komponen yang bermakna secara biologis dapat diperoleh jika sebagian besar kebisingan (*noise*) telah dihilangkan dari vektor *loading* terkait. Dengan mengidentifikasi vektor *loading non-Gaussian* dari data biologis, IPCA lebih baik dalam menggambarkan struktur internal data dibandingkan dengan PCA dan ICA.

3.11.1. *Principal Component Analysis*

Principal Component Analysis (PCA) adalah alat reduksi dimensi klasik dan alat ekstraksi fitur dalam analisis *eksploratory* data, serta telah digunakan

dalam berbagai bidang. Ada berbagai cara untuk menyelesaikan PCA. Algoritma yang paling efisien secara komputasi yaitu dengan menggunakan *Singular value decomposition* (SVD): misalkan $\mathbf{X}$ adalah matriks $n \times p$ terpusat, di mana n adalah jumlah sampel (atau pengamatan) dan p adalah jumlah variabel atau entitas biologis yang diukur. Maka SVD dari data matriks $\mathbf{X}$ dapat didefinisikan sebagai berikut:

$$\mathbf{X} = \mathbf{UDV^T} \tag{3.3}$$

di mana:

$\mathbf{U}$ adalah matriks $n \times p$ dengan kolom yang tidak berkorelasi ($\mathbf{U^T U = I_P}$),

$\mathbf{D}$ adalah matriks diagonal $p \times p$ dengan elemen diagonal d_j.

$\mathbf{V}$ adalah matriks ortogonal $p \times p$ ($\mathbf{V^T V = I_P}$),

Hal ini menunjukkan bahwa $\mathbf{u}_j$ adalah kolom $\mathbf{U}$ dan $\mathbf{v}_j$ adalah kolom $\mathbf{V}$. Maka dengan demikian $\mathbf{u}_j d_j$ adalah *principal component* (PC) ke-j dan $\mathbf{v}_j$ adalah vektor *loading* yang sesuai (Jolliffe, 2002). PC adalah kombinasi linear dari variabel asli dan vektor *loading* yang menunjukkan bobot yang ditetapkan untuk masing-masing variabel dalam kombinasi linear. PC ke-1 menyumbang jumlah maksimal dari total varians. Demikian pula, PC ke-j $(j = 2, \dots, p)$ dapat menjelaskan jumlah varians maksimal yang tidak diperhitungkan oleh PC ke-$j-1$ sebelumnya. Oleh karena itu, sebagian besar informasi yang terkandung dalam $\mathbf{X}$ dapat direduksi menjadi beberapa PC dengan harapan bahwa sampel yang termasuk dalam kelompok biologis yang sama atau menjalani perlakuan biologis yang sama akan dikelompokkan bersama dan dipisahkan dari kelompok lain.

- **Kekurangan PCA**

  Terkadang PCA mungkin tidak dapat mengekstraksi informasi yang relevan dan karenanya dapat memberikan PC yang tidak bermakna yang tidak menggambarkan karakteristik data. Alasannya adalah bahwa transformasi liniernya melibatkan statistik *order* kedua untuk mendapatkan PC yang saling tidak ortogonal yang mungkin tidak sesuai untuk data biologis. PCA mengasumsikan bahwa data ekspresi gen memiliki sinyal *Gaussian*, sementara itu telah ditunjukkan bahwa banyak data ekspresi gen sebenarnya memiliki sinyal *'super-Gaussian'* (Lee & Batzoglou, 2003) dan (Huang & Zheng, 2006).

3.11.2. *Independent Component Analysis*

Independent principal component (ICA) pertama kali diusulkan oleh (Comon, 1994). ICA dapat mengurangi efek *noise* atau artefak kebisingan dalam data karena ICA bertujuan untuk memisahkan sinyal yang bercampuran dari sumber yang berbeda. Dengan mengasumsikan sinyal berdistribusi *non-Gaussian*, ICA memodelkan pengamatan sebagai variabel atau komponen yang berkombinasi secara linear, yang dipilih seindependen mungkin secara statistik di mana komponen yang berbeda mewakili informasi berbeda yang tidak tumpang tindih. Karena itu ICA melibatkan statistik *order* yang lebih tinggi (Kong, Vanderburg, Gunshin, Rogers, & Huang, 2008). Faktanya, ICA mencoba untuk memulihkan sinyal yang independen secara statistik dari pengamatan campuran linier yang tidak diketahui. Beberapa algoritma seperti *FastICA*, *Kernel ICA* (Bach & Jordan, 2002) dan *ProDenICA* (Hastie & Tibshirani, 2002) diusulkan untuk mengestimasi *Independent Component* (IC). Algoritma *FastICA* memaksimalkan sifat *non-Gaussian* dari masing-masing komponen, sementara *Kernel ICA* dan *ProDenICA* meminimalkan informasi timbal balik antar komponen. Pada penelitian tersebut, Yao et al. (2012) memilih untuk menggunakan algoritma *FastICA*.

Dalam algoritma *FastICA*, PCA digunakan sebagai langkah *pre-processing* untuk memutihkan matriks data. Jika persamaan (3.3) dibentuk ulang maka diperoleh:

$$\mathbf{U^T} = \mathbf{D^{-1}V^TX^T} \tag{3.4}$$

dikarenakan kolom $\mathbf{V}$ bersifat ortonormal. Baris-baris $\mathbf{U^T}$ tidak berkorelasi dan memiliki rata-rata sama dengan nol. Untuk menyelesaikan langkah pemutihan/*whitening*, maka $\mathbf{U^T}$ dapat dikalikan dengan $\sqrt{n-1}$, sehingga baris-baris $\mathbf{U^T}$ memiliki variansi *unit*. Dengan demikian, maka $\tilde{\mathbf{U}}$ menjadi PC yang telah diputihkan $\left(\tilde{\mathbf{U}} = \sqrt{\mathbf{n-1}}\mathbf{U^T}\right)$. IC diestimasi melalui persamaan berikut:

$$\mathbf{S} = \mathbf{W}\tilde{\mathbf{U}} \tag{3.5}$$

ICA mengasumsikan bahwa distribusi *Gaussian* mewakili kebisingan, dan oleh karena itu ICA bertujuan untuk mengidentifikasi komponen *non-Gaussian* dalam ruang sampel yang bersifat independen. Studi terbaru telah mengamati bahwa distribusi sinyal data *microarray* biasanya berdistribusi *super-Gaussian*

karena hanya sejumlah kecil gen yang berkontribusi besar pada proses biologis (Lee & Batzoglou, 2003) dan (Engreitz, Daigle Jr., Marshall, & Altman, 2010).

Ada dua pengukuran kuantitatif klasik untuk kasus *Gaussian* yaitu kurtosis dan *negentropy*. Kurtosis dapat didefinisikan sebagai berikut:

$$\boldsymbol{K = E\{s_i^4\} - 3}. \tag{3.6}$$

di mana:

$\mathbf{s_i}$ adalah bari dari $\mathbf{S}$, yang memiliki rata-rata sama dengan 0 dan *unit* variansi, $j = 1, \dots, n$.

Nilai kurtosis sama dengan nol jika $\mathbf{s_i}$ memiliki *probability density function* (pdf) *Gaussian*, bernilai positif jika $\mathbf{s_i}$ memiliki pdf yang runcing (*super-Gaussian*, yaitu pdf relatif besar pada titik nol) dan bernilai negatif jika $\mathbf{s_i}$ memiliki pdf yang datar (*sub-Gaussian*, yaitu pdf agak konstan mendekati nilai nol).

Pada kasus ini yang diperhatikan adalah pdf yang runcing dan datar (pdf *non-Gaussian*) karena sifat *non-Gaussian* di sini dianggap sebagai independen (Hyvärinen & Oja, 2000). Meskipun kurtosis sangat baik secara komputasi dan simpel secara teori, namun bisa sangat sensitif terhadap pencilan. Scholz, Sterling, Gatzek, Fiehn dan Selbig (2004) mengusulkan untuk mengurutkan IC berdasarkan nilai kurtosis.

Dalam algoritma *FastICA*, *negentropy* digunakan karena merupakan pengukuran *non-Gaussian* yang sangat baik. *Negentropy* sama dengan nol jika $\mathbf{s_i}$ adalah Gaussian dan bernilai positif jika $\mathbf{s_i}$ adalah *non-Gaussian*. Ini tidak hanya mudah untuk dihitung, tetapi juga sangat kuat (Hyvärinen & Oja, 2000). Namun, pengukuran ini tidak dapat membedakan antara *super-Gaussian* dan *sub-Gaussian*.

- **Kekurangan ICA**

  Mirip dengan PCA, ICA juga kewalahan ketika menghadapi kasus berdimensi tinggi, yang kadang-kadang menyebabkan ketidakmampuan IC untuk mencerminkan struktur internal data (yang diharapkan secara biologis). Selain itu, karena ICA adalah algoritma stokastik, maka terdapat masalah konvergensi dengan *local optima*, yang menyebabkan ICA menghasilkan IC yang sedikit berbeda ketika menganalisis kembali data yang sama (Himberg, Hyvarinen, & Esposito, 2004).

3.11.3. *Independent Principal Component Analysis*

Untuk mengurangi kebisingan (*noise*) dan mencerminkan struktur internal data yang dihasilkan oleh eksperimen biologis, maka Yao et al. (2012) mengusulkan pendekatan baru yang disebut *Independent Principal Component Analysis* (IPCA). Metode tersebut lebih memilih untuk mengurangi kebisingan dalam vektor *loading* daripada mengurangi kebisingan secara langsung pada data atau PC, seperti yang dilakukan metode ICA. Perlu diingat bahwa PC adalah kombinasi linear dari variabel asli yang diberi bobot berdasarkan elemen-elemen dari PC itu sendiri dalam vektor *loading* yang sesuai. Dengan demikian akan diperoleh PC yang kebisingannya telah dihilangkan dengan menggunakan metode ICA sebagai proses *denoising* dari vektor *loading* terkait.

Yao et al. (2012) mengasumsikan bahwa dalam sistem biologis, variabel yang berbeda (entitas biologis, seperti gen dan metabolit) memiliki tingkat ekspresi yang berbeda tergantung pada kondisi biologis itu sendiri. Karena itu, hanya beberapa variabel yang berkontribusi pada proses biologis. Variabel yang relevan ini harus memiliki bobot penting dalam vektor *loading* sementara variabel yang tidak relevan atau berisik (*noisy*) lainnya harus memiliki bobot yang sangat kecil. Bahkan begitu vektor-vektor *loading* telah melalui proses *denoising*, maka diharapkan vektor *loading* memiliki distribusi *super-Gaussian* yang berbeda dengan distribusi *Gaussian* ketika kebisingan dimasukkan. Memaksimalkan sifat *non-Gaussian* pada vektor *loading* akan memungkinkan untuk menghapus sebagian besar kebisingan.

Adapun langkah-langkah atau algoritma dari metode IPCA adalah sebagai berikut:

1. Menerapkan SVD pada matriks data terpusat $\mathbf{X}$ untuk menghasilkan vektor *loading* $\mathbf{V}$ yang telah diputihkan, dan memilih jumlah komponen m untuk reduksi dimensi.
2. Mengimplementasikan *FastICA* pada vektor *loading* $\mathbf{V}$ dan mendapatkan vektor *loading* independen $\mathbf{S^T}$.
3. Memproyeksikan matriks data terpusat $\mathbf{X}$ pada vektor *loading* independen $\mathbf{s}_j$ yang berukuran m dan mendapatkan PC Independen (IPC) $\tilde{\mathbf{u}}_j, j = 1, \dots, m$.

4. Mengurutkan IPC berdasarkan nilai kurtosis dari vektor *loading* independen yang sesuai.

Penjelasan dari algoritma tersebut adalah sebagai berikut:

- **Mengekstrak vektor *loading* dari PCA**

PCA diterapkan ke matriks data terpusat $\mathbf{X}(n \times p)$ menggunakan SVD untuk mengekstrak vektor *loading*:

$$\mathbf{X} = \mathbf{UDV^T}, \tag{3.7}$$

di mana:

kolom **V** berisi vektor *loading*.

Karena rata-rata setiap vektor *loading* bernilai sangat mendekati nol, maka dapat dikatakan bahwa vektor *loading* tersebut telah diputihkan. Oleh karena itu algoritma *FastICA* dapat diterapkan pada vektor *loading*.

- **Mereduksi dimensi**

Reduksi dimensi memungkinkan interpretasi yang lebih jelas tanpa beban komputasi. Oleh karena itu, hanya sejumlah kecil vektor *loading*, atau dengan kata lain hanya sejumlah kecil PC yang diperlukan untuk merangkum sebagian besar informasi yang relevan yang terkandung dalam data. Namun, belum ada kriteria yang diterapkan secara global tentang cara memilih jumlah PC yang akan disimpan. Telah ditunjukkan bahwa nilai kurtosis vektor *loading* independen memberikan indikasi *post hoc* jumlah komponen independen yang akan dipilih. Dengan demikian telah diamati secara eksperimental bahwa 2 atau 3 komponen sudah cukup untuk menyoroti karakteristik data yang bermakna dan membuang banyak kebisingan atau informasi yang tidak relevan.

- **Menerapkan ICA pada vektor *loading***

Sifat *non-Gaussian* dari vektor *loading* dapat dimaksimalkan menggunakan persamaan berikut:

$$\mathbf{S} = \mathbf{W}\tilde{\mathbf{V}}^{\mathbf{T}}, \tag{3.8}$$

di mana:

$\tilde{\mathbf{V}}$ adalah $(p \times m)$ matriks yang mengandung vektor m *loading* yang dipilih, $\mathbf{W}$ adalah $(m \times m)$ matriks yang bercampur dan

$\mathbf{S}$ adalah $(m \times p)$ matriks yang barisnya adalah vektor *loading* independen.

Independent Principal Component (IPC) baru dapat diperoleh dengan memproyeksikan $\mathbf{X}$ pada $\mathbf{S^T}$:

$$\tilde{\mathbf{U}} = \mathbf{X}\mathbf{S^T} \tag{3.9}$$

di mana:

$\tilde{\mathbf{U}}$ adalah matriks $(n \times m)$ yang kolomnya berisi IPC.

- **Mengurutkan IPC**

Perlu diingat bahwa ICA memperoleh komponen yang tidak berurutan dan bahwa nilai kurtosis menunjukkan karakteristik *Gaussian* dari pdf. Scholz et al. (2004) mengusulkan untuk menggunakan nilai kurtosis dari IC untuk mengurutkan komponen. Dalam metode IPCA diusulkan untuk mengurutkan IPC sesuai dengan nilai kurtosis dari vektor *loading* independen $\mathbf{s_j}$ $(j = 1, \dots, m)$, dikarenakan pada vektor *loading* dengan pdf runcing, yang ditunjukkan oleh nilai kurtosis yang besar.

3.12. Klasifikasi

Salah satu masalah penting dalam *data mining* atau *machine learning* adalah klasifikasi (Zhang, 2004). Klasifikasi adalah metode yang dapat digunakan untuk mengestimasi kelas pada suatu objek yang kelas/labelnya tidak diketahui (Marlina, Muslim, & Siahaan, 2016).

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek (Mulyanto, 2009). Gorunescu (2011) dalam bukunya menyatakan bahwa proses klasifikasi didasarkan pada empat komponen fundamental yaitu *classes*, *predictors*, *training data*, dan *testing data*.

3.13. *Support Vector Machine*

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di *Annual Workshop on Computational Learning Theory* (COLT-92) sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang *pattern recognition*. SVM adalah metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan

tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space* (Nugroho, Witarto, & Handoko, 2003). SVM adalah alat klasifikasi dan prediksi regresi yang menggunakan teori *machine learning* untuk memaksimalkan akurasi prediksi dan sekaligus secara otomatis menghindari masalah *overfitting* pada data (Jakkula, 2006).

SVM memiliki prinsip dasar *linier classifier* yaitu kasus klasifikasi yang hanya secara linier dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada problem non-linier dengan memasukkan konsep *kernel* pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak (*margin*) antara kelas data (Octaviani, Wilandari, & Ispriyanti, 2014).

Menurut Santosa (2007) *hyperplane* klasifikasi linier SVM dinotasikan sebagai berikut:

$$f(x) = W^T x + b \tag{3.10}$$

sehingga menurut Cortes & Vapnik (1995) diperoleh persamaan

$$[(w^T x_i + b)] \geq 1 \text{ untuk } y_i = +1 \tag{3.11}$$

$$[(w^T x_i + b)] \leq 1 \text{ untuk } y_i = -1 \tag{3.12}$$

di mana:

$\boldsymbol{x}_i$ = himpunan data *training*, i = 1, 2, ..., n dan y_i = label kelas dari $\boldsymbol{x}_i$

Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM (Nugroho et al., 2003). Untuk mendapatkan *hyperplane* terbaik adalah dengan mencari *hyperplane* yang terletak di tengah-tengah antara dua bidang pembatas kelas dan untuk mendapatkan *hyperplane* terbaik itu, sama dengan memaksimalkan *margin* atau jarak antara dua set objek dari kelas yang berbeda (Santosa, 2007). *Margin* dapat dihitung dengan $\frac{2}{\|w\|}$.

Untuk mencari *hyperplane* terbaik dapat digunakan metode *Quadratic Programm*ing (QP) *problem*, yaitu mencari titik minimal (persamaan (3.13)), dengan memperhatikan *constraint* (persamaan (3.14)).

$$\min \tau(w) = \frac{1}{2} w^T w \tag{3.13}$$

$$y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, n \quad (3.14)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, di antaranya *Lagrange Multiplier* yang dilakukan oleh Cortes & Vapnik (1995).

$$L(w, b, \alpha) = \frac{1}{2} w^T w = \sum_{i=1}^{n} \alpha_i \{y_i[(w^T x_i) + b] - 1\} \; i = 1, 2, \dots, n \quad (3.15)$$

di mana:

α_i adalah fungsi *Lagrange multipliers* yang bernilai nol atau positif ($\alpha_i \geq 0$).

Nilai optimal dari persamaan (3.15) dapat dihitung dengan meminimalkan *L* terhadap *w* dan *b* , dan memaksimalkan *L* terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal *gradient L* = 0, persamaan (3.15) dapat dimodifikasi sebagai memaksimalkan *problem* yang hanya mengandung α_i (Hastie, Tibshirani, & Friedman, 2008), sebagaimana persamaan (3.16) di bawah.

Maximize:

$$max_d \; L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (3.16)$$

dengan:

$$\alpha_i \geq 0 (i = 1, 2, \dots, n) \text{ dan } \sum_{i=1}^{n} \alpha_i y_i = 0$$

Dari hasil dari perhitungan ini diperoleh α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai *support vector* (Nugroho et al., 2003).

3.13.1. *Soft Margin*

Dalam metode SVM diasumsikan bahwa kedua belah kelas terpisah secara sempurna oleh *hyperplane*. Namun pada kenyataannya kedua buah kelas tersebut tidak selalu terpisah secara sempurna. Hal ini menyebabkan optimasi tidak terpenuhi. Untuk mengatasi masalah ini, menurut Nugroho et al., (2003) SVM dirumuskan ulang dengan menggunakan teknik *soft margin*. Dalam *soft margin* persamaan (3.14) dimodifikasi dengan memasukkan variabel *slack* $\xi_i (\xi_i > 0)$ sebagai berikut:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, n \quad (3.17)$$

Dengan demikian, persamaan (3.13) diubah menjadi persamaan berikut:

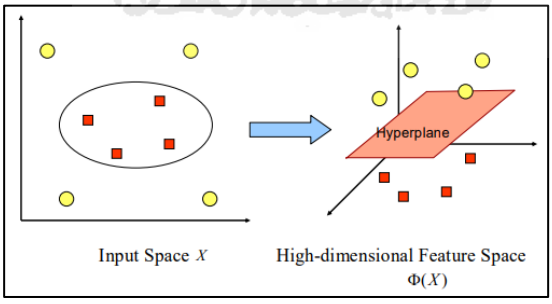
$$\min \tau(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \quad (3.18)$$

Parameter C dipilih untuk mengontrol *trade-off* antara margin dan *error* klasifikasi ξ atau nilai kesalahan pada klasifikasi. Nilai C yang besar berarti akan memberikan penalti yang lebih besar terhadap *error* klasifikasi tersebut (Nugroho et al., 2003).

3.13.2. *Kernel* SVM

Banyak teknik *machine learning* yang dikembangkan dengan asumsi kelinieran, sehingga algoritma yang dihasilkan terbatas atau hanya tersedia untuk kasus-kasus yang linier saja (Santosa, 2007). SVM dapat bekerja pada data non-linier dengan menggunakan pendekatan *kernel* pada fitur data awal himpunan data. Fungsi *kernel* digunakan untuk memetakan dimensi awal (dimensi yang lebih rendah) dari himpunan data *input* ke ruang dimensi baru (dimensi yang relatif lebih tinggi) (Octaviani et al., 2014). Ruang dimensi yang lebih tinggi ini disebut ruang *kernel* yang dapat digunakan untuk memisahkan data secara linear (Awad & Khanna, 2015).

Dalam non linear SVM, pertama-tama data X dipetakan oleh fungsi $\Phi(X)$ ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, *hyperplane* yang memisahkan kedua *class* tersebut dapat dikonstruksikan (Nugroho et al., 2003).



Sumber: Nugroho et al. (2003)

Gambar 3.1 Ilustrasi fungsi $\Phi(X)$

Gambar di atas merupakan ilustrasi kinerja *kernel*. Pada gambar (kiri) diperlihatkan data pada *class* kuning dan data pada *class* merah yang berada pada *input space* berdimensi dua tidak dapat dipisahkan secara linear. Selanjutnya gambar (kanan) menunjukkan bahwa fungsi Φ memetakan tiap data pada *input*

space tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), di mana kedua *class* dapat dipisahkan secara linear oleh sebuah *hyperplane*.

Menurut (Prasetyo, 2012), secara umum, fungsi *kernel* yang sering digunakan adalah sebagai berikut :

1) Linear : $x^T x$,

   Kernel Linear merupakan fungsi *kernel* yang paling sederhana, digunakan ketika data yang akan diklasifikasikan telah dapat dipisahkan dengan *hyperplane*.

2) *Polynomial*

   Polynomial merupakan sebuah metode populer untuk model non Linear.

   $K(x, x') = ((x, x') + 1)^d$ atau $K(x, x') = (x, x')^d$ (3.19)

3) *Gaussian Radial Basis Function* (RBF)

   $K(x, x') = exp\left(\frac{\|x, x'\|}{2\sigma^2}\right)$ (3.20)

Fungsi *kernel* mana yang harus digunakan untuk substitusi *dot product* dalam *feature space* sangat tergantung pada data karena fungsi *kernel* ini akan menentukan fitur baru di mana *hyperplane* akan dicari (Santosa, 2007).

3.14. *Naive Bayes Classifier*

Naive Bayes Classifier (NBC) adalah salah satu algoritma yang terkenal yang merupakan klasifikasi probabilistik berdasarkan Teorema Bayes (Dey, Chakraborty, Biswas, Bose, & Tiwari, 2016). NBC merupakan algoritma yang paling populer karena sederhana, efisien, dan memiliki kinerja yang baik dalam klasifikasi data (Jabbar & Samreen, 2016) dan juga memiliki tingkat efisiensi pembelajaran yang sangat tinggi dan dapat digunakan untuk memperkirakan semua kemungkinan dalam mempelajari data *training* (An, Sun, & Wang, 2017).

Menurut Gorunescu (2011), konsep dasar dari metode ini adalah Teorema Bayes, dalam statistika teorema ini banyak digunakan untuk menghitung peluang. Bila diketahui $A_1, A_2, \dots, A_n$ adalah sebuah kejadian yang independen dalam ruang sampel Ω, sehingga $\sum_{i=1}^{n} A_i = \Omega$. B merupakan sebuah kejadian random dan kejadian $A_1 \cap B, A_2 \cap B, \dots, A_n \cap B$ merupakan partisi dalam B.

$P(B) = \sum_{i=1}^{n} P(A_i \cap B)$ (3.21)

Bila $P(A_i) > 0$ saat $i = 1,2, \dots, n$ maka $P(A_i \cap B) = P(B|A_i)P(A_i)$.

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{3.22}$$

Kejadian random $A_1, A_2, \dots, A_n$ dan B adalah partisi dari ruang sampel Ω . Jika $P(B) > 0$ dan $P(A_i) > 0$ untuk $i = 1,2, \dots, n$, maka:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{3.23}$$

$P(A_i|B)$ merupakan *posterior probability* karena nilai $P(A_i|B)$ bergantung pada nilai B. $P(A_i)$ disebut *prior probability* karena nilainya tidak bergantung pada nilai B, sedangkan $P(B|A_i)$ adalah fungsi *likelihood* dan $P(B)$ merupakan keterangan.

Metode NBC menggunakan konsep dari Teorema Bayes. Dengan demikian bila diberikan $A_1, A_2, \dots, A_n$ maka atribut tersebut akan digunakan untuk menentukan kelas C, dengan menggunakan Teorema Bayes maka perhitungan *posterior probability* untuk setiap kelas C adalah sebagai berikut (Gorunescu, 2011):

$$P(C_j|A_1, A_2, \dots, A_n) = \frac{P(A_1,A_2,\dots,A_n|C_j)P(C_j)}{\sum_{i=1}^{n} P(A_1,A_2,\dots,A_n)} \tag{3.24}$$

Apabila kelas tersebut memaksimalkan nilai $P(C_j|A_1, A_2, \dots, A_n)$ atau memaksimalkan nilai $P(A_1, A_2, \dots, A_n|C_j)$, maka kelas tersebut yang dipilih. Berdasarkan persamaan di atas, diperlukan perhitungan $P(A_1, A_2, \dots, A_n|C_j)$. Setiap atribut diasumsikan independen untuk setiap kelas C. Apabila terdapat atribut yang memiliki sifat kuantitatif atau kontinu, maka $P(A_i|C_j)$ dihitung dengan pendekatan distribusi normal sebagai berikut:

$$P(A_i|C_j) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} exp\left(\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}\right) \tag{3.25}$$

Dugaan peluang $P(A_i|C_j)$ dapat dihitung untuk setiap atribut A_i dan kelas C_j, sehingga data baru dapat diklasifikasikan ke dalam kelas k C_k jika peluang yang didapat merupakan yang terbesar di antara peluang lainnya.

3.15. Evaluasi Kinerja Klasifikasi

Menurut Xhemali, Hinde dan Stone (2009) *confusion matrix* adalah tabel prediktif klasifikasi yang dapat digunakan untuk mengukur hasil evaluasi dari klasifikasi tersebut.

3.15.1. *Confusion Matrix*

Confusion matrix menurut Han dan Kamber (2006) adalah suatu alat yang memiliki fungsi untuk melakukan analisis yang bertujuan untuk mengetahui apakah metode *classifier* yang digunakan dapat membedakan kelas klasifikasi dengan baik atau tidak. *Confusion matrix* adalah alat ukur berbentuk matriks yang digunakan untuk mendapatkan jumlah ketepatan/akurasi klasifikasi terhadap kelas.

Tabel 3.1 *Confusion matrix* untuk dua kelas klasifikasi

Confusion Matrix		**Nilai Aktual**	
		Positif	Negatif
Nilai Prediksi	Positif	TP	FN
	Negatif	FP	TN

Confusion matrix berukuran $n \times n$ terkait dengan hasil klasifikasi yang memperlihatkan nilai aktual dan nilai prediksi, di mana n adalah banyak kelas. Tabel di atas merupakan *confusion matrix* untuk $n = 2$, di mana nilai yang terkandung di dalamnya adalah sebagai berikut (Visa, Ramsay, Ralescu, & Knaap, 2011):

Keterangan :

- *True Positive* (TP) adalah jumlah prediksi positif yang benar;
- *False Positive* (FP) adalah jumlah prediksi positif yang salah;
- *False Negative* (FN) adalah jumlah prediksi negatif yang salah;
- *True Negative* (TN) adalah jumlah prediksi negatif yang benar.

Nilai dari TN dan TP memberikan informasi ketika *classifier* dalam melakukan klasifikasi data bernilai benar, sedangkan FP dan FN memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data.

Dari hasil *confusion matrix* dapat digunakan untuk atau menghitung nilai *accuracy*, *precision*, *recall* dan *f-score*. Nilai *accuracy* adalah nilai kedekatan antara nilai prediksi dengan nilai aktual; nilai *precision* adalah ketepatan atau

ketelitian dalam pengklasifikasian; dan nilai *recall* adalah proporsi positif aktual yang benar diidentifikasi (Sasongko, 2016). Nilai *recall* juga sering disebut sebagai nilai *sensitivity*.

F-Score digunakan karena meskipun nilai *precision* dan *recall* masing-masing sudah memiliki nilai yang valid, akan tetapi salah satu di antara kedua nilai tersebut dapat dioptimalkan dengan mengorbankan nilai yang lain. *F-Score* hanya mendapatkan hasil yang tinggi jika *precision* dan *recall* sama-sama mempunyai nilai seimbang, jadi nilai ini sangat signifikan (Turney, 1999). Adapun rumus perhitungan nilai-nilai tersebut adalah sebagai berikut (Xhemali et al., 2009):

- *Accuracy*: Proporsi dari jumlah total prediksi yang benar:

  $$Accuracy(\%) = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (3.26)$$

- *Precision*: Proporsi dari *class* 'positif' yang diprediksi dengan benar:

  $$Precision(\%) = \frac{TP}{(TP+FP)} \quad (3.27)$$

- *Recall/Sensitivity*: Proporsi *class* 'positif' yang diidentifikasi dengan benar:

  $$Recall(\%) = \frac{TP}{(TP+FN)} \quad (3.28)$$

- *F-Score*: nilai keseimbangan dari nilai *precision* dan nilai *recall*:

  $$F - Score(\%) = \frac{(2 \times Precision \times Recall)}{(Precision+Recall)} \quad (3.29)$$

Selain pengukuran tersebut juga terdapat pengukuran lain yang dapat digunakan yaitu *specificity* dan AUC. Pengukuran ini berguna sebagai ukuran statistik dari kinerja klasifikasi biner, mengukur model terbaik dan memilih model yang paling efisien. Berbeda dengan *sensitivity* (*recall*) yang mengukur proporsi positif yang benar diidentifikasi, *Specificity* sebaliknya mengukur proporsi negatif yang benar diidentifikasi (Bramer, 2013). *Specificity* berbanding terbalik dengan *sensitivity* dalam arti *specificity* meningkat ketika *sensitivity* menurun di berbagai ambang batas (Kumar & Indrayan, 2011).

Nilai *area under the curve* (AUC) bertujuan untuk mengukur kinerja diskriminatif dengan memperkirakan probabilitas *output* dari sampel yang dipilih secara acak dari populasi berupa sampel positif atau negatif. AUC sering digunakan untuk mengukur kualitas *classifier* probabilistik (Eng, 2005). Oleh karena itu nilai AUC akan selalu berkisar antara 0 sampai 1, semakin besar nilai AUC maka

semakin kuat klasifikasi yang digunakan. Level pengukuran kualitas *classifier* menggunakan AUC dilihat berdasarkan akurasi dengan rentang yang diperlihatkan pada tabel berikut:

Tabel 3.2 Nilai AUC

Rentang Akurasi	**Kualitas *Classifier***
0,91 – 1	Excellent
0,81 – 0,90	*Good*
0,71 – 0,80	Fair
0,61 – 0,70	*Poor*
≤ 0,60	*Failure*

Sumber: (Defiyanti & Jajuli, 2015)

Perhitungan nilai *specificity* dan AUC dapat dilakukan menggunakan rumus berikut:

- *Specificity*: Proporsi *class* 'negatif' yang diidentifikasi dengan benar:

$$Specificity(\%) = \frac{TN}{(TN+FP)} \tag{3.30}$$

- AUC: Proporsi *class* 'positif' yang diidentifikasi dengan benar:

$$AUC = \frac{(Recall+Specificity)}{2} \tag{3.31}$$

Tabel 3.3 *Confusion matrix* untuk klasifikasi *multi class*

Confusion Matrix		**Nilai Aktual**			
		Kelas 1	Kelas 2	...	Kelas *m*
Nilai Prediksi	Kelas 1	x_{11}	x_{12}	...	x_{1n}
	Kelas 2	x_{21}	x_{22}	...	x_{2n}
	⋮	⋮	⋮	⋱	⋮
	Kelas *m*	x_{m1}	x_{m2}	...	x_{mn}

Sumber: (Manliguez, 2016)

Tabel di atas merupakan *confusion matrix* untuk klasifikasi *multi class* atau jumlah kelas $n \geq 3$. Adapun nilai *accuracy*, *precision*, *recall*, dan *specificity* untuk klasifikasi *multi class* dapat dihitung menggunakan rumus berikut (Sokolova & Lapalme, 2009):

$$accuracy(\%) = \frac{\sum_{i=1}^{I} TP_i}{N} \tag{3.32}$$

$$precision(\%) = \frac{\sum_{i=1}^{I} \frac{TP_i}{TP_i + FP_i}}{I} \quad (3.33)$$

$$recall(\%) = \frac{\sum_{i=1}^{I} \frac{TP_i}{TP_i + FN_i}}{I} \quad (3.34)$$

$$Specificity(\%) = \frac{\sum_{i=1}^{I} \frac{TN_i}{TN_i + FP_i}}{I} \quad (3.35)$$

Keterangan (Sokolova & Lapalme, 2009):

- TP_i adalah jumlah prediksi positif yang benar pada kelas ke-i;
- TN_i adalah jumlah prediksi negatif yang benar pada kelas ke-i;
- FN_i adalah jumlah prediksi negatif yang salah pada kelas ke-i;
- FP_i adalah jumlah prediksi positif yang salah pada kelas ke-i;
- I adalah jumlah kelas
- N adalah total keseluruhan prediksi

3.15.2. Kurva *Receiver Operating Characteristic*

Kurva *Receiver Operating Characteristic* (ROC) adalah salah satu teknik yang dapat memvisualisasikan, mengorganisasi, dan memilih *classifier* berdasarkan performanya (Vuk & Curk, 2006). ROC merupakan hasil dari pengukuran klasifikasi yang berbentuk 2 dimensi di mana garis horizontal menggambarkan nilai *false positive* (FP) dan garis vertikal sebagai *true positive* (TP) (Hastuti, 2012). Visualisasi hasil perhitungan digambarkan dengan menggunakan *area under ROC curve* (AUC) (Hidayatullah, Prasetyo, Sari, & Pratiwi, 2014). *Area Under Curve* (AUC) adalah ukuran yang efektif untuk menilai validitas dari tes diagnostik (Kumar & Indrayan, 2011). Ulasan tentang AUC telah dijelaskan pada bagian sebelumnya.

BAB IV
METODOLOGI PENELITIAN

4.1. Tempat dan Waktu Penelitian

Penelitian dilakukan di kampus terpadu Universitas Islam Indonesia yang bertempat di Yogyakarta. Waktu pelaksanaan penelitian dilakukan dari bulan Agustus 2020 sampai dengan bulan Oktober 2020.

4.2. Data dan Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari *website National Center of Biotechnology Information* (NCBI) https://www.ncbi.nlm.nih.gov/. Objek Penelitian adalah data *microarray* yang berisi tentang informasi bioteknologi seperti DNA, protein, senyawa aktif dan taksonomi. Data yang digunakan adalah data *microarry* dengan kode akses GEO *Series* (GSE) yaitu GSE43837, GSE16515, dan GSE48060.

Tabel 4.1 Informasi data

Set Data	Kode	Sampel	Gen	Kelas
Kanker Payudara	GSE43837	38	61359	2
Kanker Pankreas	GSE16515	52	54675	2
Penyakit AMI	GSE48060	52	54675	3

GSE43837 adalah data yang berisi ekspresi gen kanker payudara yang terdiri dari dua kelas yaitu pasien yang terkena metastasis otak dan tidak terkena metastasis otak dengan jumlah sampel sebanyak 38 dan jumlah ekspresi gen sebanyak 61359. Data GSE16515 merupakan kumpulan sampel data ekspresi gen dari Mayo Clinic yang terdiri dari dua kategori yaitu sampel tumor dan sampel normal pada pasien kanker pankreas. Set data kanker pankreas merupakan kumpulan data ekspresi gen sebanyak 54675 dan terdiri dari 52 sampel pasien yang menderita kanker pankreas yang dibedakan berdasarkan sampel jaringan tumor dan sampel jaringan normal. Sedangkan GSE48060 adalah data ekspresi gen dari pasien *Acute Myocardial Infarction* (AMI) dengan jumlah sampel sebanyak 52 dan ekspresi gen sebanyak 54675. Data ini terdiri dari tiga kategori yaitu: (1) Pasien dengan kejadian berulang (*Patient with recurrent events*); (2) Pasien tanpa kejadian berulang (*Patient without recurrent events*) dan (3) *Normal control*.

4.3. Variabel Penelitian

Adapun variabel yang digunakan dalam penelitian ini beserta definisi operasional variabel dijelaskan pada tabel berikut:

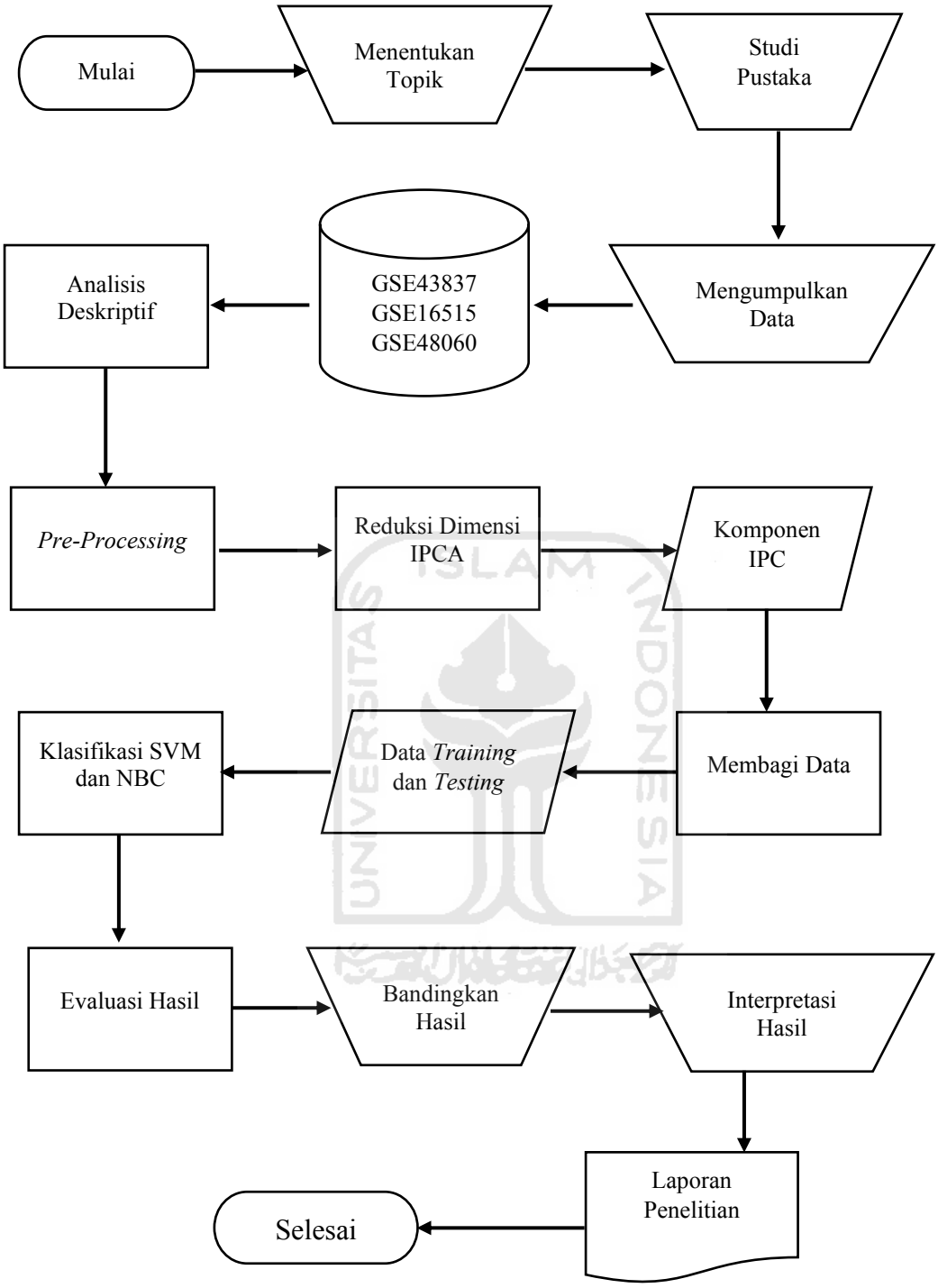
Tabel 4.2 Variabel dan definisi operasional variabel

Variabel	Definisi Operasional
Gen	Nama gen yang mempunyai nilai intensitas ekspresi gen.
Tissue (Jaringan)	Variabel yang menunjukkan sekumpulan sel yang menyusun setiap tubuh manusia. Variabel jaringan ini berbeda pada tiap data yang digunakan.
Usia	Variabel usia menjelaskan tentang usia pasien yang tercatat dalam set data. Variabel usia tersedia pada set data kanker payudara dan kanker pankreas.
Jenis Kelamin	Variabel yang menjelaskan identitas jenis kelamin yang terdapat pada set data kanker pankreas
Disease Status (Status Penyakit)	Variabel *disease status* menjelaskan tentang kategori status penyakit yang diderita oleh pasien yang terdapat pada set data. Pada set data kanker payudara dan kanker pankreas masing-masing terdapat 2 status penyakit dan pada set data penyakit AMI terdapat 3 status penyakit.

4.4. Metode Analisis Data

Metode analisis data yang digunakan dalam penelitian ini adalah analisis deskriptif untuk melihat gambaran secara umum mengenai set data yang digunakan. Kemudian menerapkan metode *Independent Principal Component Analysis* (IPCA) untuk melakukan proses reduksi dimensi terhadap data. Metode IPCA diterapkan dalam R menggunakan *package mixomics* yang didapatkan dari http://mixomics.qfab.org. Setelah itu diterapkan metode *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC) untuk melakukan proses klasifikasi. *Software* yang digunakan untuk pengolahan data adalah *Rstudio version 3.6.1*.

4.5. Langkah Penelitian



Gambar 4.1 *Flowchart* penelitian

Keterangan:

a) Tahap pertama yang dilakukan peneliti sebelum memulai penelitian adalah menentukan topik penelitian kemudian mencari studi pustaka atau referensi tentang penelitian-penelitian yang pernah dilakukan sebelumnya yang sesuai dengan topik yang telah ditentukan. Setelah itu peneliti mengidentifikasi dan merumuskan masalah penelitian serta menentukan tujuan yang ingin dicapai dari penelitian yang akan dilakukan.

b) Pengumpulan Data

Pada tahap ini, peneliti mengumpulkan data yang dibutuhkan sebagai objek penelitian. Data tersebut diambil dari situs web NCBI yang berupa data *microarray*.

c) Analisis Deskriptif

Tahap ini dilakukan untuk melihat deskripsi data yang diperoleh dari *pheno* data yang berisi variabel jumlah gen, kategori jaringan, kategori penyakit yang ada, umur, jenis kelamin dan lainnya.

d) *Pre-Processing*

Setelah mengetahui informasi apa saja yang ada pada data, selanjutnya adalah tahap *pre-processing*. Tahap ini dilakukan agar data yang dimiliki bersifat normal dan tidak terdapat data pengganggu seperti data kosong, *outlier* dan lainnya.

e) Reduksi Dimensi

Tahap selanjutnya adalah melakukan proses reduksi dimensi terhadap data menggunakan metode IPCA dan memilih jumlah komponen IPC yang akan digunakan pada proses klasifikasi.

f) Klasifikasi

Sebelum melakukan klasifikasi data, peneliti terlebih dahulu menentukan data *training* dan data *testing* dari set data hasil reduksi dimensi. Setelah itu proses klasifikasi dilakukan menggunakan metode SVM dan metode NBC.

g) Evaluasi

Setelah hasil klasifikasi didapatkan maka langkah selanjutnya yaitu melakukan evaluasi dengan melihat nilai akurasi dari masing-masing metode klasifikasi dan membandingkan hasil tersebut.

h) Interpretasi Hasil dan penulisan laporan

Setelah mendapatkan hasil yang diinginkan, maka peneliti melakukan interpretasi hasil yang dituangkan dalam penulisan laporan akhir penelitian.

BAB V
HASIL DAN PEMBAHASAN

5.1. Analisis Deskriptif

Analisis deskriptif adalah analisis yang dilakukan untuk menilai karakteristik dari sebuah sampel atau gambaran umum tentang suatu data yang akan diamati. Dalam penelitian ini terdapat tiga set data ekspresi gen yang digunakan untuk membantu proses analisis. Ketiga data tersebut memiliki kode *series database* GSE yang berbeda yaitu GSE43837 untuk set data ekspresi gen kanker payudara, GSE16515 untuk set data ekspresi gen kanker pankreas dan GSE48060 untuk set data ekspresi gen penyakit AMI.

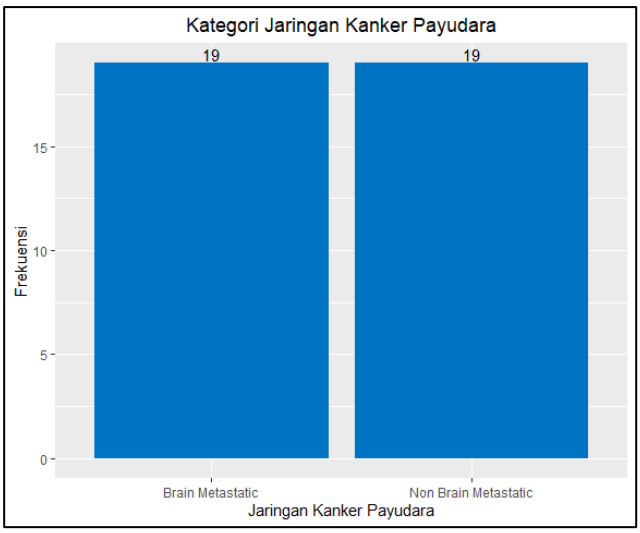
Informasi tentang deskripsi data dapat diperoleh dari *pheno* data dengan menerapkan perintah `phenoData()` pada *series* data GSE. Dalam *pheno* data tersebut terdapat berbagai macam variabel-variabel yang mengandung beragam informasi terkait dengan *series* data GSE. Untuk menampilkan deskripsi data yang diperlukan, peneliti hanya terfokus pada beberapa variabel terkait dari masing-masing *series* yang tersedia. Variabel tersebut dapat dilihat pada tabel berikut:

Tabel 5.1 Variabel *pheno* data

Data	Kode	Variabel	Keterangan
Kanker Payudara	GSE43837	`age:ch1`	Usia
		`disease state:ch1`	Status Penyakit
Kanker Pankreas	GSE16515	`age:ch1`	Usia
		`sex:ch1`	Jenis Kelamin
		`tissue:ch1`	Status Penyakit
AMI	GSE48060	`disease status:ch1`	Status Penyakit

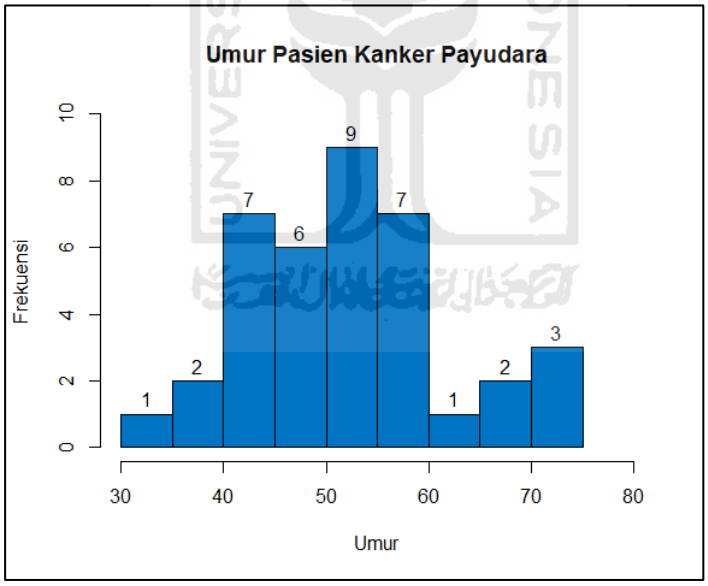
5.1.1. Kanker Payudara

Dalam set data kanker payudara terdapat dua variabel yang berasal dari *pheno* data yang digunakan untuk melihat deskripsi data, yaitu status penyakit (`disease state:ch1`) dan juga usia pasien (`age:ch1`).



Gambar 5.1 Histogram *diseases status*

Gambar di atas menunjukkan bahwa pada set data kanker payudara terdapat 38 sampel dengan dua kategori berbeda, 50% di antaranya memiliki penyakit *brain metastatic breast cancer* yaitu sebanyak 19 pasien dan 50% atau 19 pasien lainnya memiliki penyakit *non-metastatic breast cancer*.



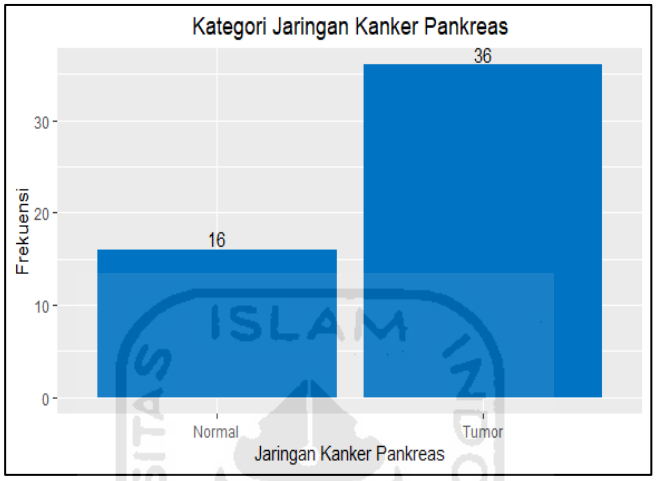
Gambar 5.2 Histogram variabel *pheno* umur pasien

Gambar 5.2 mendeskripsikan informasi terkait usia pasien yang menderita kanker payudara. Usia pasien yang menderita kanker payudara berkisar dari usia 32 tahun hingga 71 tahun. Pasien yang paling banyak terkena kanker payudara berada pada rentang usia 40-60 tahun yaitu sebanyak 29 pasien. Pasien termuda kanker pankreas pada usia 32 tahun dan tertua pada usia 73 tahun. Dari hasil

tersebut dapat diambil kesimpulan bahwa seseorang yang telah didiagnosis terkena kanker payudara dapat terjadi pada usia dewasa sampai dengan usia lanjut.

5.1.2. Kanker Pankreas

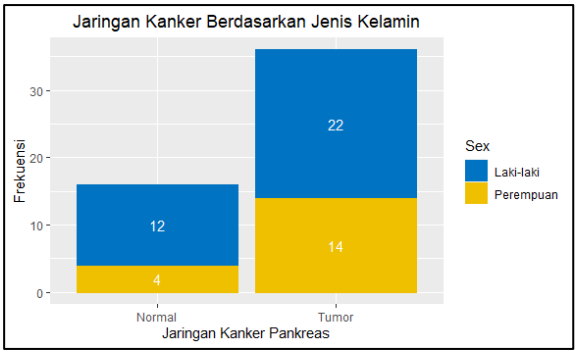
Pada set dara kanker pankreas terdapat tiga variabel *pheno* yang dapat diperhatikan yaitu variabel jenis jaringan kanker (`tissue:ch1`), jenis kelamin (`sex:ch1`) dan variabel usia (`age:ch1`).



Gambar 5.3 Histogram kategori jaringan kanker pankreas

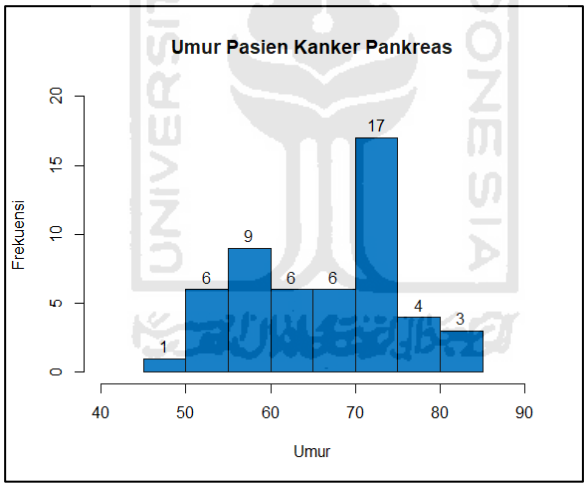
Dari Gambar 5.3 diperoleh informasi persebaran penyakit kanker pankreas pada jaringan tumor dan jaringan normal. Sebanyak 69% data merupakan jaringan tumor pada kanker pankreas dengan jumlah 36 sampel sedangkan 31% lainnya merupakan jaringan normal pada kanker pankreas dengan jumlah sampel sebanyak 16.

Pada Gambar 5.4 berdasarkan jenis kelamin pada pasien kanker pankreas lebih banyak dialami oleh pasien laki-laki, di mana jumlah pasien laki-laki sebanyak 34 lebih banyak dari pasien perempuan yang hanya berjumlah 18. Terdapat beberapa faktor penyebab terjadinya kanker pankreas, salah satunya adalah kebiasaan merokok yang lumrah dilakukan oleh laki-laki sehingga hal ini dapat diduga menjadi salah satu penyebab laki-laki mempunyai jumlah penderita kanker pankreas lebih banyak dari perempuan. Grafik tersebut juga menjelaskan bahwa dari 34 pasien laki-laki, 22 di antaranya merupakan pasien kanker pankreas yang mempunyai jaringan tumor sedangkan 12 pasien lainnya memiliki jaringan normal. Sedangkan pasien perempuan yang memiliki jaringan tumor sebanyak 14 pasien dan hanya 4 pasien yang memiliki jaringan normal.



Gambar 5.4 Kategori jaringan kanker berdasarkan jenis kelamin

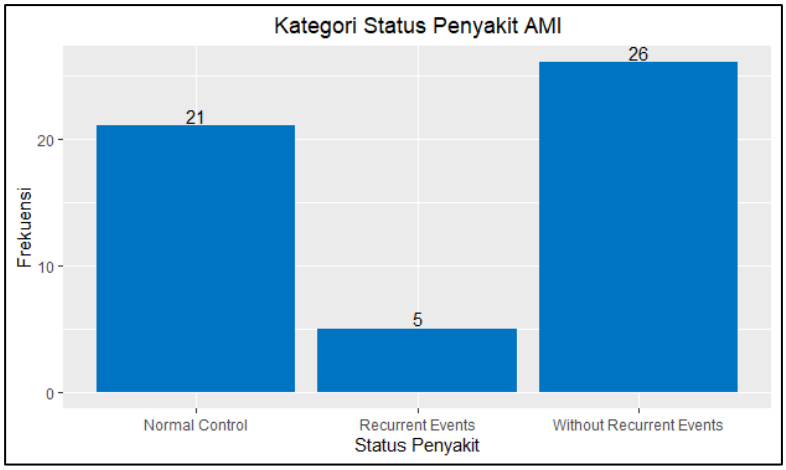
Pada Gambar 5.5. di bawah ini mendeskripsikan informasi terkait usia pasien yang menderita kanker pankreas. Usia pasien yang menderita kanker pankreas berkisar dari 49 tahun hingga 84 tahun. Pasien yang paling banyak terkena kanker pankreas berada pada rentang usia 70-75 tahun sebanyak 17 pasien. Pasien termuda kanker pankreas pada usia 49 tahun dan tertua pada usia 84 tahun. Berdasarkan faktor penyebab penyakit kanker pankreas adalah umur dapat terjadi di umur 50 tahun dan puncaknya pada dekade ke tujuh yaitu pada umur 70 tahun.



Gambar 5.5 Histogram usia pasien kanker pankreas

5.1.3. Penyakit AMI

Data Pasien AMI yang diamati pada penelitian ini terbagi menjadi tiga kelas, yaitu pasien normal, pasien dengan kejadian berulang dan pasien tanpa kejadian berulang. Pasien normal merupakan pasien yang tidak memiliki riwayat penyakit jantung dengan hasil *ekokardiogram* normal, sedangkan untuk pasien dengan kejadian berulang dan pasien tanpa kejadian berulang merupakan pasien yang sebelumnya terdaftar dalam program rehabilitasi jantung dan memiliki riwayat penyakit *kardiovaskular* (Suresh, et al., 2014). Adapun jumlah masing-masing kelas dapat dilihat pada gambar berikut:

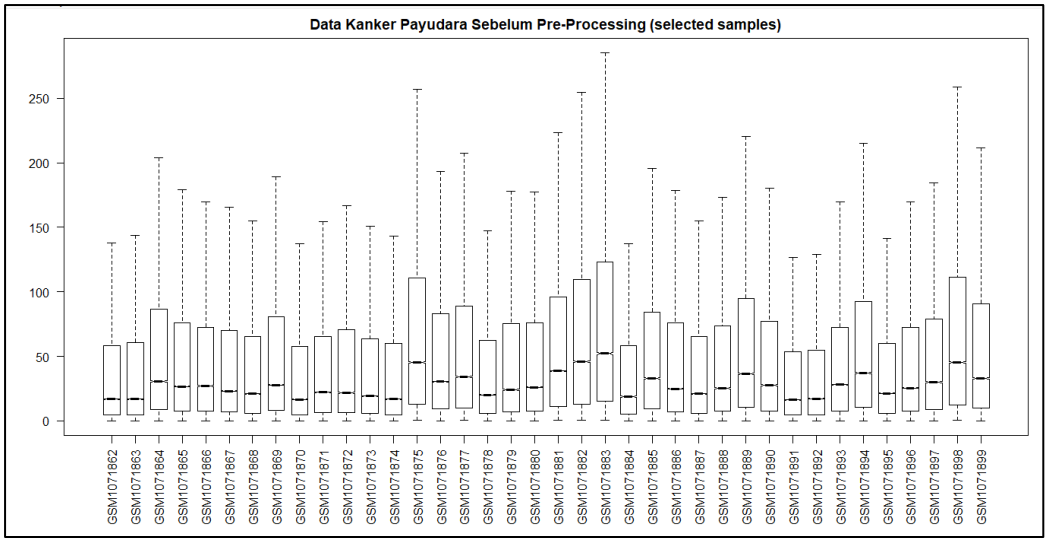


Gambar 5.6 *Bar Plot Status Disease*

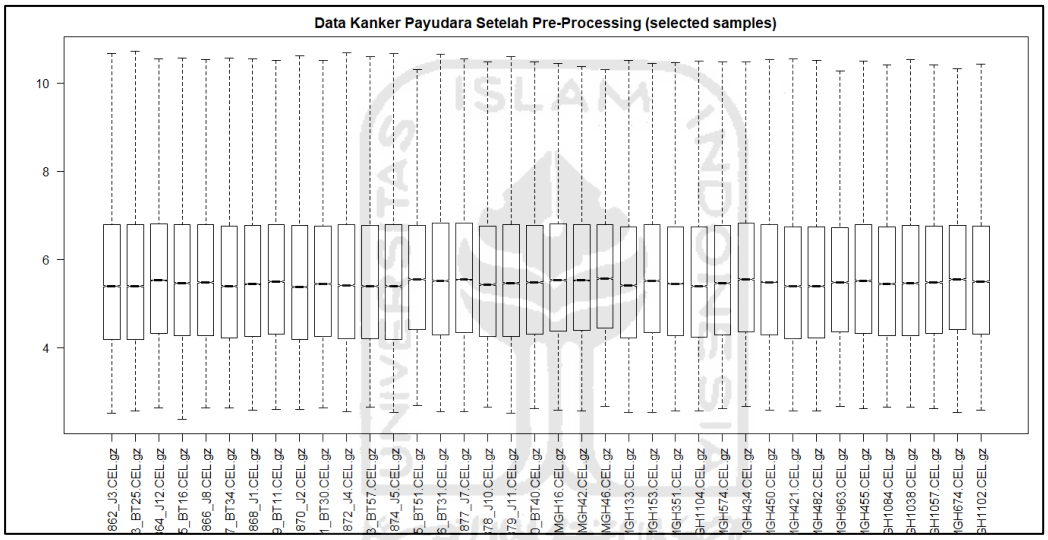
Pada gambar di atas terlihat bahwa set data penyakit AMI terbagi menjadi tiga kelompok atau kategori. Kelompok paling banyak adalah pasien dengan status penyakit tanpa kejadian berulang sebanyak 26 pasien atau 50% dari keseluruhan sampel. Kelompok paling sedikit adalah kelompok pasien dengan kejadian berulang sebanyak 5 pasien atau 9,61% dan selebihnya adalah pasien normal *control* sebanyak 21 pasien atau 40,38%.

5.2. *Pre-Processing*

Pada tahap ini, *pre-processing* dilakukan dengan menggunakan *package* `affyPLM` dengan *function* `threestep()`. Fungsi tersebut merupakan satu metode *pre-processing* yang di dalamnya terdapat tiga langkah *pre-processing* yaitu: penyesuaian sinyal/*background*, normalisasi lalu kemudian *summarization*. Hasil *output* pada tahap in dapat ditampikan menggunakan grafik *boxplot*. *Boxplot* adalah salah satu cara dalam statistika untuk menggambarkan data dalam bentuk grafik. Terdapat beberapa ukuran yang dapat diketahui dari *boxplot* di antaranya adalah kuartil pertama (Q1), median/kuartil tengah (Q2), kuartil ketiga (Q3) dan *interquartile range* (IQR). Nilai IQR menentukan panjang dari *box* pada *boxplot*, semakin tinggi atau lebar bidang IQR menandakan semakin menyebar data yang dimiliki.



Gambar 5.7 *Boxplot* data sebelum *pre-processing*



Gambar 5.8 *Boxplot* data setelah *pre-processing*

Gambar 5.7 dan Gambar 5.8 adalah hasil perbandingan data kanker payudara antara sebelum dan setelah melalui tahap *pre-processing* metode *three spteps* yang memperlihatkan perbedaan yang sangat signifikan. Hal ini terjadi karena pada tahap *pre-processing* terdapat metode normalisasi yang membuat data menjadi normal atau seragam. Tahap *pre-processing* ini juga diterapkan pada data kanker pankreas dan data penyakit AMI.

5.3. Reduksi Dimensi IPCA

Untuk mengurangi dimensi yang besar pada data *microarray* diperlukan sebuah teknik reduksi dimensi agar kesalahan dapat diminimalkan dan *running time* ketika proses klasifikasi menjadi lebih baik dan lebih efisien. Pada Tabel 4.1 terlihat bahwa jumlah fitur data ekspresi gen sebagai data masukan (*input*) sangat besar,

maka dari itu perlu dilakukan reduksi dimensi yang bertujuan untuk mengurangi jumlah fitur tanpa mengurangi karakteristik biologis yang terkandung dalam data secara signifikan. Pada penelitian ini, peneliti menggunakan teknik reduksi dimensi yang diusulkan oleh Yao et al. (2012) yaitu *Independent Principal Component Analysis* (IPCA).

Jumlah fitur yang dipilih dari hasil transformasi IPCA dilihat berdasarkan besaran nilai variansi yang mampu dijelaskan oleh masing-masing fitur. Pada Tabel 5.2 ditampilkan variansi yang mampu dijelaskan oleh tiap fitur hasil transformasi IPCA. Pada tabel tersebut terlihat bahwa jika hanya digunakan 1 fitur hasil IPCA, maka 1 fitur tersebut dapat menjelaskan 24,20% varians dari data awal pada set data kanker payudara. Jika digunakan 2 fitur hasil IPCA, maka 2 fitur tersebut dapat menjelaskan 33,62% variansi dari data awal begitu pun seterusnya. Hal ini juga berlaku untuk set data kanker pankreas dan penyakit AMI.

Tabel 5.2 Nilai *explained variance* tiap fitur IPC

Set Data	Jumlah Komponen	*Explained Variance* Tiap Fitur						Total
		Ke-1	Ke-2	Ke-3	Ke-4	⋯	Ke-30	
Kanker Payudara	IPC1	24,20%	-	-	-	⋯	-	24,20%
	IPC2	9,74%	23,88%	-	-	⋯	-	33,62%
	IPC3	6,47%	9,94%	22,53%	-	⋯	-	38,94%
	IPC5	4,52%	5,55%	13,00%	14,44%	⋯	-	47,03%
	IPC10	4,76%	6,24%	4,74%	9,59%	⋯	-	60,35%
	IPC15	5,68%	6,64%	2,69%	4,87%	⋯	-	70,64%
	IPC30	5,01%	6,31%	3,57%	1,87%	⋯	2,79%	93,38%
Kanker Pankreas	IPC1	30,58%	-	-	-	⋯	-	30,58%
	IPC2	23,17%	18,51%	-	-	⋯	-	41,68%
	IPC3	24,11%	14,84%	9,08%	-	⋯	-	48,03%
	IPC5	19,79%	7,21%	7,64%	13,32%	⋯	-	57,09%
	IPC10	17,77%	8,07%	7,72%	13,41%	⋯	-	70,65%
	IPC15	18,19%	3,05%	8,55%	5,87%	⋯	-	78,53%
	IPC30	17,55%	3,14%	4,00%	5,10%	⋯	4,10%	91,26%
Penyakit AMI	IPC1	31,54%	-	-	-	⋯	-	31,54%
	IPC2	12,34%	30,73%	-	-	⋯	-	43,07%
	IPC3	10,85%	8,70%	28,31%	-	⋯	-	47,86%
	IPC5	6,43%	12,57%	11,27%	8,84%	⋯	-	55,82%
	IPC10	6,40%	5,32%	13,75%	3,21%	⋯	-	67,02%
	IPC15	6,44%	2,84%	3,76%	13,83%	⋯	-	73,92%
	IPC30	6,36%	3,14%	3,79%	14,12%	⋯	1,91%	87,17%

Tabel di atas memperlihatkan bahwa semakin banyak jumlah komponen yang digunakan maka semakin tinggi pula nilai persentase variansi yang mampu dijelaskan. Selain itu juga terlihat bahwa dari sekian banyak jumlah fitur data awal (lihat Tabel 4.1), ketika ditransformasikan menggunakan IPCA, cukup dengan 30 fitur telah memberikan varians total rata-rata dari ketiga set data didapatkan nilai

sebesar 90,60%. Artinya, IPCA dalam kasus kanker payudara, kanker pankreas dan penyakit AMI mampu mengurangi jumlah fitur yang diperlukan atau mereduksi dimensi secara signifikan untuk proses klasifikasi.

Dengan demikian, didapatkan hasil bahwa dengan menggunakan 10 atau lebih komponen IPC, maka didapatkan hasil variansi yang mampu dijelaskan lebih dari 60%. Pada set data kanker payudara senilai 60,35%, kanker pankreas 70,65% dan pada set data penyakit AMI senilai 67,02%. Berdasarkan hasil yang diperoleh maka dapat disimpulkan bahwa 10 komponen IPC sudah tergolong baik dalam merepresentasikan data asal dan sudah layak untuk dijadikan data *input* pada tahap klasifikasi. Dengan demikian untuk tahap klasifikasi, peneliti akan menggunakan IPC10, IPC15 dan IPC30 sebagai perbandingan untuk melihat apakah jumlah komponen berpengaruh terhadap hasil klasifikasi dari kedua metode yang akan diterapkan.

5.4. Pembagian Data *Training* dan *Testing*

Untuk melakukan proses klasifikasi, data hasil reduksi dimensi terlebih dahulu dibagi menjadi dua bagian yaitu data *training* (latih) dan data *testing* (uji). Pembagian data *training* dan *testing* dibagi dengan rasio 80% untuk data *training* dan 20% untuk data *testing* dari total sampel data yang siap dianalisis untuk masing-masing kelas. Adapun pembagian dari masing-masing set data dapat dilihat pada tabel berikut:

Tabel 5.3 Pembagian data *training* dan data *testing*

Set Data	*Training*	*Testing*
Kanker Payudara	30	8
Kanker Pankreas	41	11
Penyakit AMI	41	11

5.5. Klasifikasi

Pada bagian ini akan dibahas mengenai hasil klasifikasi terhadap tiga set data yang digunakan. Ketiga set data tersebut telah melalui proses reduksi dimensi menggunakan metode IPCA. Hasil transformasi IPCA akan dijadikan data *input* pada proses klasifikasi menggunakan metode SVM dan NBC. Data yang akan diklasifikasi sebelumnya sudah dibagi menjadi dua bagian yaitu data *training* dan data *testing*. Adapun jumlah komponen IPC yang digunakan berbeda-beda

dimaksudkan untuk melihat apakah terdapat berbedaan hasil akurasi klasifikasi berdasarkan jumlah komponen.

5.5.1. Hasil Klasifikasi *Support Vector Machine*

Pada Klasifikasi SVM dilakukan percobaan menggunakan beberapa *kernel* yang ada dalam metode klasifikasi SVM. *Kernel* yang dimaksudkan ada tiga yaitu *kernel linear*, *kernel polynomial* dan *kernel Radial Basis Function* (RBF). Ketiga metode *kernel* SVM tersebut diterapkan terhadap data hasil reduksi dimensi dengan jumlah komponen yang terdiri dari 10, 15 dan 30. Tabel berikut ini adalah hasil perbandingan akurasi dari ketiga *kernel* SVM yang diterapkan.

Tabel 5.4 Perbandingan akurasi tiap *kernel* SVM

Data	Jumlah Komponen	Klasifikasi SVM		
		Linear	*Polynomial*	RBF
Kanker Payudara	IPC10	50,00%	100,00%	75,00%
	IPC15	37,50%	37,50%	62,50%
	IPC30	75,00%	37,50%	62,50%
Kanker Pankreas	IPC10	54,55%	72,73%	90,91%
	IPC15	54,55%	72,73%	90,91%
	IPC30	54,55%	54,55%	54,55%
Penyakit AMI	IPC10	36,36%	45,45%	81,82%
	IPC15	36,36%	54,55%	72,73%
	IPC30	54,55%	36,36%	72,73%
Rata-rata		50,38%	56,82%	73,74%

Berdasarkan perbandingan menggunakan beberapa *kernel* pada Tabel 5.4 menunjukkan bahwa *kernel* dengan akurasi terbaik didapatkan pada *kernel* RBF. Hal ini dapat dilihat dari nilai rata-rata akurasi yang dihasilkan oleh *kernel* RBF yaitu sebesar 73,74% lebih tinggi dibanding dua *kernel* lainnya yaitu *kernel polynomial* dengan rata-rata akurasi sebesar 56.82% dan *kernel linear* yang memiliki nilai rata-rata akurasi hanya sebesar 50,38%. Dengan demikian peneliti memutuskan untuk menggunakan hasil klasifikasi SVM *kernel* RBF sebagai bahan perbandingan yang akan dibandingkan dengan metode klasifikasi NBC.

Pada tabel di atas juga terlihat bahwa data yang digunakan dengan jumlah komponen IPC10 merupakan data yang dapat diklasifikasikan dengan baik oleh ketiga *kernel* SVM. Bahkan nilai akurasi klasifikasi pada *kernel polynomial* terhadap IPC10 memiliki nilai akurasi tertinggi yaitu 100%. Hal ini menunjukkan bahwa banyaknya jumlah komponen IPC yang dipilih tidak menjamin hasil akurasi

klasifikasi yang tinggi. Bahkan hasil akurasi terendah dihasilkan dari data dengan jumlah IPC sebanyak 30 yaitu sebesar 36,36% yang diklasifikasikan menggunakan *kernel polynomial*.

5.5.2. Hasil Klasifikasi *Naive Bayes Classifier*

Analisis NBC dilakukan setelah menentukan data *testing* dan data *training*, pembagian data *testing* dan *training* dapat dilihat pada Tabel 5.3. Analisis NBC menggunakan nilai peluang bersyarat dalam menentukan kelasnya. Oleh karena itu, klasifikasi pada analisis NBC sebelumnya harus memiliki *prior probabilities*. *Prior probabilities* adalah komponen utama pada konsep Naive Bayes. Nilai *prior probabilities* dapat dilihat pada Tabel 5.8.

Tabel 5.5 Nilai *Prior Probabilities*

Data	Kelas 0	Kelas 1	Kelas 2
Kanker Payudara	16 (53,3%)	14 (46,7%)	-
Kanker Pankreas	30 (73,2%)	11 (26,8%)	-
Penyakit AMI	4 (9,8%)	22 (53,7%)	15 (36,6%)

Prior probabilities merupakan tahapan untuk mencari nilai probabilitas pada masing-masing pengamatan yang akan menghasilkan klasifikasi, juga dapat dikatakan bahwa nilai *prior* adalah nilai suatu peluang kejadian. Nilai tersebut merupakan nilai proporsi masing-masing kelas pada data *training*. Nilai tersebut didapatkan dari banyaknya jumlah pada kategori kelas 'positif' (contohnya kelas *Brain Metastatic* pada data kanker payudara) dibagi dengan total dari data *training* begitu pun dengan kategori kelas lainnya. Untuk itu, nilai *prior* ini bertujuan untuk menentukan besarnya peluang atau probabilitas dari masing-masing kategori kelas.

Tabel 5.6 Nilai akurasi klasifikasi metode NBC

Data	Jumlah Komponen	Klasifikasi NBC
Kanker Payudara	IPC10	87,50%
	IPC15	62,50%
	IPC30	75,00%
Kanker Pankreas	IPC10	90,91%
	IPC15	81,82%
	IPC30	72,73%
Penyakit AMI	IPC10	63,64%
	IPC15	54,55%
	IPC30	72,73%
Rata-rata		73,48%

Adapun hasil akurasi yang didapatkan dari metode klasifikasi NBC terhadap data kanker payudara, kanker pankreas dan penyakit AMI diperlihatkan pada Tabel 5.6. Pada tabel tersebut terlihat bahwa nilai rata-rata dari keseluruhan hasil akurasi klasifikasi NBC adalah sebesar 73,48%. Akurasi terbaik dihasilkan dari pengklasifikasian data kanker pankreas dengan jumlah fitur IPC sebanyak 10 dengan nilai akurasi sebesar 90,91%, sedangkan hasil akurasi terkecil yaitu 54,55% yang merupakan data penyakit AMI dengan jumlah fitur IPC sebanyak 15. Sama dengan hasil akurasi SVM yang dijelaskan pada bagian sebelumnya, pada bagian ini IPC10 juga mendominasi tingkat akurasi hasil klasifikasi NBC. Pada data kanker payudara dan data kanker pankreas masing-masing memiliki nilai akurasi sebesar 87,50% dan 90,91%.

5.6. Hasil Evaluasi Klasifikasi

Pada bagian ini akan dibandingkan hasil akurasi dari tiap metode klasifikasi yang digunakan terhadap data kanker payudara, kanker pankreas dan data AMI berdasarkan data hasil reduksi dimensi dengan jumlah komponen sebanyak 10 fitur (IPC10). Hal ini dilakukan berdasarkan penjelasan pada bagian-bagian sebelumnya yang menyimpulkan bahwa IPC10 merupakan data yang mampu diklasifikasikan dengan baik dibandingkan dengan IPC15 dan IPC30.

Tabel 5.7 Hasil akurasi IPC10

Classifier	Kanker Payudara	Kanker Pankreas	AMI
SVM (RBF)	75,00%	90,91%	81,82%
NBC	87,50%	90,91%	63,64%

Tabel 5.7 memperlihatkan bahwa data kanker payudara diklasifikasikan dengan baik oleh metode NBC dengan nilai akurasi sebesar 87,50% lebih besar dibandingkan dengan metode SVM *kernel* RBF yang hanya menghasilkan tingkat akurasi sebesar 75,00%. Berbanding terbalik dengan data kanker payudara, data AMI justru diklasifikasikan dengan baik oleh metode SVM dengan tingkat akurasi 81,82% sedangkan oleh metode NBC hanya 63,64% yang merupakan tingkat akurasi terendah dari seluruh data untuk IPC10. Berdasarkan pada kedua hasil tersebut dapat disimpulkan bahwa nilai akurasi tidak hanya tergantung dari berapa jumlah IPC yang digunakan tapi juga tergantung pada jenis data itu sendiri. Suatu

metode klasifikasi yang memiliki kinerja yang baik pada satu set data belum tentu memiliki kinerja baik terhadap data yang lain.

Sedangkan untuk data kanker pankreas dapat diklasifikasikan dengan baik oleh kedua metode dengan tingkat akurasi yang dihasilkan sama yaitu 90,91% yang merupakan tingkat akurasi tertinggi. Dengan demikian, sulit untuk menentukan metode terbaik jika hanya melihat dari nilai hasil akurasi. Untuk itu, peneliti akan melihat nilai evaluasi lain seperti nilai *accuracy, precision*, *recall*, *f-score, specificity* dan nilai AUC pada *confusion matrix* hasil klasifikasi SVM *kernel* RBF dan NBC terhadap IPC10 data kanker pankreas.

Tabel 5.8 *Confusion matrix* SVM *kernel* RBF dan NBC terhadap IPC10 kanker pankreas

Data Kanker Pankreas		**Testing**			
		SVM *kernel* RBF		**NBC**	
		Tumor	**Normal**	**Tumor**	**Normal**
Prediksi	**Tumor**	6	1	5	0
	Normal	0	4	1	5

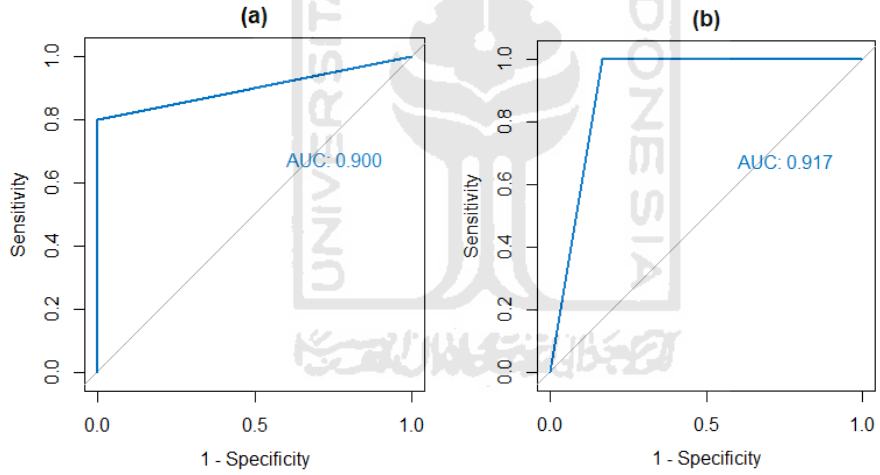
Tabel di atas adalah *Confusion matrix* yang dihasilkan dari metode SVM *kernel* RBF dan NBC dalam melakukan klasifikasi terhadap data IPC10 kanker pankreas. Kedua klasifikasi tersebut menghasilkan nilai akurasi yang sama yaitu 90,91%. Untuk itu, berikut ini akan dihitung ukuran kinerja lainnya yang bertujuan untuk mengetahui metode mana yang memiliki kinerja terbaik untuk data kanker pankreas. Hasil perhitungan tersebut ditampilkan pada tabel berikut:

Tabel 5.9 Hasil perhitungan ukuran kinerja klasifikasi

	K. Pankreas	
	SVM *kernel* RBF	NBC
Accuracy	90,91%	90,91%
Precision	85,71%	100,00%
Recall	100,00%	83,33%
F-Score	92,31%	90,91%
Spesificity	80,00%	100,00%
FPR	20,00%	0,00%
AUC	90,00%	91,67%
Error	9,09%	9,09%

Nilai *accuracy* dan nilai *error* yang dihasilkan dari kedua metode klasifikasi bernilai sama. Untuk mengetahui model klasifikasi terbaik maka peneliti akan

membandingkan nilai *f-score* dan nilai AUC. Nilai *f-score* pada metode SVM sebesar 92,31% lebih tinggi dibanding nilai *f-score* pada klasifikasi metode NBC yaitu 90,91%. Hal ini menunjukkan bahwa *precision* dan *recall* atau *sensitivity* pada klasifikasi SVM lebih seimbang dibandingkan pada klasifikasi NBC. Dengan demikian berdasarkan nilai *f-score*, klasifikasi SVM *kernel* RBF lebih baik dibandingkan klasifikasi NBC. Akan tetapi jika dilihat dari nilai AUC, maka klasifikasi NBC lebih baik dibandingkan SVM. Hal ini disimpulkan berdasarkan kriteria penilaian AUC (lihat Tabel 3.2). Klasifikasi NBC mempunyai nilai AUC sebesar 91,67% atau 0,916 yang merupakan *excellent classifier*, sedangkan nilai AUC dari klasifikasi SVM hanya 90% atau 0,9 yang merupakan *good classifier*. Berdasarkan perbandingan tersebut maka dapat disimpulkan bahwa klasifikasi NBC dan SVM *kernel* RBF sama-sama memiliki kinerja yang baik dalam melakukan klasifikasi terhadap data ekspresi gen kanker pankreas. Nilai AUC dapat dilihat pada gambar kurva ROC berikut.



Gambar 5.9 ROC *Curve* Klasifikasi SVM *kernel* RBF (a) dan NBC (b) berdasarkan data IPC10 kanker pankreas

BAB VI
PENUTUP

6.1. Kesimpulan

Berdasarkan hasil tahapan analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut:

1) Berdasarkan hasil analisis deskriptif, pasien yang paling banyak terkena kanker payudara berada pada rentang usia 40-60 tahun atau terjadi pada usia dewasa sampai dengan usia lanjut. Sedangkan pada data pasien kanker pankreas didapatkan bahwa laki-laki mempunyai jumlah penderita kanker pankreas lebih banyak dari perempuan dan banyak terjadi pada pasien umur 50 tahun ke atas.
2) Metode *Independent Principal Component Analysis* (IPCA) dapat mereduksi dimensi data ekspresi gen dengan baik tanpa mengurangi karakteristik biologis yang terkandung dalam data meskipun jumlah fitur menjadi lebih sedikit. Pemilihan jumlah fitur IPC dapat dilakukan menggunakan besaran nilai *explained variance* dan semakin banyaknya jumlah fitur IPC yang dipilih tidak menjamin akan memperoleh hasil akurasi klasifikasi yang tinggi.
3) Metode klasifikasi *Support Vector Machine* (SVM) dan *Naive Bayes Classifier* (NBC) mampu memperoleh hasil klasifikasi yang baik ketika diterapkan pada data yang telah melalui proses reduksi dimensi menggunakan IPCA. Pada metode SVM, *kernel* yang memperoleh akurasi yang baik adalah *kernel Radial Basis Function* (RBF).
4) Pada hasil klasifikasi didapatkan bahwa data ekspresi gen kanker payudara diklasifikasikan dengan baik oleh metode NBC dengan nilai akurasi 87,50% lebih tinggi dibandingkan dengan metode SVM dengan tingkat akurasi 75,00%. Berbeda dengan data ekspresi gen penyakit AMI yang justru dapat diklasifikasikan dengan baik oleh metode SVM dengan tingkat akurasi 81,82% lebih tinggi dibandingkan dengan metode NBC dengan tingkat akurasi 63,64%. Sedangkan untuk data ekspresi gen kanker pankreas dapat diklasifikasikan dengan baik oleh kedua metode dengan tingkat akurasi yang dihasilkan bernilai sama yaitu 90,91% yang merupakan tingkat akurasi tertinggi. Klasifikasi NBC dan SVM *kernel* RBF sama-sama memiliki kinerja

yang baik dalam melakukan klasifikasi terhadap data ekspresi gen kanker pankreas. Suatu metode klasifikasi yang memiliki kinerja yang baik pada satu set data belum tentu memiliki kinerja baik terhadap data yang lain.

6.2. Saran

Adapun saran pada penelitian ini sebagai upaya perbaikan dan pengembangan atau penelitian lanjutan adalah sebagai berikut:

1) Penelitian selanjutnya dapat menggunakan set data ekspresi gen lain untuk melihat efektivitas metode *Independent Principal Component Analysis* (IPCA) dalam melakukan reduksi dimensi terhadap data ekspresi gen.
2) Penelitian selanjutnya menggunakan metode klasifikasi lain sebagai bahan perbandingan terhadap kinerja IPCA dalam melakukan reduksi dimensi pada data ekspresi gen sebelum diklasifikasi.
3) Mencari kombinasi metode IPCA dan metode klasifikasi yang tepat dalam melakukan reduksi dimensi untuk mendapatkan hasil akurasi yang lebih maksimal dengan nilai *error* yang relatif kecil.

DAFTAR PUSTAKA

Adiwijaya. (2018). Deteksi kanker berdasarkan klasifikasi microarray data. *Media Informatika Budidarma,* 2(4), 181-186. Doi:10.30865/mib.v2i4.1043.

Ammu, P. K., & Preeja, V. (2013). Review on feature selection techniques of DNA microarray data. *International Journal of Computer Applications,* 61(12), 39-44. Doi:10.5120/9983-4814.

An, Y., Sun, S., & Wang, S. (2017). Naive Bayes classifiers for music emotion classification based on lyrics. *IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS),* 1, 635-638. Doi:10.1109/ICIS.2017.7960070.

Astuti, W., & Adiwijaya. (2019). Principal Component Analysis sebagai ekstraksi fitur data microarray untuk deteksi kanker berbasis linear discriminant analysis. *Jurnal Media Informatika Budidarma,* 3(2), 72-77. Doi:10.30865/mib.v3i2.1161.

Awad, M., & Khanna, R. (2015). Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress. Doi:10.1007/978-1-4302-5990-9.

Bach, F. R., & Jordan, M. I. (2002). Kernel Independent Component Analysis. *Journal of Machine Learning Research,* 3, 1-48. Doi:10.1162/153244303768966085.

Bayat, A. (2002). Bioinformatics: Science, medicine, and the future. *BMJ,* 324(7344), 1018-1022. Doi:10.1136/bmj.324.7344.1018.

Berg, D. D. et al. (2018). Modes and timing of death in 66252 patients with non-ST-segment elevation acute coronary syndromes enrolled in 14 TIMI trials. *European Heart Journal,* 39(42), 3810-3820. Doi:10.1093/eurheartj/ehy556.

Bolstad, B. M. (2004). Low-level analysis of high-density oligonucleotide array data: Background, Normalization and Summarization (Doctoral Dissertation, University of California, 2004). Diakses dari https://bmbolstad.com/Dissertation/Bolstad_2004_Dissertation.pdf.

Boluwade, A., & Madramootoo, C. A. (2016). Independent principal component analysis for simulation of soil water content and bulk density in a Canadian Watershed. *International Soil and Water Conservation Research,* 4(3), 151-158. Doi:10.1016/j.iswcr.2016.09.001.

Bramer, M. (2013). Principles of Data Mining (2nd Ed.). London: Springer. Doi:10.1007/978-1-4471-4884-5.

Campbell, N. A., Reece, J. B., & Mitchell, L. G. (2002). Biologi Jilid 1 (Ed. 5). Jakarta: Erlangga.

CMScripst. (2017, April 24). Pancreatic Cancer Issue 9. *Council for Medical Scheme (CMS) Newaletters*. Diakses dari https://www.topmed.co.za/.

Comon, P. (1994). Signal Processing: Independent component analysis, A new concept?. *Signal Processing,* 36(3), 287-314. Doi:10.1016/0165-1684(94)90029-9.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20, 273–297. Doi:10.1007/BF00994018.

Dale, J. W., & Park, S. F. (2004). Molecular Genetics of Bacteria (4th Ed.). Chichester: John Wiley & Sons Ltd.

Defiyanti, S., & Jajuli, M. (2015). Integrasi metode klasifikasi dan clustering dalam data mining. *Konferensi Nasional Informatika (KNIF)*, (hal. 39-44). Diakses dari https://www.researchgate.net/.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using Naïve Bayes‘ and K-NN classifier. *International Jurnal of Information Engineering and Electronic Business,* 8(4), 54-62. Doi:10.5815/ijieeb.2016.04.07.

Dufva, M. (2009). Introduction to microarray technology. Dalam *DNA Microarrays for Biomedical Research* (Vol. 529, hal. 1-22). Humana Press. Doi:10.1007/978-1-59745-538-1.

Eng, J. (2005). Receiver operating characteristic analysis: a primer. *Academic Radiology,* 12(7), 909-916. Doi:10.1016/j.acra.2005.04.005.

Engreitz, J. M., Daigle Jr., B. j., Marshall, J. J., & Altman, R. B. (2010). Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics,* 43(6), 932–944. Doi:10.1016/j.jbi.2010.07.001.

Eveline, L., Purwanto, H., & Lestari, P. (2017). Faktor klinis dan histopatologi serta hubungannya dengan kekambuhan pasca-operasi pada pasien kanker payudara di RSUD Dr. Soetomo, Januari-Juni 2015. *Indonesian Journal of Cancer,* 11(2), 55-60. Doi:10.33371/ijoc.v11i2.502.

Fatchiyah. (2008). Pengantar bioinformatika kedokteran. Malang: Universitas Brawijaya.

Globocan. (2019, Mei). Diakses dari The Global Cancer Observatory (GCO): https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf.

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Berlin: Springer. Doi: 10.1007/978-3-642-19721-5.

Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy and Bioallied Sciences,* 4, 310-312. Doi:10.4103/0975-7406.100283.

Hall, J., & Guyton, A. C. (2006). Buku ajar fisiologi kedokteran (Edisi ke-11). (Irawati, D. Ramdani, & F. Indriyani, Penerj.) Jakarta: Penerbit Buku Kedokteran EGC.

Han, J., & Kamber, M. (2006). Data mining: concepts and techniques (2nd ed.). San Francisco: Elsevier Inc.

Hania, A. A. (2017). Mengenal artificial intelligence, machine learning, neural network, dan deep learning. *Jurnal Teknologi Indonesia*. Diakses dari https://www.researchgate.net/.

Harianto, Mutiara, R., & Surachmat, H. (2005). Risiko penggunaan pil kontrasepsi kombinasi terhadap kejadian kanker payudara pada reseptor KB di Perjan RS. Dr. Cipto Mangunkusumo. *Majalah Ilmu Kefarmasian,* 2(1), 84-99. Doi:10.7454/psr.v2i2.3386.

Harteminka, A. J., Gifford, D. K., Jaakkola, T. S., & Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE - The International Society for Optical Engineering,* 4266. Doi:10.1117/12.427981.

Hastie, T., & Tibshirani, R. (2002). Independent Components Analysis through Product Density Estimation . *Conference of Advances in Neural Information Processing Systems 15: Neural Information Processing*

Systems, NIPS 2002, 9-14 December 2002. Vancouver, British Columbia, Canada.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Stanford: Springer.

Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012*, 241-249. Diakses dari http://eprints.dinus.ac.id.

Hidayatullah, A. F., Prasetyo, A. D., Sari, D. P., & Pratiwi, I. (2014). Analisis kualitas data dan klasifikasi data pasien kanker. *Seminar Nasional Informatika Medis (SNIMed) V 2014, Yogyakarta*. Diakses dari https://journal.uii.ac.id/.

Himberg, J., Hyvarinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage,* 22(3), 1214-1222. Doi:10.1016/j.neuroimage.2004.03.027.

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics,* 2015. Doi:10.1155/2015/198363.

(http://mixomics.qfab.org)

Huang, D.-S., & Zheng, C.-H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics,* 22(15), 1855–1862. Doi:10.1093/bioinformatics/btl190.

Humaera, R., & Mustofa, S. (2017). Diagnosis dan penatalaksanaan karsinoma mammae stadium 2. *Journal Medula Unila,* 7(2), 103-107. Diakses dari https://juke.kedokteran.unila.ac.id/.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks,* 13(4-5), 411-430. Doi:10.1016/s0893-6080(00)00026-5.

Irizarry, R., & Love, M. (2013). Downloading data from GEO using GEOquery. Diakses Agustus 27, 2020, dari http://genomicsclass.github.io/book/pages/GEOquery.html.

Jabbar, M. A., & Samreen, S. (2016). Heart disease prediction system based on hidden naïve bayes classifier. *2016 International Conference on Circuits,*

Controls, Communications and Computing (I4C), 1-5. Doi:10.1109/CIMCA.2016.8053261.

Jakkula, V. (2006). Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*. 37. Diakes dari https://www.semanticscholar.org/.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). New York: Springer.

Kementerian Kesehatan RI. (2019, Januari 31). Hari kanker sedunia 2019. Diakses dari KemKes: https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html.

Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., & Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. *Biotechniques,* 45(5), 501-520. Doi:10.2144/000112950.

Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) Curve for medical researchers. *Indian Pediatrics,* 48, 277-287. Doi:10.1007/s13312-011-0055-4.

Lee, S., & Batzoglou, S. (2003). Application of independent component analysis to microarrays. *Genome Biology,* 4(11). Doi:Doi.org/10.1186/gb-2003-4-11-r76.

Lehninger, A. L. (1994). Dasar-Dasar Biokimia (Jilid 3). (M. Thenawijaya, Penerj.) Jakarta: Erlangga.

Lu, H. H.-S., Schölkopf, B., & Zhao, H. (2011). Handbook of Statistical Bioinformatics. Berlin: Springer.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? an introduction and overview. *Yearbook of medical informatics,* 10(1), 83-100. Doi:10.1055/s-0038-1638103.

Ma'ruf, F. A., Adiwijaya, & Wisesty, U. N. (2018). Analisis pengaruh metode reduksi dimensi minimum redundancy maximum relevancepada klasifikasi kanker berdasarkan data microarray menggunakan classifier support vector machine. *e-Proceeding of Engineering,* 5(1), 1499-1506. Diakses dari https://libraryeproceeding.telkomuniversity.ac.id/.

Ma'wa, S., Adiwijaya, & Rohmawati, A. (2019). Klasifikasi k-nearest neighbor untuk data microarray dengan seleksi genetic algorithm. *e-Proceeding of Engineering,* 6(2), 9838-9847. Diakses dari https://libraryeproceeding.telkomuniversity.ac.id/.

Madigan, M. T., Martinko, J. M., Dunlap, P. V., & Clark, D. P. (2008). Brock Biology of Microorganisms (12th ed.). Benjamin Cummings.

Manliguez, G. (2016). Generalized confusion matrix for multiple classes. Doi:10.13140/RG.2.2.31150.51523.

Marlina, L., Muslim, & Siahaan, A. U. (2016). Data mining classification comparison (naïve bayes and c4.5 algorithms). *International Journal of Engineering Trends and Technology,* 38(7), 380-383. Doi:10.14445/22315381/IJETT-V38P268.

Massberg, S., & Polzin, A. (2018). Update ESC-Guideline 2017: Dual antiplatelet therapy. *Deutsche Medizinische Wochenschrift,* 143(15), 1090-1093. Doi:10.1055/a-0549-8230.

Miller, J. H., Griffiths, A. J., Lewontin, R. C., Suzuki, D. T., & Gelbart, W. M. (2002). An introduction to genetic analysis (7th ed.). New York: W. H. Freema.

Mulyanto, A. (2009). Sistem informasi konsep & aplikasi. Yogyakarta: Pustaka Pelajar.

Munzir, A. F., Adiwijaya, & Aditsania, A. (2018). Analsis reduksi dimensi pada klasifikasi microarray menggunakan MBP powell beale. *E-Jurnal Matematika,* 7(1), 17-24. Doi:10.24843/MTK.2018.v07.i01.p179.

Nascimento, B. R., Brant, L. C., Marino, B. C., Passaglia, L. G., & Ribeiro, A. P. (2019). Implementing myocardial infarction systems of care in low/middle-income countries. *Heart,* 105(1), 20-26. Doi:10.1136/heartjnl-2018-313398.

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Support vector machine: Teori dan aplikasinya dalam bioinformatika. *IlmuKomputer.Com*. Diakses dari http://asnugroho.net/.

Nurfalah, A., Adiwijaya, & Suryani, A. A. (2016). Cancer detection based on microarray data classification using PCA and modified back propagation. *Far East Journal of Electronics and Communications ,* 16(2), 269-281. Doi:10.17654/EC016020269.

Octaviani, P. A., Wilandari, Y., & Ispriyanti, D. (2014). Penerapan metode klasifikasi support vector machine (SVM) pada data akreditasi sekolah dasar (SD) di kabupaten magelang. *Jurnal Gaussian,* 3(4), 811-820. Doi:10.14710/j.gauss.v3i4.8092.

Palladino, M. A., & Thieman, W. J. (2013). Introduction to biotechnology (3rd ed.). Pearson.

Prasetyo, E. (2012). Data mining konsep dan aplikasi menggunakan matlab. Yogyakarta: Andi.

Primandari, A. H. (2018, Maret 11). Membuat fungsi dalam R. Diakses dari RPubs: https://rpubs.com/arumprimandari/368914.

Primartha, R. (2018). *Belajar: Machine learning: Teori dan praktik.* Bandung: Informatika.

Rana, J. S., & Vaisla, K. S. (2012). Introduction to bioinformatics. Dalam J. S. Rana, K. S. Vaisla, & R. Bharti, *BIOINFORMATICS - Tools & Applications* (hal. 158-175). Dehradun: USBD.

Raza, K. (2010). Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering,* 1(2), 114-118 . Diakses dari https://www.researchgate.net/.

Rong Yang, B. S. (2006). Comparison of normalization methods in microarray analysis (Doctoral Thesis, McMaster University, 2006). Diakses dari https://pdfs.semanticscholar.org/fa22/b470810aac7475e62da930825a957cc053fc.pdf.

Rosai, J. (2011). Rosai and Ackerman's surgical pathology (10th ed.). Elsevier.

Rostami, R., Mittal, S., Rostami, P., Tavassoli, F., & Jabbari, B. (2016). Brain metastasis in breast cancer: a comprehensive literature review. *Journal of Neuro-Oncology,* 127(3), 407-414. Doi:10.1007/s11060-016-2075-3.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development,* 3(3), 210 - 229. Doi:10.1147/rd.33.0210.

Santosa, B. (2007). Data mining: teknik pemanfaatan data untuk keperluan bisnis. Yogyakarta: Graha Ilmu.

Santoso, M., & Setiawan, T. (2005). Penyakit jantung koroner. *Cermin Dunia Kedokteran,* 147, 6-9.

Sanz, R. G., & Sánchez, A. (2018). Introduction to microarrays technology and data analysis. Dalam J. Jaumot, C. Bedia, & R. Tauler, *Comprehensive Analytical Chemistry* (Vol. 82, hal. 37-69). Amsterdam: Elsevier. Doi:10.1016/bs.coac.2018.08.002.

Sasongko, T. B. (2016). Komparasi dan analisis kinerja model algoritma SVM dan PSO-SVM (studi kasus klasifikasi jalur minat SMA). *Jurnal Teknik Informatika dan Sistem Informasi,* 2(2). Doi:10.28932/jutisi.v2i2.476.

Scheen, A. J. (2018). From atherosclerosis to atherothrombosis : from a silent chronic pathology to an acute critical event. *Revue Medicale de Liege,* 73(5-6), 224-228. Diakses dari https://pubmed.ncbi.nlm.nih.gov/.

Scholz, M., Sterling, A. R., Gatzek, S., Fiehn, O., & Selbig, J. (2004). Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics,* 20(15), 2447–2454. Doi:10.1093/bioinformatics/bth270.

Serin, A. (2011). Biclustering analysis for large scale (Doctoral Dissertation). Tersedia dari Refubium - Freie Universitat Brlin Repository. Doi:10.17169/refubium-17150.

Silalahi, L., Rahmi, E., & Sutarni, S. (2018). Tumor otak metastasis dari payudara. *Berkala Ilmiah Kedokteran Duta Wacana,* 3(1), 48-54. Doi:10.21460/bikdw.v3i1.116.

Sirait, D. T., Adiwijaya, & Astuti, W. (2019). Analisis perbandingan reduksi dimensi principal component analysis (PCA) dan partial least square (PLS) untuk deteksi kanker menggunakan data microarray. *e-Proceeding of Engineering,* 6(2), 8570-8581. Diakses dari https://openlibrary.telkomuniversity.ac.id/.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management,* 45(4), 427-437. Doi:10.1016/j.ipm.2009.03.002.

Suresh, R., Li, X., Chiriac, A., Goel, K., Terzic, A., Perez-Terzic, C., & Nelson, T. J. (2014). Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction. *Journal of Molecular and Cellular Cardiology,* 74, 13-21. Doi:10.1016/j.yjmcc.2014.04.017.

Tan, C. S., Ting, W. S., Mohamad, M. S., Chan, W. H., Deris, S., & Shah, Z. A. (2014). A review of feature extraction software for microarray gene expression data. *BioMed Research International,* 2014, 15. Doi:10.1155/2014/213656.

Tirta, I. M. (2005). Buku panduan program statistika R. Jember: UPT Penerbitan Universitas Jember.

Trevino, V., Falciani, F., & Barrera-Saldaña, H. A. (2007). DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine,* 13(9-10), 527-541. Doi:10.2119/2006-00107.Trevino.

Turney, P. D. (1999). Learning to extract keyphrases from text. *Departmental Technical Report (Unpublished)*. Diakses dari https://www.researchgate.net/publication/220485917.

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. V. (2011). Confusion matrix-based feature selection. *The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011, Cincinnati, Ohio, USA, April 16-17, 2011*. Cincinnati: MAICS.

Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodoloski zvezki,* 3(1), 89-108. Diakses dari https://www.semanticscholar.org/.

WHO. (2018). News Room. Diakses dari World Health Organization: https://www.who.int/health-topics/cancer#tab=tab_1

Widityasari, D., & Dewi, S. M. (2014). Adenokarsinoma duktal pankreas. *MEDICINA Jurnal Ilmiah Kedokteran,* 45(3), 188-192. Diakses dari https://ojs.unud.ac.id/.

Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). Naïve bayes vs. Decision trees vs. Neural networks in the classification of training web pages. *IJCSI International Journal of Computer Science Issues,* 4(1), 16-23. Diakses dari https://www.semanticscholar.org/.

Yao, F., Coquery, J., & Cao, K.-A. L. (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics,* 13(24). Doi:10.1186/1471-2105-13-24.

Yip, W.-K., Amin, S. B., & Li, C. (2011). A survey of classification techniques for microarray data analysis. Dalam H. H.-S. Lu, B. Schölkopf, & H. Zhao, *Handbook of Statistical Bioinformatics* (hal. 193-223). Berlin: Springer. Doi:10.1007/978-3-642-16345-6_10.

Yuwono, T. (2005). Biologi molekular. Jakarta: Erlangga.

Zhang, H. (2004). The optimality of naive bayes. *17th International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*. American Association for Artificial Intelligence.

LAMPIRAN

Lampiran 1 *Script* pemrograman *software R*

```
######################## DAFTAR LIBRARY ########################
library(GEOquery)
library(affyPLM)
library(stringr)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2)
library(mixOmics)
library(writexl)
library(e1071)
library(pROC)
```

```
######################## PENGUMPULAN DATA ######################
### Download Data ###
if(length(dir("Data/GEO/", pattern = "GSE43837"))==0){
  getGEOSuppFiles("GSE43837", makeDirectory = TRUE)
  untar("GSE43837/GSE43837_RAW.tar", exdir = "GSE43837_RAW/",
        tar = Sys.getenv("TAR"))
}
if(length(dir("Data/GEO/", pattern = "GSE16515"))==0){
  getGEOSuppFiles("GSE16515", makeDirectory = TRUE)
  untar("GSE16515/GSE16515_RAW.tar", exdir = "GSE16515_RAW/",
        tar = Sys.getenv("TAR"))
}
if(length(dir("Data/GEO/", pattern = "GSE48060"))==0){
  getGEOSuppFiles("GSE48060", makeDirectory = TRUE)
  untar("GSE48060/GSE48060_RAW.tar", exdir = "GSE48060_RAW/",
        tar = Sys.getenv("TAR"))
}
### Input Data ###
gsepayu <- list.celfiles("D:/Skripsi 06-
Nov/260820/GSE43837_RAW", full.names = TRUE)
affy.datapayu <- ReadAffy(filenames = gsepayu)
affy.datapayu
gsepank <- list.celfiles("D:/Skripsi 06-
Nov/260820/GSE16515_RAW", full.names = TRUE)
affy.datapank <- ReadAffy(filenames = gsepank)
affy.datapank
gseami <- list.celfiles("D:/Skripsi 06-
Nov/260820/GSE48060_RAW", full.names = TRUE)
affy.dataami <- ReadAffy(filenames = gseami)
affy.dataami
### Input Pheno Data ###
gsetpayu <- getGEO(GEO = "GSE43837", GSEMatrix = TRUE)
data.gsepayu <- exprs(gsetpayu[[1]])
dim(data.gsepayu)
gsetpank <- getGEO(GEO = "GSE16515", GSEMatrix = TRUE)
data.gsepank <- exprs(gsetpank[[1]])
dim(data.gsepank) # 54613 dari 54675
gsetami <- getGEO(GEO = "GSE48060", GSEMatrix = TRUE)
data.gseami <- exprs(gsetami[[1]])
dim(data.gseami)
```

```
###################### ANALISIS DESKRIPTIF ######################
### Deskripsi Kanker Payudara ###
phenopayu <- pData(phenoData(gsetpayu[[1]]))
varLabels(phenoData(gsetpayu[[1]]))
table(phenopayu$'age:ch1')            # Variabel Pheno Umur
table(phenopayu$'disease state:ch1')  # Variabel Pheno Jaringan
Kanker
deskpayu <- data.frame(Umur =
str_remove_all(phenopayu$'age:ch1',"[yr]"),
                       J_Kanker = recode(phenopayu$'disease
state:ch1',
                                         "brain metastatic
breast cancer" = "Brain Metastatic",
                                         "nonmetastatic breast
cancer" = "Non Brain Metastatic"))
head(deskpayu, 4)
dim(deskpayu)
## Histogram Jaringan Kanker ##
ggplot(deskpayu) +
  aes(x = J_Kanker) +
  geom_bar(fill = "#0073C2FF") +
  geom_text(stat = 'count',aes(label =..count.., vjust = -0.2))
+
  ggtitle("Kategori Jaringan Kanker Payudara") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Jaringan Kanker Payudara") + ylab("Frekuensi")
## Histogram Umur ##
hist_umur <- hist(as.numeric(as.character(deskpayu$Umur)),
                  col = "#0073C2FF", ylim = c(0,10), xlim =
c(30,80),
                  main = "Umur Pasien Kanker Payudara", ylab
="Frekuensi",
                  xlab ="Umur")
text(hist_umur$mids,
     hist_umur$counts,
     labels = hist_umur$counts,
     adj = c(0.5, - 0.5))
### Deskripsi Kanker Pankreas ###
phenopank <- pData(phenoData(gsetpank[[1]]))
varLabels(phenoData(gsetpank[[1]]))
table(phenopank$'age:ch1')            # Variabel Pheno Umur
table(phenopank$'sex:ch1')            # Variabel Pheno Jenis
Kelamin
table(phenopank$'tissue:ch1')         # Variabel Pheno Jaringan
Kanker
deskpank <- data.frame(Umur = phenopank$'age:ch1',
                       J_Kanker =
recode(phenopank$'tissue:ch1',
                                         "Tumor Tissue in
Pancreatic Cancer Sample" = "Tumor",
                                         "Normal Tissue in
Pancreatic Cancer Sample" = "Normal"),
                       Sex = recode(phenopank$'sex:ch1',
                                    "Female" = "Perempuan",
                                    "Male" = "Laki-laki"))
head(deskpank, 4)
dim(deskpank)
```

```
## Histogram Jaringan Kanker ##
ggplot(deskpank) +
  aes(x = J_Kanker) +
  geom_bar(fill = "#0073C2FF") +
  geom_text(stat = 'count',aes(label =..count.., vjust = -0.2))
+
  ggtitle("Kategori Jaringan Kanker Pankreas") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Jaringan Kanker Pankreas") + ylab("Frekuensi")
## Histogram Jaringan Kanker dan Jenis Kelamin ##
dfpank <- deskpank %>%
  group_by(J_Kanker, Sex) %>%
  summarise(Frekuensi = n()) %>%
  arrange(J_Kanker, desc(Sex)) %>%
  mutate(lab_ypos = cumsum(Frekuensi) - 0.5 * Frekuensi)
dfpank
ggplot(dfpank, aes(x = J_Kanker, y = Frekuensi)) +
  geom_bar(aes(color = Sex, fill = Sex), stat = "identity") +
  geom_text(aes(y = lab_ypos, label = Frekuensi, group = Sex),
            color = "white") +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF")) +
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF")) +
  ggtitle("Jaringan Kanker Berdasarkan Jenis Kelamin") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Jaringan Kanker Pankreas") + ylab("Frekuensi")
## Histogram Umur ##
hist_umur <- hist(as.numeric(as.character(deskpank$Umur)),
                  col = "#0073C2FF", ylim = c(0,20), xlim =
c(40,90),
                  main = "Umur Pasien Kanker Pankreas", ylab =
"Frekuensi",
                  xlab ="Umur")
text(hist_umur$mids,
     hist_umur$counts,
     labels = hist_umur$counts,
     adj = c(0.5, - 0.5))
### Deskripsi Penyakit AMI ###
phenoami <- pData(phenoData(gsetami[[1]]))
varLabels(phenoData(gsetami[[1]]))
table(phenoami$'disease status:ch1')  # Variabel Pheno Status
Penyakit
deskami <- data.frame(J_Penyakit = recode(phenoami$'disease
status:ch1',
                                          "patient with
recurrent events" = "Recurrent Events",
                                          "patient without
recurrent events" = "Without Recurrent Events",
                                          "normal control" =
"Normal Control"))
head(deskami, 4)
dim(deskami)
## Histogram Status Penyakit ##
ggplot(deskami) +
  aes(x = J_Penyakit) +
  geom_bar(fill = "#0073C2FF") +
  geom_text(stat = 'count',aes(label = ..count.., vjust = -
0.2)) +
  ggtitle("Kategori Status Penyakit AMI") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Status Penyakit") + ylab("Frekuensi")
```

```
######################## PRE-PROCESSING ########################
### Threestep Pre-Processing ###
eset.dChippayu <- threestep(affy.datapayu, background.method =
"RMA.2",
                            normalize.method = "quantile",
                            summary.method = "median.polish")
eset.dChippank <- threestep(affy.datapank, background.method =
"RMA.2",
                            normalize.method = "quantile",
                            summary.method = "median.polish")
eset.dChipami  <- threestep(affy.dataami, background.method =
"RMA.2",
                            normalize.method = "quantile",
                            summary.method = "median.polish")
### Ekprs dan Transppose Data ###
datapayu = as.data.frame(t(exprs(eset.dChippayu)))
datapank = as.data.frame(t(exprs(eset.dChippank)))
dataami  = as.data.frame(t(exprs(eset.dChipami)))
### Boxplot Pre-Processing ###
## Boxplot Sebelum Pre-Processing ##
dev.new(width = 4+dim(gsetpayu)[[2]]/5, height = 6)
par(mar =
c(2+round(max(nchar(sampleNames(gsetpayu)))/2),4,2,1))
title <- paste("Data Kanker Payudara Sebelum Pre-Processing","
(selected samples)", sep = '')
boxplot(exprs(gsetpayu[[1]]), boxwex = 0.7, notch = T, main =
title, outline = FALSE, las = 2)
## Boxplot Setelah Pre-Processing ##
dev.new(width = 4+dim(gsetpayu)[[2]]/5, height = 6)
par(mar =
c(2+round(max(nchar(sampleNames(gsetpayu)))/2),4,2,1))
title <- paste("Data Kanker Payudara Setelah Pre-Processing","
(selected samples)", sep = '')
boxplot(exprs(eset.dChippayu), boxwex = 0.7, notch = T, main =
title, outline = FALSE, las = 2)
```

```
###################### REDUKSI DIMENSI IPCA ####################
N <- 30
Exp.Var <- matrix(0,N,3)
for (i in 1:N) {
  payu.ipca <- ipca(datapayu, ncomp = i, mode = "deflation",
scale = FALSE)
  pank.ipca <- ipca(datapank, ncomp = i, mode = "deflation",
scale = FALSE)
  ami.ipca <- ipca(dataami, ncomp = i, mode = "deflation",
scale = FALSE)
  Exp.Var[i,1] <- sum(payu.ipca$explained_variance)
  Exp.Var[i,2] <- sum(pank.ipca$explained_variance)
  Exp.Var[i,3] <- sum(ami.ipca$explained_variance)
}
colnames(Exp.Var) <- c("Payu", "Pank", "AMI")
Exp.Var <- as.data.frame(Exp.Var)
write_xlsx(Exp.Var,"Var_IPCA.xlsx")
```

```
##################### MEMILIH KOMPONEN IPC ####################
Exp.Var[c(1,2,3,5,10,15,30),]
Exp.Var[c(10,15,30),]
```

```
############## MEMBAGI DATA TRAINING DAN TESTING ##############
dataY1 = as.factor(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
                     1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))
dataY2 = as.factor(c(0,0,0,0,1,0,1,0,1,0,1,0,0,0,1,0,1,0,1,
                     0,0,0,0,1,0,0,0,0,0,1,0,1,0,1,0,0,0,0,
                     0,1,0,0,1,0,0,0,1,0,1,0,0,1))
dataY3 = as.factor(c(0,1,1,1,1,1,0,1,1,0,1,1,1,1,1,1,0,1,1,
                     1,1,1,1,1,1,0,1,1,1,1,2,2,2,2,2,2,2,2,
                     2,2,2,2,2,2,2,2,2,2,2,2,1,2))
set.seed(978)
rasio = 8/10
train1 = sample(length(dataY1), size =
floor(rasio*length(dataY1)))
set.seed(978)
rasio = 8/10
train2 = sample(length(dataY2), size =
floor(rasio*length(dataY2)))
set.seed(978)
rasio = 8/10
train3 = sample(length(dataY3), size =
floor(rasio*length(dataY3)))
```

```
#################### KLASIFIKASI DATA PAYU ####################
for (i in c(10,15,30)) {
  print("############ KLASIFIKASI DATA PAYU ############")
  print(paste("Klasifikasi Berdasarkan Jumlah IPC Sebanyak:",
i))
  payu.ipca <- ipca(datapayu, ncomp = i, mode = "deflation",
scale = FALSE)
  payu <- as.data.frame(payu.ipca$x)
  databaru <- as.data.frame(cbind(payu, dataY1))
  datatrain <- databaru[train1, ]
  print("Dimensi data training")
  print(dim(datatrain))
  datatest <- databaru[-train1, ]
  print("Dimensi data testing")
  print(dim(datatest))
  ### Klasifikasi SVM ###
  print("##### KLASIFIKASI SVM #####")
  ## kernel Linier ##
  tunaslin <- svm(dataY1 ~ . ,
                  data = datatrain, kernel = "linear", types =
"C-clasification",
                  ranges = list(cost =
c(0.1,0.01,0.001,1,10,100)))
  model <- svm(dataY1~. , datatrain, kernel = "linear",
               cost = 1, scale = F, types = "C-clasification",
               decision.value = T)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Linear")
  print(table(Prediksi, datatest$dataY1))
  print(paste("Akurasi Klasifikasi SVM Kernel Linear =",
              mean(Prediksi == datatest$dataY1)))
```

```
  ## kernel Polynomial ##
  tuning <- svm(dataY1~. , data = datatrain, kernel =
"polynomial",
                types = "C-clasification", ranges = list( cost
= c(100,200,300,400,500)))
  model <- svm(dataY1~. , datatrain, kernel = "polynomial",
degree = 3,
               cost = 1, types = "C-clasification",
decision.value = T)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Polynomial")
  print(table(Prediksi, datatest$dataY1))
  print(paste("Akurasi Klasifikasi SVM Kernel Polynomial =",
              mean(Prediksi == datatest$dataY1)))
  ## kernel RBF ##
  tun <- svm(dataY1~. , data = datatrain,
             kernel ="radial", types = "C-clasification",
             ranges = list(cost = c(0.1,0.01,0.001,1,10,100),
                           gamma = c(1,2,3,4,5)))
  model <- svm(dataY1~. , datatrain, kernel = "radial", cost =
1)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel RBF")
  print(table(Prediksi, datatest$dataY1))
  print(paste("Akurasi Klasifikasi SVM Kernel RBF =",
              mean(Prediksi == datatest$dataY1)))
  ### Klasifikasi NBC ###
  print("##### KLASIFIKASI NBC #####")
  NBayes <- naiveBayes(datatrain[1:i], datatrain$dataY1,
laplace = 0)
  print(NBayes$apriori)
  Prediksi <- predict(NBayes, datatest)
  print("Confusion Matrix Data Test Naive Bayes")
  print(table(Prediksi, datatest$dataY1))
  print(paste("Akurasi Klasifikasi Naive Bayes =",
              mean(Prediksi == datatest$dataY1)))
  if (i == 10){
    if (file.exists("Data_IPC10_Payudara.xlsx")  == FALSE){
      write_xlsx(payu,"Data_IPC10_Payudara.xlsx")
    }
  }
}
```

```
##################### KLASIFIKASI DATA PANK ###################
for (i in c(10,15,30)) {
  print("############ KLASIFIKASI DATA PANK ############")
  print(paste("Klasifikasi Berdasarkan Jumlah IPC Sebanyak:",
i))
  pank.ipca <- ipca(datapank, ncomp = i, mode = "deflation",
scale = FALSE)
  pank <- as.data.frame(pank.ipca$x)
  databaru <- as.data.frame(cbind(pank, dataY2))
  datatrain <- databaru[train2, ]
  print("Dimensi data training")
  print(dim(datatrain))
  datatest <- databaru[-train2, ]
  print("Dimensi data testing")
  print(dim(datatest))
```

```
  ### Klasifikasi SVM ###
  print("##### KLASIFIKASI SVM #####")
  ## kernel Linier ##
  tunaslin <- svm(dataY2 ~ . ,
                  data = datatrain, kernel = "linear", types =
"C-clasification",
                  ranges = list(cost =
c(0.1,0.01,0.001,1,10,100)))
  model <- svm(dataY2~. , datatrain, kernel = "linear",
               cost = 1, scale = F, types = "C-clasification",
               decision.value = T)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Linear")
  print(table(Prediksi, datatest$dataY2))
  print(paste("Akurasi Klasifikasi SVM Kernel Linear =",
              mean(Prediksi == datatest$dataY2)))
  ## kernel Polynomial ##
  tuning <- svm(dataY2~. , data = datatrain, kernel =
"polynomial",
                types = "C-clasification", ranges = list( cost
= c(100,200,300,400,500)))
  model <- svm(dataY2~. , datatrain, kernel = "polynomial",
degree = 3,
               cost = 1, types = "C-clasification",
decision.value = T)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Polynomial")
  print(table(Prediksi, datatest$dataY2))
  print(paste("Akurasi Klasifikasi SVM Kernel Polynomial =",
              mean(Prediksi == datatest$dataY2)))
  ## kernel RBF ##
  tun <- svm(dataY2~. , data = datatrain,
             kernel ="radial", types = "C-clasification",
             ranges = list(cost = c(0.1,0.01,0.001,1,10,100),
                           gamma = c(1,2,3,4,5)))
  model <- svm(dataY2~. , datatrain, kernel = "radial", cost =
1)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel RBF")
  print(table(Prediksi, datatest$dataY2))
  print(paste("Akurasi Klasifikasi SVM Kernel RBF =",
              mean(Prediksi == datatest$dataY2)))
  if (i == 10){
    ## ROC Curve ##
    forauc = as.numeric(predict(model, datatest))
    forauc
    auc = multiclass.roc(datatest$dataY, forauc)
    auc$auc
    plot.roc(auc[["rocs"]][[1]],print.auc = T,
             col = "#0073C2FF",print.auc.x = 0.4,
             print.auc.y = 0.7, legacy.axes = T,
             main = "(a)")
  }
      ### Klasifikasi NBC ###
  print("##### KLASIFIKASI NBC #####")
  NBayes <- naiveBayes(datatrain[1:i], datatrain$dataY2,
laplace = 0)
  print(NBayes$apriori)
  Prediksi <- predict(NBayes, datatest)
  print("Confusion Matrix Data Test Naive Bayes")
```

```
  print(table(Prediksi, datatest$dataY2))
  print(paste("Akurasi Klasifikasi Naive Bayes =",
              mean(Prediksi == datatest$dataY2)))
  if (i == 10){
    if (file.exists("Data_IPC10_Pankreas.xlsx")  == FALSE){
      write_xlsx(pank,"Data_IPC10_Pankreas.xlsx")
    }
    ## ROC Curve ##
    forauc = as.numeric(predict(NBayes, datatest))
    forauc
    auc = multiclass.roc(datatest$dataY, forauc)
    auc$auc
    plot.roc(auc[["rocs"]][[1]],print.auc = T,
             col = "#0073C2FF",print.auc.x = 0.4,
             print.auc.y = 0.7, legacy.axes = T,
             main = "(b)")
  }
}
```

```
###################### KLASIFIKASI DATA AMI ####################
for (i in c(10,15,30)) {
  print("############ KLASIFIKASI DATA AMI ###########")
  print(paste("Klasifikasi Berdasarkan Jumlah IPC Sebanyak:",
i))
  ami.ipca <- ipca(dataami, ncomp = i, mode = "deflation",
scale = FALSE)
  ami <- as.data.frame(ami.ipca$x)
  databaru <- as.data.frame(cbind(ami, dataY3))
  datatrain <- databaru[train3, ]
  print("Dimensi data training")
  print(dim(datatrain))
  datatest <- databaru[-train3, ]
  print("Dimensi data testing")
  print(dim(datatest))
  ### Klasifikasi SVM ###
  print("##### KLASIFIKASI SVM #####")
  ## kernel Linier ##
  tunaslin <- svm(dataY3 ~ . ,
                  data = datatrain, kernel = "linear", types =
"C-clasification",
                  ranges = list(cost =
c(0.1,0.01,0.001,1,10,100)))
  model <- svm(dataY3~. , datatrain, kernel = "linear",
               cost = 1, scale = F, types = "C-clasification",
               decision.value = T)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Linear")
  print(table(Prediksi, datatest$dataY3))
  print(paste("Akurasi Klasifikasi SVM Kernel Linear =",
              mean(Prediksi == datatest$dataY3)))
  ## kernel Polynomial ##
  tuning <- svm(dataY3~. , data = datatrain, kernel =
"polynomial",
                types = "C-clasification", ranges = list( cost
= c(100,200,300,400,500)))
  model <- svm(dataY3~. , datatrain, kernel = "polynomial",
degree = 3,
               cost = 1, types = "C-clasification",
decision.value = T)
```

```
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel Polynomial")
  print(table(Prediksi, datatest$dataY3))
  print(paste("Akurasi Klasifikasi SVM Kernel Polynomial =",
              mean(Prediksi == datatest$dataY3)))
  ## kernel RBF ##
  tun <- svm(dataY3~. , data = datatrain,
             kernel ="radial", types = "C-clasification",
             ranges = list(cost = c(0.1,0.01,0.001,1,10,100),
                           gamma = c(1,2,3,4,5)))
  model <- svm(dataY3~. , datatrain, kernel = "radial", cost =
1)
  Prediksi <- predict(model, datatest)
  print("Confusion Matrix Data Test SVM Kernel RBF")
  print(table(Prediksi, datatest$dataY3))
  print(paste("Akurasi Klasifikasi SVM Kernel RBF =",
              mean(Prediksi == datatest$dataY3)))
  ### Klasifikasi NBC ###
  print("##### KLASIFIKASI NBC #####")
  NBayes <- naiveBayes(datatrain[1:i], datatrain$dataY3,
laplace = 0)
  print(NBayes$apriori)
  Prediksi <- predict(NBayes, datatest)
  print("Confusion Matrix Data Test Naive Bayes")
  print(table(Prediksi, datatest$dataY3))
  print(paste("Akurasi Klasifikasi Naive Bayes =",
              mean(Prediksi == datatest$dataY3)))
  if (i == 10){
    if (file.exists("Data_IPC10_AMI.xlsx")  == FALSE){
      write_xlsx(ami,"Data_IPC10_AMI.xlsx")
    }
  }
}

#################################################################
```

Lampiran 2 Nilai *explained variance* IPCA berdasarkan jumlah komponen IPC

IPC	Payu	Pank	AMI	IPC	Payu	Pank	AMI
IPC1	24,20%	30,58%	31,54%	IPC16	72,57%	79,75%	75,05%
IPC2	33,62%	41,68%	43,07%	IPC17	74,44%	80,89%	76,14%
IPC3	38,94%	48,03%	47,86%	IPC18	76,29%	81,97%	77,15%
IPC4	43,20%	52,82%	52,00%	IPC19	78,00%	82,96%	78,14%
IPC5	47,03%	57,09%	55,82%	IPC20	79,62%	83,93%	79,07%
IPC6	50,20%	60,63%	59,03%	IPC21	81,22%	84,81%	79,99%
IPC7	53,09%	63,60%	61,46%	IPC22	82,73%	85,63%	80,89%
IPC8	55,63%	66,09%	63,60%	IPC23	84,21%	86,46%	81,74%
IPC9	58,12%	68,51%	65,35%	IPC24	85,64%	87,23%	82,56%
IPC10	60,35%	70,65%	67,02%	IPC25	87,05%	87,96%	83,37%
IPC11	62,57%	72,61%	68,66%	IPC26	88,42%	88,68%	84,15%
IPC12	64,68%	74,26%	70,08%	IPC27	89,76%	89,35%	84,92%
IPC13	66,69%	75,81%	71,46%	IPC28	91,02%	90,00%	85,68%
IPC14	68,69%	77,19%	72,72%	IPC29	92,23%	90,64%	86,43%
IPC15	70,64%	78,53%	73,92%	IPC30	93,38%	91,26%	87,17%

Lampiran 3 *Output* hasil reduksi dimensi IPCA data kanker payudara dengan 10 komponen IPC

Data Kanker Payudara										
No.	**IPC1**	**IPC2**	**IPC3**	**IPC4**	**IPC5**	**IPC6**	**IPC7**	**IPC8**	**IPC9**	**IPC10**
1	-0,173	-0,062	0,157	0,049	0,153	-0,035	-0,097	0,156	0,191	-0,184
2	0,086	-0,016	0,052	0,131	0,271	0,083	0,010	0,187	0,235	-0,079
3	0,061	0,077	-0,036	0,074	-0,243	0,044	-0,131	-0,003	0,168	-0,012
4	-0,411	0,189	-0,161	0,116	0,250	0,043	0,038	0,036	-0,284	0,325
5	0,037	0,030	-0,050	0,075	-0,203	-0,148	-0,057	-0,025	0,123	0,047
6	0,151	-0,107	-0,080	0,103	0,112	0,136	0,580	-0,175	0,040	0,050
7	0,153	0,178	-0,070	-0,047	0,266	-0,159	0,040	0,162	-0,017	-0,054
8	-0,227	0,039	-0,094	0,080	-0,209	-0,034	-0,175	0,012	0,066	0,066
9	-0,391	0,112	-0,044	-0,017	0,092	-0,231	0,079	-0,130	0,079	-0,205
10	0,087	-0,060	0,025	0,202	-0,130	-0,040	-0,253	-0,014	0,069	0,032
11	0,249	0,114	-0,038	0,004	0,211	-0,028	0,216	0,013	0,276	-0,065
12	0,173	-0,055	0,077	0,038	0,251	0,184	-0,214	0,034	0,127	-0,250
13	-0,380	0,073	-0,084	-0,149	0,180	-0,161	0,052	0,001	0,257	-0,179
14	0,053	0,179	-0,016	-0,062	-0,142	0,164	0,001	0,022	-0,260	-0,294
15	-0,090	0,223	0,892	-0,142	-0,025	0,103	0,024	-0,087	-0,016	0,161
16	0,110	0,156	-0,085	-0,014	0,147	0,150	-0,129	0,152	-0,054	0,338
17	0,067	-0,147	0,041	0,157	-0,139	-0,120	-0,079	-0,092	0,074	0,017
18	0,068	0,130	-0,157	-0,229	0,176	0,249	-0,225	-0,769	0,082	0,060
19	0,104	0,046	-0,020	0,067	-0,196	-0,109	-0,014	-0,095	0,048	0,012
20	-0,142	0,037	-0,091	-0,091	-0,158	0,206	0,001	0,105	0,024	-0,059
21	0,074	0,149	-0,072	-0,083	-0,110	0,219	-0,049	0,106	0,015	-0,147
22	-0,158	0,210	-0,098	-0,101	-0,147	0,246	0,074	0,151	-0,140	-0,115
23	-0,289	-0,241	-0,011	0,105	-0,090	-0,080	-0,022	-0,036	0,040	0,005
24	0,149	0,130	-0,096	-0,245	0,014	-0,252	0,003	0,082	-0,081	0,078
25	0,091	0,100	0,102	-0,026	-0,085	-0,353	0,231	-0,076	-0,019	0,112
26	0,006	-0,143	0,033	0,147	-0,031	-0,074	0,242	-0,209	-0,047	-0,122
27	-0,002	-0,174	0,004	0,121	-0,195	0,004	-0,131	-0,138	0,006	0,021
28	0,108	0,095	-0,108	-0,013	-0,146	0,080	-0,066	0,058	0,040	0,043
29	0,081	-0,022	0,033	0,090	0,092	0,019	-0,084	0,164	-0,108	0,203
30	-0,016	-0,178	0,045	0,246	0,157	0,058	-0,019	0,073	-0,022	-0,029
31	0,109	-0,123	0,062	0,262	0,065	-0,038	-0,125	0,036	-0,063	-0,148
32	0,166	0,071	0,007	-0,215	0,013	-0,381	-0,170	0,005	-0,408	-0,300
33	0,119	0,085	-0,111	-0,144	-0,069	-0,120	-0,005	0,158	0,268	0,378
34	0,087	-0,089	-0,003	0,159	0,077	-0,027	0,109	-0,093	-0,419	-0,008
35	-0,066	-0,137	-0,015	0,168	0,122	0,145	-0,074	0,000	-0,243	0,183
36	-0,022	-0,629	0,015	-0,579	0,123	-0,007	-0,105	0,075	-0,041	0,119
37	-0,004	-0,143	-0,005	-0,194	-0,222	0,329	0,287	0,161	-0,016	-0,161
38	-0,018	-0,097	0,001	-0,045	-0,234	-0,067	0,236	-0,007	0,013	0,160

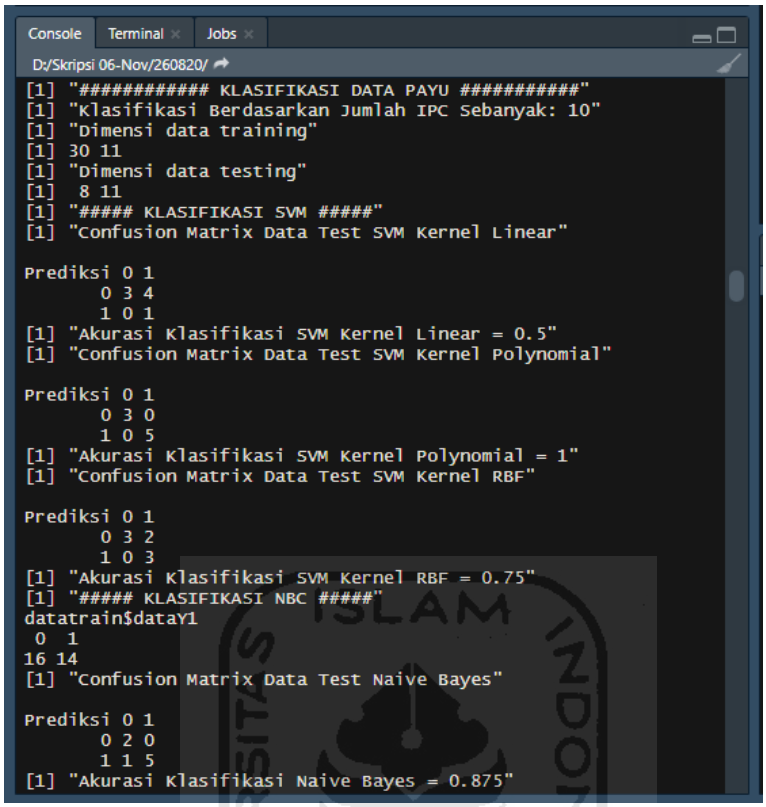
Lampiran 4 *Output* hasil reduksi dimensi IPCA data kanker pankreas dengan 10 komponen IPC

Data Kanker Pankreas										
No.	**IPC1**	**IPC2**	**IPC3**	**IPC4**	**IPC5**	**IPC6**	**IPC7**	**IPC8**	**IPC9**	**IPC10**
1	0,085	0,274	0,022	0,063	0,021	0,034	-0,026	-0,136	0,162	0,050
2	0,065	0,177	0,125	0,033	-0,082	-0,075	-0,120	-0,114	0,190	0,082
3	-0,066	0,119	-0,011	0,122	-0,010	-0,109	-0,151	-0,322	0,218	0,055
4	0,247	0,001	-0,081	0,046	0,079	-0,037	0,154	-0,035	-0,018	0,158
5	0,255	0,045	-0,070	0,007	0,123	0,033	0,165	-0,025	0,020	0,137
6	-0,175	0,160	-0,236	0,049	-0,046	0,030	-0,021	0,003	0,072	0,081
7	0,162	-0,108	-0,069	0,171	-0,009	-0,137	0,139	-0,203	0,093	0,050
8	-0,169	0,053	0,157	0,148	-0,102	-0,024	0,087	0,000	-0,040	-0,004
9	0,076	-0,090	0,046	-0,191	-0,006	0,082	0,010	-0,064	0,027	-0,072
10	-0,154	0,093	0,104	0,233	-0,077	-0,081	0,377	-0,178	-0,078	-0,061
11	-0,230	-0,416	-0,121	-0,129	-0,132	-0,168	0,069	-0,009	0,146	-0,122
12	0,020	-0,024	0,100	-0,112	0,021	-0,048	-0,016	-0,042	0,099	0,037
13	-0,130	0,006	0,109	0,009	-0,101	-0,006	0,243	0,193	-0,074	0,120
14	-0,084	0,069	0,059	-0,084	0,001	-0,026	0,177	0,095	-0,107	0,077
15	-0,211	0,233	-0,768	-0,118	0,061	0,070	-0,021	-0,006	-0,023	0,095
16	-0,057	0,043	0,098	-0,036	0,000	-0,043	0,021	-0,014	-0,034	0,053
17	0,002	-0,175	-0,028	-0,149	-0,002	-0,103	0,082	-0,164	0,234	-0,026
18	0,059	0,171	0,074	-0,061	0,007	0,005	-0,028	0,017	0,096	0,033
19	0,165	-0,102	-0,110	0,209	-0,016	-0,120	0,042	-0,055	-0,053	0,079
20	-0,201	-0,035	0,083	0,021	-0,089	-0,105	0,049	0,014	0,031	-0,043
21	-0,213	-0,210	0,078	0,221	0,812	0,149	0,010	0,240	0,212	0,010
22	0,093	-0,177	-0,047	0,163	-0,397	0,699	-0,006	0,247	0,351	0,102
23	-0,017	0,023	0,164	0,103	-0,065	-0,110	-0,128	-0,081	0,334	0,098
24	0,179	-0,062	-0,066	0,140	0,003	-0,099	0,095	-0,038	0,080	-0,018
25	-0,072	0,176	0,120	0,052	-0,021	-0,008	0,063	0,095	0,009	-0,039
26	0,163	0,004	0,111	-0,118	0,020	-0,015	0,071	0,039	0,034	-0,010
27	0,097	0,041	0,099	-0,151	0,062	0,012	0,065	-0,033	0,029	0,065
28	-0,079	0,042	0,005	-0,100	-0,050	-0,013	0,347	0,112	0,028	0,027
29	0,054	-0,042	0,017	-0,211	0,036	0,053	-0,162	-0,022	-0,082	0,056
30	-0,181	-0,350	-0,073	-0,199	-0,050	-0,113	-0,124	-0,008	0,032	0,062
31	0,009	0,067	0,093	-0,121	-0,041	0,070	-0,213	0,061	-0,138	0,065
32	0,071	-0,072	0,013	-0,243	0,012	0,032	-0,137	-0,023	-0,089	0,077
33	-0,190	-0,118	0,083	-0,067	-0,087	0,011	-0,002	0,056	-0,212	0,123
34	0,163	-0,093	-0,065	0,028	-0,023	0,001	-0,016	0,018	-0,105	-0,268
35	-0,168	0,149	-0,050	0,058	-0,054	-0,016	-0,099	0,089	-0,054	0,012
36	-0,144	-0,002	-0,035	0,022	0,061	-0,026	-0,180	-0,202	-0,013	-0,021
37	-0,138	0,153	-0,033	0,097	-0,007	0,124	-0,074	-0,022	0,031	-0,476
38	0,086	0,069	0,064	-0,204	0,020	0,027	-0,020	0,053	-0,188	-0,022
39	0,156	0,051	-0,014	-0,037	0,046	0,076	-0,096	-0,117	0,082	-0,537
40	0,118	-0,147	-0,020	-0,193	0,031	0,010	0,015	-0,097	-0,017	-0,071
41	0,045	0,079	0,066	-0,121	0,048	0,065	-0,081	-0,066	-0,038	-0,018
42	-0,218	-0,091	0,094	0,103	-0,117	-0,095	-0,272	-0,047	-0,044	0,132
43	0,149	-0,133	-0,103	0,291	-0,056	-0,129	-0,089	0,071	-0,198	0,140
44	-0,080	-0,097	0,090	0,087	0,131	0,483	0,056	-0,506	-0,420	0,128
45	0,036	0,115	-0,107	-0,211	0,035	-0,020	0,355	0,079	0,039	-0,046
46	0,095	0,193	0,010	-0,024	0,099	-0,009	-0,163	0,176	-0,020	0,077
47	0,099	-0,108	0,012	-0,132	0,042	-0,011	-0,066	0,134	0,041	-0,027
48	-0,164	0,061	0,130	0,051	-0,080	0,015	0,056	0,180	-0,135	-0,251
49	0,124	-0,130	-0,135	0,228	-0,049	-0,061	-0,045	0,078	-0,224	-0,288
50	-0,057	0,160	0,065	0,021	0,021	-0,071	-0,182	0,175	-0,037	0,040
51	0,158	0,077	0,059	-0,018	0,014	-0,078	-0,083	0,204	-0,102	0,031
52	0,168	-0,117	-0,106	0,253	-0,038	-0,125	-0,130	0,198	-0,145	0,067

Lampiran 5 *Output* hasil reduksi dimensi IPCA data penyakit AMI dengan 10 komponen IPC

Data Penyakit AMI										
No.	**IPC1**	**IPC2**	**IPC3**	**IPC4**	**IPC5**	**IPC6**	**IPC7**	**IPC8**	**IPC9**	**IPC10**
1	-0,171	-0,117	-0,268	-0,059	0,005	0,143	-0,124	-0,159	0,187	-0,004
2	-0,103	-0,125	-0,102	-0,067	-0,009	-0,023	0,062	0,047	0,010	-0,155
3	0,153	-0,248	0,144	-0,005	-0,202	0,012	-0,030	0,046	-0,028	0,114
4	0,130	-0,085	-0,217	0,100	-0,026	0,534	0,078	0,032	0,201	0,199
5	0,036	0,168	0,161	-0,092	-0,408	0,296	0,254	0,241	0,374	-0,239
6	0,084	-0,132	0,080	0,000	0,005	-0,082	0,081	-0,143	-0,061	-0,026
7	0,083	-0,094	-0,129	0,046	0,054	0,072	0,125	0,075	-0,011	0,200
8	-0,200	-0,260	-0,012	-0,017	0,107	-0,097	0,000	0,024	0,096	0,111
9	-0,104	-0,124	-0,078	0,024	0,032	-0,092	0,425	0,066	-0,176	-0,273
10	-0,191	-0,236	0,145	0,384	0,276	0,126	-0,129	-0,022	0,248	-0,074
11	-0,118	0,150	-0,238	-0,056	-0,030	-0,091	-0,104	-0,023	-0,023	-0,019
12	-0,195	-0,194	0,091	0,220	-0,117	-0,174	0,254	-0,191	0,111	0,075
13	0,110	0,059	0,041	0,179	-0,035	-0,095	0,145	-0,148	0,028	-0,228
14	0,134	-0,059	0,023	0,001	-0,101	-0,011	-0,077	-0,102	0,006	0,146
15	0,113	0,070	-0,135	0,154	0,040	-0,025	0,140	-0,048	0,183	0,098
16	0,127	0,136	0,116	0,091	0,125	0,166	0,054	-0,286	-0,022	-0,142
17	0,090	-0,092	-0,235	0,159	-0,021	-0,176	0,270	0,125	-0,153	0,166
18	0,149	0,086	0,048	0,134	-0,218	0,030	0,018	-0,172	-0,035	0,157
19	0,065	0,099	0,021	-0,096	-0,044	-0,054	0,216	-0,157	-0,038	0,127
20	-0,162	0,151	-0,240	0,019	-0,077	-0,126	0,102	0,123	-0,109	0,105
21	0,119	0,031	-0,024	0,135	0,047	0,102	-0,142	0,072	-0,097	0,062
22	0,116	0,010	0,145	0,182	-0,179	-0,175	-0,191	0,168	-0,120	0,020
23	-0,154	0,194	-0,132	-0,115	-0,016	0,028	0,015	-0,103	-0,024	0,105
24	-0,222	-0,055	-0,089	0,022	0,177	0,189	-0,174	0,244	-0,110	0,174
25	-0,231	0,260	-0,069	0,028	-0,023	0,010	-0,172	-0,305	-0,016	-0,053
26	0,092	0,053	-0,176	0,006	0,023	0,128	0,051	-0,116	-0,171	-0,080
27	-0,156	0,047	0,069	0,316	-0,070	0,033	-0,018	-0,138	-0,341	0,020
28	0,055	-0,136	-0,117	-0,043	0,177	-0,082	-0,009	0,318	-0,022	-0,373
29	0,098	-0,175	-0,091	0,010	-0,039	-0,062	-0,045	0,086	0,016	0,160
30	-0,132	-0,175	0,117	-0,150	-0,174	-0,069	0,028	-0,020	-0,017	0,005
31	0,108	0,104	0,205	0,020	0,108	0,054	-0,033	0,082	0,244	0,087
32	-0,186	-0,306	0,128	-0,406	-0,027	0,085	0,061	-0,216	0,037	0,007
33	0,143	0,048	-0,227	0,039	0,123	-0,125	-0,097	0,010	0,221	-0,174
34	0,143	0,155	0,163	-0,072	-0,005	-0,151	0,012	0,061	0,048	0,202
35	0,141	-0,031	-0,187	-0,102	0,045	-0,196	-0,124	-0,054	0,124	-0,157
36	0,105	-0,076	0,128	-0,254	-0,008	-0,116	-0,053	0,044	-0,023	0,062
37	0,130	0,210	-0,125	-0,016	0,015	-0,242	-0,103	-0,121	0,270	-0,060
38	-0,164	0,053	0,226	-0,032	0,042	-0,207	0,001	0,004	0,103	0,049
39	0,108	-0,082	0,041	-0,142	-0,014	0,015	-0,030	-0,022	0,010	-0,137
40	-0,182	0,251	0,201	-0,061	0,225	-0,053	0,085	0,098	0,030	0,117
41	-0,173	0,047	0,025	-0,227	-0,020	-0,004	-0,157	-0,012	0,067	0,070
42	0,089	0,001	0,014	-0,075	0,031	-0,040	-0,054	0,143	-0,057	-0,058
43	-0,105	0,280	0,168	-0,049	0,191	0,098	0,154	0,283	-0,079	0,013
44	0,133	-0,039	0,026	-0,029	0,006	-0,123	-0,261	0,000	0,032	0,178
45	0,129	-0,081	0,032	-0,093	0,108	0,083	-0,079	-0,003	-0,176	-0,053
46	-0,154	0,099	0,136	0,093	0,157	0,003	0,092	0,033	0,064	0,048
47	0,095	0,072	-0,097	-0,140	0,178	-0,020	0,080	0,072	-0,059	0,022
48	0,096	0,075	-0,008	-0,257	0,108	0,265	0,030	-0,194	-0,299	-0,050
49	0,122	-0,071	0,243	0,132	0,195	0,070	-0,165	-0,104	-0,106	-0,269
50	0,142	-0,006	0,138	0,010	-0,096	0,061	-0,035	0,080	-0,164	0,006
51	-0,117	0,004	-0,030	0,076	-0,236	0,081	-0,222	0,231	-0,108	-0,044
52	-0,219	0,087	-0,047	0,075	-0,404	0,026	-0,207	0,050	-0,065	-0,239

Lampiran 6 Output hasil *confusion matrix* klasifikasi data kanker payudara (IPC10)



```
Console  Terminal  Jobs
D:/Skripsi 06-Nov/260820/
[1] "########### KLASIFIKASI DATA PAYU ##########"
[1] "Klasifikasi Berdasarkan Jumlah IPC Sebanyak: 10"
[1] "Dimensi data training"
[1] 30 11
[1] "Dimensi data testing"
[1]  8 11
[1] "##### KLASIFIKASI SVM #####"
[1] "Confusion Matrix Data Test SVM Kernel Linear"

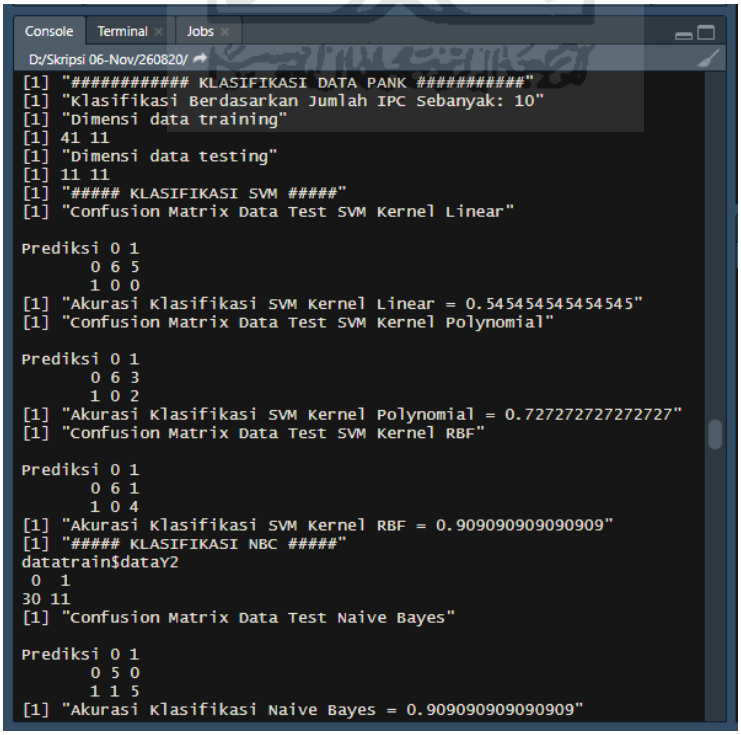
Prediksi 0 1
       0 3 4
       1 0 1
[1] "Akurasi Klasifikasi SVM Kernel Linear = 0.5"
[1] "Confusion Matrix Data Test SVM Kernel Polynomial"

Prediksi 0 1
       0 3 0
       1 0 5
[1] "Akurasi Klasifikasi SVM Kernel Polynomial = 1"
[1] "Confusion Matrix Data Test SVM Kernel RBF"

Prediksi 0 1
       0 3 2
       1 0 3
[1] "Akurasi Klasifikasi SVM Kernel RBF = 0.75"
[1] "##### KLASIFIKASI NBC #####"
datatrain$dataY1
 0  1
16 14
[1] "Confusion Matrix Data Test Naive Bayes"

Prediksi 0 1
       0 2 0
       1 1 5
[1] "Akurasi Klasifikasi Naive Bayes = 0.875"
```

Lampiran 7 *Output* hasil *confusion matrix* klasifikasi data kanker pankreas (IPC10)



```
Console  Terminal  Jobs
D:/Skripsi 06-Nov/260820/
[1] "########### KLASIFIKASI DATA PANK ##########"
[1] "Klasifikasi Berdasarkan Jumlah IPC Sebanyak: 10"
[1] "Dimensi data training"
[1] 41 11
[1] "Dimensi data testing"
[1] 11 11
[1] "##### KLASIFIKASI SVM #####"
[1] "Confusion Matrix Data Test SVM Kernel Linear"

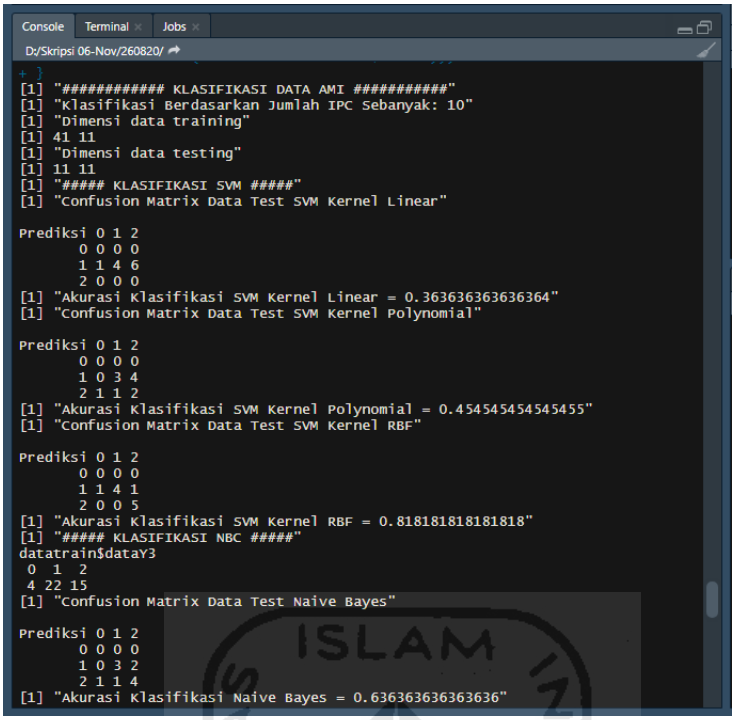
Prediksi 0 1
       0 6 5
       1 0 0
[1] "Akurasi Klasifikasi SVM Kernel Linear = 0.545454545454545"
[1] "Confusion Matrix Data Test SVM Kernel Polynomial"

Prediksi 0 1
       0 6 3
       1 0 2
[1] "Akurasi Klasifikasi SVM Kernel Polynomial = 0.727272727272727"
[1] "Confusion Matrix Data Test SVM Kernel RBF"

Prediksi 0 1
       0 6 1
       1 0 4
[1] "Akurasi Klasifikasi SVM Kernel RBF = 0.909090909090909"
[1] "##### KLASIFIKASI NBC #####"
datatrain$dataY2
 0  1
30 11
[1] "Confusion Matrix Data Test Naive Bayes"

Prediksi 0 1
       0 5 0
       1 1 5
[1] "Akurasi Klasifikasi Naive Bayes = 0.909090909090909"
```

Lampiran 8 *Output* hasil *confusion matrix* klasifikasi data penyakit AMI (IPC10)



```
Console  Terminal  Jobs
D:/Skripsi 06-Nov/260820/
+ }
[1] "########### KLASIFIKASI DATA AMI ##########"
[1] "Klasifikasi Berdasarkan Jumlah IPC Sebanyak: 10"
[1] "Dimensi data training"
[1] 41 11
[1] "Dimensi data testing"
[1] 11 11
[1] "##### KLASIFIKASI SVM #####"
[1] "Confusion Matrix Data Test SVM Kernel Linear"

Prediksi 0 1 2
       0 0 0 0
       1 1 4 6
       2 0 0 0
[1] "Akurasi Klasifikasi SVM Kernel Linear = 0.363636363636364"
[1] "Confusion Matrix Data Test SVM Kernel Polynomial"

Prediksi 0 1 2
       0 0 0 0
       1 0 3 4
       2 1 1 2
[1] "Akurasi Klasifikasi SVM Kernel Polynomial = 0.454545454545455"
[1] "Confusion Matrix Data Test SVM Kernel RBF"

Prediksi 0 1 2
       0 0 0 0
       1 1 4 1
       2 0 0 5
[1] "Akurasi Klasifikasi SVM Kernel RBF = 0.818181818181818"
[1] "##### KLASIFIKASI NBC #####"
datatrain$dataY3
 0  1  2
 4 22 15
[1] "Confusion Matrix Data Test Naive Bayes"

Prediksi 0 1 2
       0 0 0 0
       1 0 3 2
       2 1 1 4
[1] "Akurasi Klasifikasi Naive Bayes = 0.636363636363636"
```