BAB III
LANDASAN TEORI

3.1 Bioinformatika

Menurut Bayat (2002) bioinformatika ialah ilmu yang menerapkan sebuah teknik komputasi untuk menganalisa dan mengolah informasi biologis. Bioinformatik memiliki arti gabungan dari biologi dan ilmu teknik informasi yang asal katanya “bio” dan “informatika”. Bioinformatika merupakan ilmu baru yang mencangkup berbagai bidang yang saling menunjang yakni matematika, fisika, biologi, kedokteran, dan ilmu komputer.

Tumpukan informasi berbagai gen dimulai dari mahluk hidup tingkat rendah maupun tinggi merupakan hasil dari *genome project* yang dilakukan di seluruh dunia. Genom dari manusia yang berjumlah 2,91 juta bp (*base-pair*, pasangan basa) sudah selesai dibaca. Genom adalah sebuah cetak biru dari suatu informasi genetik yang dapat mendefinisikan sifat dari setiap mahluk hidup yang telah dikodekan dalam bentuk pita molekul DNA yang nama latinya asam deoksiribo-nukleat (Venter dkk, 2001).

3.2 DNA dan *Gene Expression*

DNA atau *Deoxyribonucleic acid* menurut Saefudin (2007) adalah substansi dibalik adegium ‘sejenis menghasilkan sejenis’. DNA merupakan molekul paling terkenal saat ini, karena molekul ini merupakan substansi penurunan sifat. DNA merupakan suatu polimer heliks ganda yang terdiri dari nukleotida, setiap nukleotida terdiri dari tiga komponen; satu basa nitrogen, satu gula pentosa yang disebut deoksiribosa, dan satu gugus fosfat. Basa nitrogennya bisa Adenin (A), Timin (T), Guanin (G), atau Sitosin (S). Adenin dan dan guanin adalah purin, basa nitrogen dengan dua cincin organik. Sebaliknya, sitosin dan timin adalah anggota famili basa nitrogen yang dikenal sebagai pirimidin, yang mempunyai satu cincin tunggal.

Pernyataan yang telah dibuktikan dengan percobaan bahwa DNA merupakan materi genetik sel. Bukti lainnya sebelum mengalami mitosis, sel eukariotik dengan tepat menggandakan kandungan DNA-nya, dan selama mitosis, DNA ini akan terdistribusi tepat sama ke dua sel anaknya. Selain itu, kromosom diploid mempunyai DNA dua kali lebih banyak daripada kromosom haploid yang ditemukan di dalam gamet-gamet organisme yang sama. Bukti lainnya hasil penelitian Erwin Chargaff, pada tahun 1947 ia melaporkan bahwa komposisi DNA berbeda-beda antara satu spesies dengan spesies lainnya. Dalam spesies apapun yang dipilih, banyaknya keempat basa nitrogen tidaklah sama tetapi hadir dalam rasio yang khas. Chargaff juga menemukan adanya keteraturan yang agak ganjil dalam rasio dari basa-basa nukelotida ini. Dalam DNA setiap spesies yang dipelajarinya, jumlah adenin kurang lebih sama dengan jumlah timin, dan jumlah guanin kurang lebih sama dengan sitosin. Contohnya, pada DNA manusia keempat basa nitrogen hadir dalam persentase: A=30,9% dan T= 29,4%; G=19,9% dan C=19,8%. Kesamaan A=T dan G=C yang kemudian dikenal sebagai aturan Chargaff.

RNA (*Ribonucleic acid*) menurut Dale, dan Thenawijaya (1994) merupakan suatu asam ribonukleat yang terdapat dalam alur informasi genetik organisme yang berupa dogma sentral dari DNA —> RNA —> Protein, yaitu DNA ditranskripsi menjadi RNA, dan selanjutnya RNA ditranslasi menjadi protein. Di dalam sel terdapat tiga jenis RNA yaitu mRNA, tRNA dan rRNA. Dalam keadaan normal, RNA merupakan untaian tunggal yang pada kenyataannya dapat membentuk dupleks dengan membentuk ikatan hidrogen yang sama seperti DNA jika terdapat untaian yang komplemen dalam urutan basa nukleotidanya. Bentuk dupleks RNA akan mengakibatkan terhalangnya proses translasi sehingga sintesis protein terganggu, atau posttranscriptional *gene silencing* (PTGS), atau *gene silencing* (Agrawal dkk, 2003; Provost dkk, 2002).

Definisi gen semakin berevolusi seiring dengan sejarah genetika. Pada konsep menurut Mendelian, suatu gen dapat digambarkan sebagai unit penurunan sifat yang mempunyai ciri tersendiri mempengaruhi karakter

fenotipik. Morgan dan koleganya menempatkan gen- gen seperti itu pada lokus-lokus tertentu di dalam kromosom dan bahwa ahli-ahli genetik menggunakan lokus sebagai nama lain untuk gen. Kemudian para ahli lainnya melihat suatu gen sebagai daerah urutan nukleotida spesifik di sepanjang molekul DNA. Akhirnya para ahli menggunakan defenisi fungsional dari gen sebagai urutan DNA yang mengkode rantai polipeptida tertentu (Saefudin, 2007).

Untuk membuat protein tertentu, gen berperan memberi perintah pada sintesis protein yang tidak dapat secara langsung membangun protein. RNA merupakan jembatan antara dan sintesis protein Secara kimiawi RNA serupa dengan DNA yang terkecuali bahwa RNA mengandung ribosa, bukan deoksiribosa sebagai gulanya dan memiliki basa nitrogen urasil bukan timin. Oleh karna itu setiap nukleotida di sepanjang untai DNA memiliki deoksiribosa sebagai gulanya dan A, G, C, atau T sebagai basanya. Sedangkan setiap nukleotida di sepanjang untai RNA memiliki ribosa sebagai gulanya dan A, G, C, atau U sebagai basanya. Suatu molekul RNA hampir selalu terdiri dari satu untai tunggal.

Pada monomer DNA atau RNA merupakan keempat jenis nukleotida yang berbeda-beda dalam basa nitrogennya. Panjang pada gen biasanya mencapai ratusan atau ribuan nukleotida yang masing-masing basa memiliki urutan yang spesifik. Setiap polipeptida dari suatu protein juga memiliki monomer yang tersusun dalam tatanan linear tertentu, tetapi monomernya adalah dua puluh asam amino. Oleh karena itu asam nukleat dan protein berisi informasi yang ditulis dalam dua bahasa kimiawi yang berbeda. Untuk beralih dari DNA, yang ditulis dalam satu bahasa, ke protein, yang ditulis dalam bahasa lain, membutuhkan dua tahapan utama, yaitu trankripsi dan translasi.

Transkripsi adalah sintesis RNA yang didasarkan arahan DNA. Kedua asam nukleat menggunakan bahasa yang sama, dan informasinya tinggal ditranskripsi, atau disalin, dari satu molekul ke molekul yang lain. Proses kerja transkripsi dengan menyediakan suatu cetakan untuk penyusunan urutan neklotida RNA. Molekul RNA yang dihasilkan merupakan transkrip penuh

dari instruksi-instruksi pembangun protein dari gen itu. Proses inisiasi, elongasi, dan terminasi merupakan tiga tahapan transkripsi secara garis besar.

Sedangkan pengertian translasi sendiri ialah sintesis polipeptida yang sesungguhnya, yang terjadi berdasarkan arahan mRNA. Sel tersebut mampu menerjemahkan urutan basa molekul mRNA ke dalam urutan asam amino polipeptida selama tahap perubahan bahasa. Ribosom, partikel kompleks yang memfasilitasi rangkaian secara teratur pada asam amino menjadi rantai polipeptida pada tempat translasi. Sama seperti transkripsi, translasi pun memiliki tiga tahapan yakni inisiasi, elongasi, dan terminasi. tRNA (RNA transfer) adalah RNA lain yang juga berperan penting dalam proses translasi. Peran tRNA sendiri sebagai adaptor dalam sintesis protein yang mentranslasi kodon-kodon mRNA menjadi asam amino

3.3 *Microarray*

Menurut Nurfalah (2016) *microarray* merupakan teknologi yang dapat menyimpan ribuan ekspresi gen yang diambil dari beberapa jaringan manusia sekaligus. Sedangkan klasifikasi data *microarray* merupakan *supervised learning* yang mampu memprediksi kategori diagnostik dari sebuah sampel fenotipe ekspresi *array*. Untuk menghasilkan model *classifier* yang dapat mengklasifikasikan sampel data baru dengan penyakit yang ditetapkan dibutuhkan labeling sample data *gene expression*.

Pada microarray DNA sangat berperan karena mampu mengukur ribuan *gene expression* secara bersamaan. Dalam penelitian ini bertujuan menilai tingkat ekspresi sejumlah besar gen yang berbeda dalam sampel sel tertentu. Ini biasanya dilakukan untuk memutuskan gen mana yang terlibat dalam transformasi dalam fungsi sel seperti keadaan sehat atau tidak sehat. Peralatan *microarray* DNA bisa digunakan berbagai penggunaan yang luas, di mana penting untuk dapat mengukur jumlah relatif atau absolut dari sejumlah besar hasil DNA yang eksplisit, ini terdiri dari aplikasi seperti sekuensing DNA. Pada penelitian ini dilakukan expresi data gen *microarray* untuk klasifikasi penyakit HIV/AIDS yang berisi informasi gen.

3.4 HIV / AIDS

3.4.1 Definisi

HIV atau *Human Immunodeficiency Virus* merupakan virus yang dapat menyerang sel darah putih (limfosit) dalam tubuh yang berdampak turunnya system kekebalan tubuh bagi yang terinfeksi. Bagi orang yang terinfeksi virus HIV dalam darahnya terlihat sehat, namun belum dibutuhkannya pengobatan walaupun orang tersebut menularkan virusnya ke orang lain seperti penggunaan jarum suntik bersama ataupun melakukan hubungan seks. (KPAD Kab. Jember, 2015)

HIV dapat diklasifikasikan sebagai amily retroviridae, subfamily lentivirinae, dan genus lentivirus yang merupakan virus sitopatik yang terdiri dari 2 grup yakni HIV-1 dan HIV-2 dengan berat molekul 9,7 kb. Pada grup HIV-1 paling ganas di seluruh dunia dan banyak kelainan yang ditimbulkan. Apabila seseorang kontak dengan ODHA (Orang Dengan HIV-AIDS) maka akan terinfeksi olehnya. (Sofro dkk, 2011)

Seseorang yang terkena HIV belum tentu terjangkit AIDS sedangkan orang yang terkena AIDS sudah pasti terjangkit HIV (HIV adalah virusnya sedangkan AIDS merupakan penyakitnya). Pengertian AIDS yaitu sekumpulan gejala penyakit yang terjadi karena kerusakan sistem imunitas tubuh (limfosit T) yang diakibatkan oleh HIV. Jika tidak mendapatkan terapi antiretroviral (ART) dapat menyebabkan rentannya terhadap penyakit yang dikenal sebagai infeksi oportunistik (IO) yang mengakibatkan rusaknya sistem imunitas, dan sepanjang hidupnya akan menjadi infeksi sehingga dapat menularkan virus melalui cairan tubuh bagi orang yang terinfeksi. (Kummar dkk, 2015)

3.4.2 Epidemiologi

Saat ini masalah kesehatan dan sosial yang paling utama disebabkan oleh infeksi HIV/AIDS. Di Negara berkembang lebih dari 95% kasus AIDS ditemukan. Pada 20% kasus yang disebabkan penderita infeksi HIV-1 tahap lanjut diperlukannya data epidemiologik yang lebih

baik, khususnya negara berkembang karena laporan yang ada masih bervariasi (Roos dkk, 2005).

Setiap tahun jumlah yang terinfeksi HIV-1 terus meningkat dan di tahun 2010 jumlah yang terinfeksi sebanyak 34 juta jiwa menurut data WHO (WHO, 2011). meningkatnya penggunaan ART bisa dikaitkan dengan rendahnya prevelensi HAD di negara maju, tetapi penurunan angka kejadian HAD digantikan dengan peningkatan angka kejadian gangguan kognitif yang lebih ringan terkait infeksi HIV-1.(Duarte dkk, 2006).

3.4.3 Patogenesis

Pada hakikatnya HIV/AIDS dapat tertular melalui cairan tubuh yang terjangkit HIV baik dari persalinan, laktasi dari ibu yang terjangkit ke bayinya, hubungan seks dan paparan parental. (Veronique dkk, 2016).

Virus dapat menyerang sel limfosit T helper yang mengandung marker molekul CD4 yang merupakan target utama dari infeksi HIV bagi system imun. Selanjutnya HIV mengikat diri ke molukul CD4, virus akan masuk ke dalam target dan melepas bungkusnya. Selanjutnya dengan enzim reverse transcriptase virus akan merubah bentuk RNA agar bisa bergabung dengan DNA sel target. Setelah itu sel yang berkembang biak mengandung bahan genetik dan membentuk virus baru yang akan menginfeksi sel *host* lainnya. Dengan demikian infeksi HIV akan menjadi *irreversible* dan berlangsung seumur hidup (Edward, 2016).

Infeksi primer, penyebaran virus ke organ limfoid, latensi klinis, peningkatan ekspresi HIV, penyakit klinis dan kematian merupakan suatu perjalanan yang khas dari infeksi virus HIV yang terdisi dari beberapa tahapan. Pada masa awal infeksi virus HIV tidak akan menyebabkan kematian dari sel tersebut, namun diawali dengan replikasi dalm tubuh penderita yang lambat laun akan merusak limfosit T-CD4. Durasi yang dibutuhkan mulai infeksi primer sampai penyakit klinis rata-rata sekitar 10 tahun. Sejak seseorang terpapar virus sampai gejala AIDS merupakan

masa inkubasinya. Pada masa itu virus HIV tidak akan terdeteksi oleh pemeriksaan laboratorium ± 3 bulan mulai tertularnya virus. Masa inkubasi tersebut dikenal dengan *window period* (Kummar dkk, 2015)

3.4.4 Gejala Klinis HIV dan AIDS

Menurut Widoyono (2011) gejala klinis HIV/AIDS adalah sebagai berikut

a) Window period (6-8 minggu) merupakan waktu tubuh terinfeksi HIV namun belum terdeteksi pada pemeriksaan laboratorium.
b) Masa inkubasi (6 bulan-5 tahun).
c) Pada penderita HIV hanya akan bertahan sampai dengan 5 tahun dan jika tidak diobati akan bermanifestasi menjadi AIDS.
d) Gejala klinis muncul sebagai penyakit yang tidak khas seperti :
    1. Kandidiasis mulu yang luas
    2. Diare kronis
    3. Pneumonia interstisialis limfositik
    4. Ensefalopati kronik.
    5. Pneumocystis carinii

3.4.5 Cara Penularan HIV / AIDS

Menurut Nursalam dan kurniawati (2007) HIV dan AIDS dapat ditularkan dengan berbagai cara yakni :

a) Hubungan seks dengan penderita HIV/AIDS

    Apabila seseorang dengan pengidap HIV/AIDS melakukan hubungan seksual akan mengakibatkan air mani, cairan vagina, darah terkena lapisan mukosa pada selaput lender penis, vagina, dubur dan bahkan mulut. Oleh karna itu virus dapat masuk ke dalam aliran apabila terdapat lesi mikro pada kulit.
b) Darah yang terkena HIV/AIDS

    Penularan melalui darah dapat terjadi secara cepat dan menyebar langsung ke seluruh tubuh.
c) Ibu ke bayi

Pada masa kehamilan terjadi akan sangat rentan terjadi penularan dari ibu ke bayi baik proses kelahiran maupun pemberian ASI. Pada masa hamil seorang penderita memiliki resiko dapat menularkan HIV/AIDS ke bayinya sebesar 50 % dan pada masa pemberian ASI penderita memiliki resiko 10% untuk menularkan.

d) Penggunaan jarum suntik bersamaan

Salah satu potensi menularnya HIV dan AIDS yakni dengan pemakaian bersama/bergantian baik dalam fasilitas kesehatan maupun pengguna narkoba (*Injection Drug User*-IDU).

e) Alat kesehatan tidak steril

Pada penggunaan peralatan kesehatan seperti *speculum*, jarum suntik, tenakulum, ataupun katereter harus dalam keadaan steril sehingga butuh perhatian khusus bagi petugas kesehatan. Penggunaan untuk pasien HIV/AIDS haruslah satu kali sehingga tidak memiliki resiko tertularnya penyakit tersebut.

f) Peralatan untuk menorah kulit

Resiko tertularnya HIV juga dapat disebabkan pada penggunaan alat-alat menorah kulit seperti tato, perlengkapan khitan, alat cukur, dan lain-lain sehingga perlunya sterilisasi sebelum menggunakan alat tersebut.

3.4.6 Pecegahan dan Pengobatan HIV/AIDS

Sampai saat ini masuk belum adanya vaksin yang dapat mencegah atau menyembuhkan infeksi HIV/AIDS. Oleh karna itu menurut widoyono (2011), pencegahan HIV/AIDS dapat dilakukan dengan cara :

1. Pencegahan HIV/AIDS :
   a. Menghindari melakukan hubungan seksual pada penderita HIV/AIDS.
   b. Tidak melakukan hubungan seksual secara berganti-ganti.
   c. Tidak berhubungan seksual dengan pecandu narkotik dan obat suntik.

d. Melarang melakukan donor darah kepada orang yang beresiko tinggi.
e. Pemberian tranfusi darah kepada pasien yang benar-benar membutuhkan.
f. Alat suntik harus di sterilisasi sebelum digunakan.

2. Pengobatan HIV/AIDS
   a. Pemberian obat anti virus
   b. Pengobatan suportif
   c. Penanggulangan penyakit oportunistik
   d. Penanggulangan dampak psikosial

3.4.7 Penatalaksanaan

Penatalaksanaan orang yang terkena HIV/AIDS secara umum terdiri dari beberapa jenis, yakni sebagai berikut.

a. Pengobatan yang dilakukan dengan cara menekan replikasi virus HIV dengan ART.
b. Pengobatan yang dilakukan dengan cara mengatasi berbagai penyakit infeksi dan kanker yang menyertai infeksi HIV/AIDS.
c. Pengobatan dengan makanan yang memiliki nilai gizi yang bagus dan pengobatan pendukung lainnya.

Ada 4 jenis obat antiretroviral (ART) yang digunakan INI (*integrase inhibitor*), PI (*protease inhibitor*), NRTI, dan NNRTI (Veronique dkk, 2016).

Menurut Ditjen PPM dan PL (2011) pada pasien yang dinyatakan terinfeksi HIV akan dilanjutkan ke tahap penilaian stadium klinis, penilaian imunologis, dan penilaian virologi yang bertujuan menentukan apakah pasien telah memenuhi syarat untuk dilakukannya terapi antiretroviral, menilai status imun pasien, menentukan infeksi oportunistik yang pernah dan sedang terjadi, dan menentukan paduan obat antiretroviral yang sesuai.

a. Stadium klinis

Pada tahap stadium klinis untuk menentukan terapi antiretroviral dengan lebih tepat waktu harus dinilai pada saat kunjungan awal dan setiap kali kunjungan.

b. Penilaian imunologi

Untuk menilai status orang yang terkena HIV/AIDS dengan melihat jumlah CD4. Untuk memeriksa CD4 dengan cara pemeriksaan klinis untuk menentukan pasien yang memerlukan pengobatan infeksi oportunistik (IO) dan terapi antiretroviral (ART). Pada peningkatan setalah diberi antiretroviral antara 50-100 sel/mm^3/tahun terjadi penurunan rata-rata CD4 sekitar 70-100 sel/mm^3/tahun.

3.5 *Pre-processing*

Pre-processing memiliki pengertian tahap untuk penyesuaian dan konversi data *affybatch* ke dalam bentuk *expression* set. Tidak hanya itu saja, menurut Bolstad (2004) *pre-processing* tidak hanya melakukan konversi data *affybbatch*. Proses *background correction*, *summarization*, dan *normalization* merupakan tahap *pre-processing* juga. Penyesuaian berbagai macam metode yang harus dilakukan merupakan istilah pada *background correction* yang meliputi .

1) Memperbaiki efek pengolahan dan *background noise*
2) Penyesuaian estimasi ekspresi sehingga berhubungan linear dan berskala tepat
3) Penyesuaian hibridasi pengikat DNA *non*-spesifik pada *array*

Setelah tahap *background correction* yakni *normalization*. Pengertiannya sendiri yaitu tahap penghapusan akhiran *non* biologis yang tidak diinginkan diantara *chip* dalam eksperimen *microarray*. Tahap setelah *normalization* adalah *summarization* yang merupakan tahap unntuk mengukur *gene expression*.

3.6 *Filtering*

Filtering pada data *microarray* merupakan proses pemilihan *subset* dari *probe* yang tersedia untuk pengecualian atau penyertaan dalam analisis (Bourgon, 2010). Pada tahap penelitian dengan menggunakan program R, proses penyaringan menggunakan package *genefilter*. Dengan menggunakan fungsi `nsFilter` yang dapat menyediakan opsi serba ada (*one-stop shop*) untuk berbagai pilihan pada *filtering* fitur dari *expression* set. Pada fitur *filtering* sangat berguna untuk proses analisis selanjutnya dikarenakan dapat menunjukan beberapa varian kecil atau ketimpangan data seacara konsisten di seluruh sampel.

3.7 *Feature Selection*

Feature selection yakni suatu cara membuat klasifikasi menjadi efektif dan efesien dengan mengurangi jumlah data yang dianalisis, atau mempertimbangkan proses pembelajaran dengan mengidentifikasi beberapa fitur yang sesuai. Menurut Moraes, dkk (2013) untuk memilih gen yang layak digunakan untuk analisis klasifikasi dibutuhkan sebuah pendekatan yakni *feature selection* pada bidang bioinformatika seperti *T-test*, *F-test*, ANOVA, dan Wilcoxon. Namun pada penelitian ini menggunakan uji *T-test* dikarenakan data yang digunakan mempunyai dua sampel.

3.8 Uji T (*Independent Sample T-Test*)

Uji T yakni uji yang digunakan untuk mengetahui adakah perbedaan rata-rata atau mean yang bermakna dari dua kelompok bebas yang berskala interval dan rasio. Untuk menghasilkan suatu data yang paling berpengaruh dari pengolahan data diperlukan nilai α yang lebih kecil, selain itu alasan digunakan nilai α lebih kecil ialah agar lebih efisien dalam melakukan penelitian dan juga efektifitas dari komputer yang digunakan peneliti. Menurut Priyanto (2008) dua kelompok bebas yakni dua kelompok yang tidak serasi, dapat diartikan juga sampel berasal dari dua sumber data berbeda. Uji hipotesis dari uji *t-test* yang digunakan yakni :

1. Hipotesis :

   $H_0 : \mu_1 = \mu_2$ (Rata-rata dua kelompok yang sama)

H_1 : $\mu_1 \neq \mu_2$ (Rata-rata dua kelompok yang tidak sama)

2. Tingkat Signifikansi

   α = 0.0000001

3. Daerah Kritis :

   Tolak H_0 jika *P-Value* < α atau

   Tolak H_0 jika T-hitung > T-tabel

4. Statistik Uji :

$$\text{T-hitung} = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 + \sum_{i=1}^{m} Y_i^2 - m\bar{Y}^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}} \quad (3.1)$$

   Keterangan :

   $\bar{X}$ = Nilai rata-rata kelompok 1

   $\bar{Y}$ = Nilai rata-rata kelompok 2

   $\sum_{i=1}^{n} x_i^2$ = Jumlah kuadrat kelompok 1

   n = Banyaknya sampel kelompok 1

   $\sum_{i=1}^{m} y_i^2$ = Jumlah kuadrat kelompok 2

   m = Banyaknya sampel kelompok 2

5. Penarikan kesimpulan berdasarkan pengujian hipotesis.

3.9 Klasifikasi

Menurut Han dan Kamber (2006) Klasifikasi ialah sebuah proses untuk menemukan sebuah model dari fungsi untuk membedakan dan menggambarkan kelas data atau dalam pengertian lain memprediksikan kelas dari suatu objek yang labelnya tidak diketahui yang ditujukan dalam sebuah konsep. Prosesnya dengan memetakan *input* x (*predictor*) ke dalam satu kelas y yang diartikan sebelumnya. Banyaknya *record (training set)* dari setiap *record* tersebut terbentuk dari sekumpulan *predictor* yang salah satunya ialah kelas yang akan menjadi inputnya. Menurut Gorunnescu (2011) klasifikasi dibagi menjadi beberapa komponen yakni.

1. Kelas

   Merepresentasikan label yang berada dalam objek dari variabel dependent yang kategorikal.

2. *Predictor*

   Karakteristik dari atribut data hasil representasi variabel independent.

3. *Training* dataset

   Hasil dari komponen kelas dan *predictor* yang menghasilkan satu set data yang bertujuan menentukan kelas yang cocok dari *predictor*.

4. *Testing* Dataset

   Klasifikasi model yang sudah dibuat memiliki akurasi klasifikasi yang telah dievaluasi berdasarkan data baru.

Dalam mengklasifikasikan suatu data dapat dilalui dengan beberapa tahapan yang terbagi menjadi 3 bagian yaitu :

1. Pembangunan Model

   Pada tahap ini model yang dibangun berdasarkan *training* set dengan membuat sebuah model supaya dapat menyelesaikan klasifikasi kelas atau atribut dalam data.

2. Penerapan Model

   Pada tahap ini model yang telah dibuat sebelumnya dapat digunakan untuk menentukan *predictor* (variabel x) dari data baru yang *predictor* nya belum diketahui. Tujuannya supaya dapat diterapkan hasil dari pembangunan atau pengembangan dan menggunakan algoritma, struktur data, dan visualisasi yang efisien. Sehingga dapat menggambarkan kelas prediksi dari objek yang label kelasnya telah diketahui.

3. Evaluasi

   Pada tahap ini model yang sudah di pakai akan di evaluasi dengan parameter terukur agar dapat ditentukan apakah model layak digunakan atau tidak.

3.10 *Support Vector Machine* (SVM)

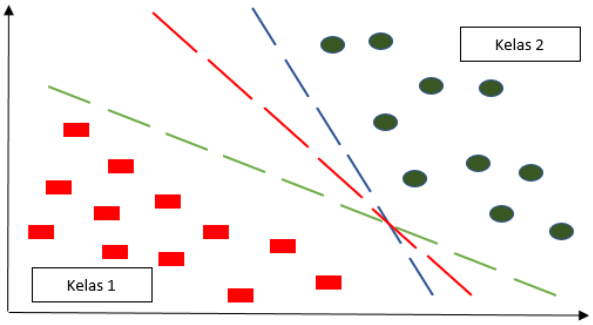
3.10.1 Pendahuluan

SVM yang kepanjangannya *Support Vector Machine* adalah sistem pembelajaran terbimbing yang pengklasifikasiannya menggunakan ruang hipotesis berupa fungsi-fungsi linear dalah sebuah ruang fitur berdimensi

tinggi (Santosa, 2007). SVM pertama kali diperkanalkan pada tahun 1992 oleh Vapnik sebagai rangkaian dari beberapa konsep – konsep unggulan dalam bidang *pattern recognition*. (Elly Susilowati, 2015). Menurut Siagian (2011) untuk penggunaan fungsi *kernel* dan parameter tergantung dari tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan dengan SVM. Metode SVM dibagi menjadi dua berdasarkan karakteristiknya yakni SVM linear dan SVM *non* linear.(F R, 2012).

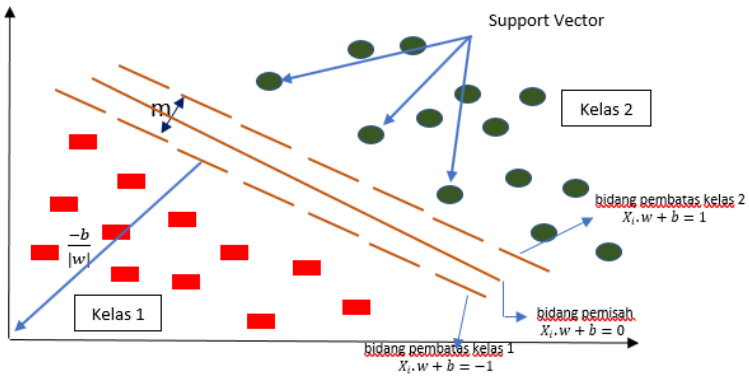
3.10.2 SVM Linear

Secara sederhana SVM bisa diartikan usaha mendapatkan *hyperlane* terbaik yang memiliki fungsi sebagai pemisah dari dua buah kelas pada ruang *input (input space).*



Gambar 3.1 Klasifikasi SVM linear

Menurut Nomleni (2015) yang tercantum dalam sebuah tesisnya, pada gambar diatas terlihat dua kelas yaitu +1 dan -1 yang masing-masing *pattern* yang digambarkan *patern* -1 berwarna merah dan *pattern* +1 berwarna hijau.



Gambar 3.2 Klasifikasi SVM linear

Pada Gambar 3.2 diatas, untuk mendapatkan hasil klasifikasi yang baik pada *hyperlane* yang digunakan dengan memisahkannya menjadi dua kelas yakni dengan mengukur *margin hyperlane* tersebut dan mencari titik maksimalnya. Pengertian *margin* sendiri ialah jarak *hyperlane* terdekat dengan *pattern* terdekat dari masing-masing kelas. Sedangkan *support vector* adalah *pattern* yang paling dekat dengan *hyperlane*. Pada gambar 3.2 diatas terlihat garis menyambung di tengah-tengah kedua kelas, sedangkan *pattern* yang berpotongan dengan garis yang putus-putus merupakan *support vector*. Berdasarkan gambar 3.2 bidang pemisah dirumuskan sebagai berikut :

m = jarak antara dua bidang

w = bidang normal

b = posisi *relative* terhadap origin

wx+b rumusan dari jarak garis sedangkan (c-b)/|w| rumusan origin

$$m = \frac{1-b-(-1-b)}{|w|} = \frac{2}{|w|} \tag{3.2}$$

Untuk memenuhi konstrain 2 bidang pembatas yang sejajar dengan memaksimalkan *margin* m dan *support vector* bisa disebut data yang ada pada bidang pembatas. Bidang pembatas pada kelas pertama membatasi kelas pertama dan bidang pembatas kelas kedua membatasi kelas kedua, sehingga diperoleh persamaan :

$$x_i.w + b \geq +1,\ y_i = +1$$

$$x_i.w + b \leq -1,\ y_i = -1 \tag{3.3}$$

Pada notasi (3.2) dan (3.3) nilai maksimal harus memenuhi dan untuk menghasilkan nilai *margin* yang dikalikan dengan konstanta yang sama perlu mengkalikan sebuah konstanta dengan nilai w dan b. pengertian konstrain sendiri ialah *scaling constraint* dengan dipenuhi *rescaling* w dan b. oleh karena itu untuk memaksimalkan dan meminimalkan w perlunya sebuah rumusan pertidaksamaan yakni :

$$y_i\,(x_i.w + b) - 1 \geq 0 \tag{3.4}$$

Untuk menghasilkan nilai m yang kemudian dikalikan dengan konstanta perlu mengkalikan b dan sebuah konstanta. *Scaling constraint* yang dipenuhi dengan *rescaling* w dan b pengertian dari konstrain. Maksimalkan $\frac{1}{|w|}$ = minimumkan $|w|^2$.

Merumuskan masalah optimasi konstrain dengan cara mencari nilai margin terbesar untuk pion pemisah terbaik, yakni :

$$min_{w,b}^{Lp} \text{ (w,b,a)} = \frac{1}{|2|} |w|^2 - \sum_{i=1}^{n} \alpha_i y_i \, (x_i . w + b) + \sum_{i=1}^{n} \alpha_i \qquad (3.5)$$

Tambahan konstrain $\alpha_i \geq 0$ (nilai dari koefisien lagrange). Dengan meminimalkan Lp terhadap w dan b, maka dari $\frac{\partial}{\partial b} Lp$ (w,b,a) = 0

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (3.6)$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad (3.7)$$

Vektor w sering bernilai tak hingga bahkan besar. Tetapi nilai α_i, maka formula lagrangian Lp (*primal problem*) harus diubah kedalam *dual problem* LD dengan mendistribusikan persamaan (3.7) ke Lp untuk memperoleh *dual problem* dengan konstrain berbeda.

$$Lp\,(\alpha) = \alpha - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \, \alpha_j y_i y_j x_i x_j \qquad (3.8)$$

Sehingga dapat dirumuskan permasalahan bidang pemisah terbaik, yakni :

$$max_{\alpha} L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \, \alpha_j y_i y_j x_i x_j$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0, \alpha_i \geq 0 \qquad (3.9)$$

Agar mendapatkan α_i untuk setiap data yang digunakan untuk pelatihan dengan cara menemukan nilai α_i yang digunakan untuk menemukan w. data *support vector* yakni data yang untuk pelatihannya mempunyai nilai α_i > 0, sedangkan sisanya yang memiliki nilai α_i = 0. Sehingga *support vector* mempengaruhi fungsi untuk keputusan yang dihasilkan.

Formula pencarian untuk bidang pemisah terbaik adalah masalah *quadratic programming*, sehingga agar selalu ditemukan nilai *maksimum global* dari α_i dengan setelah menemukan solusi permasalahan *quadratic programing* ditemukan (nilai α_i), sehingga kelas dari data pengujian x dapat ditemukan berdasarkan nilai dari fungsi keputusan.

$$f(X_d\,)\sum_{i=1}^{nx}\alpha_i\,y_i\,x_i\,x_d\;+b \tag{3.10}$$

Dapat dirumuskan :

x_i = *support vector*

nx = jumlah *support vector*

x_d = data yang diklasifikasikan

3.10.3 Soft Margin

Dua buah kelas tidak selalu terpisahkan secara sempurna yang didasarkan penjelasan SVM linear dengan asumsi bahwa kedua kelas terpisah secara sempurna oleh *hyperlane*. Pada tulisan Cahyo (2018). Untuk optimasi yang tidak diperlukan disebabkan oleh *constraint*, oleh karna itu untuk mengatasi permasalahan ini diperlukan penggunaan teknik *soft margin* untuk merumuskan ulang SVM. Dengan memodifikasi *soft margin* yakni memasukan *variabel slack* ξ_i ($\xi_i > 0$) sebagai berikut.

$$y_i(x_i.w+b)\geq 1-\;\xi_i \tag{3.11}$$

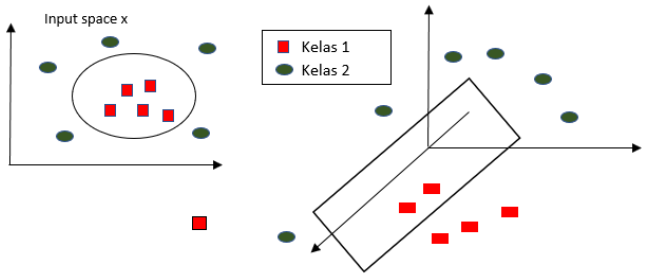
Dengan demikian pada persamaan (3.11) akan berubah menjadi

$$min_w=\tau(w)=\frac{1}{2\|w\|^2}+C\;\sum_{i=1}^{l}\varepsilon_i \tag{3.12}$$

Untuk mengontrol *trade off* antara *margin* dan *error* klasifikasi ξ pada parameter C. Nilai besar C akan memberikan *penalty* yang lebih besar terhadap *error* klasifikasi tersebut.

3.10.4 SVM *Non* Linear

Pada klasifikasi *non* linear, data yang berada pada sebuah ruang vektor awal haruslah dipindahkan ke dalam ruang vektor baru yang dimensinya lebih tinggi. Sebagai contoh fungsi pemetaan dinotasikan sebagai $\Phi(x)$. Mempresentasikan data ke format yang dipisahkan secara linear pada ruang vektor baru merupakan tujuan dari pemetaan (Nomleni (2015).



Gambar 3.3 Klasifikasi SVM *non* linear

Pada Gambar 3.3 diatas, diperlukannya sebuah perhitungan *dot product* dua buah contoh pada ruang vektor baru pada fase tersebut. Dinotasikan dengan $\Phi(x_i)$, $\Phi(x_j)$ untuk *dot product* kedua buah *vector* x_i dan x_j. Secara tidak langsung nilai *dot product* kedua vektor tersebut bisa dihitung yakni dengan cara tidak mengetahui proses transformasi Φ.

Dengan memodifikasi formula yang menyelesaikan permasalahan tersebut karena tidak ada solusi yang lain merupakan sebuah cara melakukan klasifikasi yang tidak bisa dipisahkan secara linear SVM. Pada dua bidang pembatas perlu dirubah agar tidak kaku dalam sebuah kondisi tertentu dengan penambahan variabel (ξ $\xi_i \geq 0 \forall_i$: ξ =0) jika diklasifikasikan dengan benar menjadi $x_i.w + b \geq 1 - \xi_i$ untuk kelas 1 dan $x_i.w + b \geq 1 + \xi_i$ untuk kelas 2. Untuk menambahkan ξ_i yang disebut *soft margin hyperlane* perlu mencari bidang pemisah terbaik, dan formula untuk mencari bidang pemisah terbaik berubah menjadi.

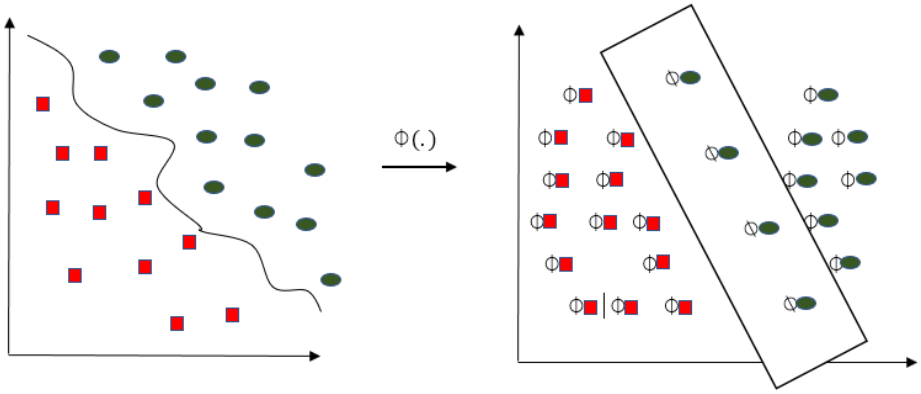
$$min \frac{1}{2} |w|^2 + (\sum_{i=1}^{n} \xi_i)$$

$$s.t. \quad y_i \ (w.x_i + b) \geq - \xi_i$$

$$\xi \geq 0 \qquad (3.13)$$

C adalah parameter yang menentukan besarnya penalti yang disebabkan oleh kesalahan klasifikasi data dan pengguna biasanya menentukan nilai. Pada permasalahan (3.14) memenuhi prinsip dari SVM yang meminimumkan $\frac{1}{2}\ |w|^2$ akuivalen dengan meminimumkan dimensi VC dan meminimumkan C($\sum_{i=1}^{0} \xi_i$) yang berarti meminimlkan *error* pada data pelatihan yang merubah bentuk primal problem menjadi.

$$min_{w.b} L_p\ (w,b,a) = \frac{1}{2}\,|\mathrm{w}|^2 + \mathrm{C}\ (\sum_{i=1}^{n}) - \sum_{i=1}^{n} \alpha_i\ \{y_i\ (x_i.\mathrm{w}+\mathrm{b}) - 1 + \xi_i\} - \sum_{i=1}^{n} \mu_i\ \xi_i \tag{3.14}$$



Gambar 3.4 Transformasi dari vektor input ke feature space

Pada Gambar 3.4 diatas bisa dilihat data dipetakan dengan menggunakan fungsi pemetaan atau transformasi caranya. Dengan menggunakan fungsi pemetaan data dipetakan ke dalam fitur ruang $x_k \rightarrow \Phi(x_k)$ sehingga bidang pemisah dapat memisahkan data sesuai dengan kelasnya. Sebagai contoh *feature space* merupakan data set yang mempunyai dua kelas dan atribut yakni kelas positif dan kelas negatif yang memisahkan data dengan kelasnya. Contoh data yang memiliki kelas positif (2,2),(2,-2),(-2,2),(-2,-2) dan data yang memiliki kelas negative (1,1),(1,-1),(-1,1),(-1,-1). Data tidak bisa dipisahkan secara linear jika digambarkan dalam ruang dua dimensi. Oleh karna itu digunakan fungsi tranformasi yakni.

$$\Phi(x_1,x_2)\begin{cases} \sqrt{x_1^2} + \sqrt{x_2^2} > 2 \rightarrow (4 - x_2 + |x_1 - x_2|, 4 - x_1 + |x_1 - x_2|) \\ \sqrt{x_1^2 x_2^2} \leq 2 \rightarrow (x_1, x_2) \end{cases} \tag{3.15}$$

Hasil transformasi data adalah data sesudah di transformasi {(2,2),(6,2),(6,6),(2,6)} untuk kelas negatif dan {(1,1), (1,-1), (-1,1), (-1,-1)} lalu dicarikan bidang pemisah terbaik dari data ini. Penggunaan fungsi transformasi $x_k \rightarrow \phi(x_k)$ maka nilai $w = \sum_{i=1}^{nx} \alpha_i y_i \phi(x_i)$ hasil pembelajarannya yakni.

$$f(X_d) = \sum_{i=1}^{nx} \alpha_i y_i \phi(x_i) \phi(x_d) + b \tag{3.16}$$

Ruang fitur dapat mempunyai jumlah fitur yang tak terhingga dan sulit mengetahui fungsi transformasi yang tepat sehingga SVM perlu menggunakan *kernel trick* dari persamaan (3.16) terlihat *dot product* $\phi(x_i)\phi(x_d)$. Jika mempunyai sebuah fungsi *kernel* K yang jika terdapat fungsi *kernel* K maka $K(x_i, x_d) = \phi(x_i).\phi(x_d)$, fungsi yang berasal dari transformasi $\phi(x_k)$ diabaikan atau tidak perlu di (3.17) secara jelas. Fungsi dari pelatihan yang dihasilkan yakni.

$$f(X_d) = \sum_{i=1}^{nx} \alpha_i y_i K(x_i, x_d) + b \ (x_i = support\ vector) \tag{3.17}$$

- *Kernel* Linear

$$K(x_i, x) = x_i^T x \tag{3.18}$$

- Polynomial *Kernel*

$$K(x_i, x) = (\gamma . x_i^T x + r)^p, \gamma > 0 \tag{3.19}$$

- Radial Basis Function (RBF)

$$K(x_i, x) = \exp(-\gamma |x_i - x|^2), \gamma > 0 \tag{3.20}$$

- Sigmoid *Kernel*

$$K(x_i, x) = \tanh(\gamma x_i^T\ x + r) \tag{3.21}$$

Confusion Matrix

Confusion matrix bertujuan memperkirakan suatu objek yang salah atau benar dalam suatu pengujian. Menurut Gorunnescu (2011) hasil dari *confusion matrix* ialah tabel yang bertujuan mengklasifikasikan hasil dari set data uji dimana

kelas yang diprediksi pada tabulasi urutan pengujian pada confusion matrix terletak pada bagian atas matriks dan dibagian kiri ialah kelas yang akan diamati. Setiap sel yang berada di dalam tabel berisi angka yang memperlihatkan banyaknya kasus yang sebenarnya dari kelas yang akan diamati untuk diprediksi.

Tabel 3.1 *Confusion Matrix* dimensi 2x2

		Aktual Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

Keterangan :

- TP : Jumlah data positif yang benar pada hasil klasifikasi
- FP : Jumlah data positif yang salah pada hasil klasifikasi
- FN : Jumlah data negatif yang benar pada hasil klasifikasi
- TN : Jumlah data negatif yang salah pada hasil klasifikasi

Nilai pada akurasi, presisi, dan juga *recall* merupakan hasil dari *confusion matrix*. Rumus penentuan model dalam menghitung nilai akurasi, presisi, dan juga *recall* yakni :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.22}$$

Presisi ialah semua item yang terpilih berdasarkan rasio item relevan yang dipilih. Presisi bisa juga diartikan kecocokan permintaan informasi dan jawaban dari permintaan informasi. Rumus yang digunakan ialah :

$$Presisi = \frac{TP}{FP + TP} \tag{3.23}$$

Recall ialah rasio item relevan yang terpilih dari hasil jumlah item relevan yang tersedia. Rumus yang digunakan ialah :

$$Recall = \frac{Tp}{TP + FN} \tag{3.24}$$

Presisi dan *recall* ialah hitungan persentase yang digunakan dari nilai dalam bentuk angka. Hasil dikatakann baik jika nilai presisi dan *recall* tinggi.

Selain itu ada perhitungan lain yakni *specificity*, FPR, dan AUC untuk pembentukan model klasifikasi yakni.

$$\textit{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \times 100\% \quad (3.25)$$

$$\mathrm{FPR} = 1 - \mathrm{Specificity} \quad (3.26)$$

$$\mathrm{AUC} = \frac{1+\mathrm{Recall}-\mathrm{FPR}}{2} \quad (3.27)$$

Menurut Sofi & Jajuli (2015) AUC ialah kinerja diskriminatif untuk memperkirakan peluang output dari sampel yang dipilih dengan acak dari populasi positif atau negatif. Hasil nilai AUC selalu diantara 0-1, makin besar nilI AUC yang terbentuk makin kuat klasifikasi yang dipakai, acuan untuk klasifikasi AUC yakni.

Tabel 3.2 Klasifikasi AUC

Nilai AUC	Keterangan
0.91-1.00	Klasifikasi Sangat Baik
0.81-0.90	Klasifikasi Baik
0,71-0.80	Klasifikasi Cukup
0.61-0.70	Klasifikasi Buruk
≤ 0.60	Klasifikasi Salah