BAB II
KAJIAN LITERATUR

2.1 Kanker Payudara

Kanker adalah penyakit yang disebabkan oleh pertumbuhan sel yang tidak normal. Sel-sel ini ada karena perubahan ekspresi gen, maka kanker akan berkembang menjadi populasi sel yang dapat menyerang jaringan tertentu (Ruddon R, 2007). Perubahan penampilan gen yang meluas ke dalam sel dapat menyebabkan pergeseran fungsional sel-sel ini. Ini sangat berbahaya karena dapat menyebabkan kematian. Berdasarkan data statistik Global Cancer (GLOBOCAN) bagian dari Badan Penelitian Internasional tentang Kanker (IARC) pada tahun 2018, ada 18,1 juta kasus kanker di dunia dan 9,6 juta di antaranya telah meninggal. Dalam 18,1 juta kasus kanker, kasus kanker yang paling umum dialami oleh pria adalah kasus kanker prostat, sedangkan kasus kanker yang paling umum dialami oleh wanita adalah kasus kanker payudara (IARC, 2018).

Kanker Payudara adalah jenis kanker di mana ada pertumbuhan sel kanker yang tidak terkendali yang terbentuk di jaringan payudara. Pertumbuhan sel kanker akan membentuk benjolan yang dapat menyebar ke jaringan lain di dalam tubuh, yang juga dikenal sebagai tumor ganas. Sebagian besar kanker di payudara mulai tumbuh di kelenjar untuk produksi susu yang disebut lobulus, dan di saluran yang menghubungkan lobulus dengan puting **(** NCBI, 2019 **).** Kanker payudara terbagi atas 2 yaitu *malignant* (ganas) atau *benign* (jinak). Tumor ganas berkembang ke sel-sel tetangga, yang dapat menyebabkan metastasis atau mencapai bagian lain, sedangkan tumor jinak tidak dapat berkembang ke jaringan lain, ekspansi kemudian hanya terbatas pada tumor jinak (L.A. Altone et al, 1998)

Faktor-faktor yang menyebabkan kanker payudara menurut *American Cancer Society* dalam (Makhfudhoh, 2014) adalah: usia, jenis kelamin, riwayat reproduksi, riwayat keluarga, *obesitas*, dan konsumsi makanan lemak tinggi. Menurut *Frank, A*. dan *Asuncion* dalam (Hermawanti, 2015) parameter-parameter kanker payudara sebagai berikut:

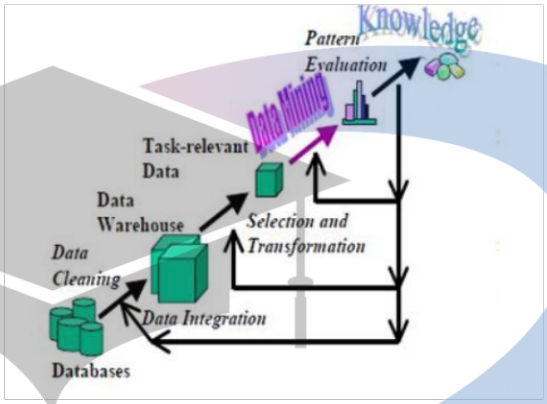
1. *Clump thickness*: sel benign cenderung dikelompokkan dalam *monolayers*, sementara sel-sel kanker sering dikelompokkan dalam multilayers.
2. *Uniformity of cell size*: sel-sel kanker mempunyai ukuran bervariasi.
3. *Uniformity of cell shape*: sel-sel kanker mempunyai bentuk bervariasi.
4. *Marginal adhesion*: sel-sel normal cenderung tetap bersama-sama.
5. *Single epithelial cell size*: sel-sel epitel yang signifikan diperbesar menjadi sel *malignant*.
6. *Bare nuclei*: adalah istilah yang digunakan untuk inti (nuclei) yang tidak dikelilingi oleh *cytoplasm* (seluruh sel). Biasanya terlihat di *benign*.
7. *Bland Chromatin*: inti "tekstur" seragam yang dilihat dalam sel benign. Dalam sel-sel kanker chromatin cenderung lebih kasar.
8. *Normal nucleoli*: *nucleoli* adalah struktur kecil yang terlihat dalam inti atom. Pada sel-sel normal nucleolus biasanya sangat kecil jika terlihat sama sekali. Dalam sel-sel cancer nucleoli menjadi lebih menonjol.
9. *Mitoses* : pembelahan satu sel menjadi dua sel.
10. *Class* : kelas

2.2 Data Mining

Data Mining adalah ekstraksi informasi yang implisit, sebelumnya tidak diketahui, dan berpotensi berguna dari data. Idenya adalah untuk membangun program komputer yang menyaring melalui database secara otomatis, mencari keteraturan atau pola. Pola yang kuat, jika ditemukan, kemungkinan akan menggeneralisasi untuk membuat prediksi akurat pada data masa depan. Apa pun yang ditemukan akan tidak tepat tetapi akan ada pengecualian untuk setiap aturan dan kasus yang tidak dicakup oleh aturan apa pun. Algoritma harus cukup kuat untuk mengatasi data yang tidak sempurna dan mengekstrak keteraturan yang tidak tepat tetapi bermanfaat (Ian & Eibe 2005).

Data Mining adalah penggunaan teknik analisis data secara otomatis untuk mengungkap hubungan yang sebelumnya tidak terdeteksi di antara item data. *Data Mining* sering melibatkan analisis data yang disimpan dalam data warehouse. Tiga dari teknik *Data Mining* utama adalah regresi, klasifikasi dan pengelompokan data. Data Mining, juga dikenal sebagai Knowledge Discovery in

Databases (KDD), mengacu pada ekstraksi nontrivial dari informasi implisit, informasi yang sebelumnya tidak diketahui dan berpotensi bermanfaat dari data dalam database. Sementara data mining dan penemuan pengetahuan dalam database sering diperlakukan sebagai sinonim, data mining sebenarnya merupakan bagian dari proses penemuan pengetahuan. Gambar 2.1 menunjukkan *Data Mining* sebagai langkah dalam proses penemuan pengetahuan iteratif.



Gambar 2. 1 *Data mining* merupakan inti dari Proses penemuan Pengetahuan (*IJCTEE,* 2015)

Tahap-tahap *data mining* adalah sebagai berikut

a. Pembersihan data (*Data cleaning*): juga dikenal sebagai pembersihan data, ini adalah fase di mana data kebisingan dan data yang tidak relevan dihapus dari koleksi.
b. Integrasi data (*Data integration*) : pada tahap ini, banyak sumber data, seringkali heterogen, dapat digabungkan dalam sumber yang sama.
c. Pemilihan data (*Data selection*): pada langkah ini, data yang relevan dengan analisis diputuskan dan diambil dari pengumpulan data.
d. Transformasi data (*Data transformation*): juga dikenal sebagai konsolidasi data, ini adalah fase di mana data yang dipilih diubah menjadi formulir yang sesuai untuk prosedur penambangan.
e. Penambangan data (*Data mining*): ini adalah langkah penting di mana teknik cerdas diterapkan untuk mengekstraksi pola yang berpotensi bermanfaat.
f. Evaluasi pola (*Pattern evaluation*): pada langkah ini, pola yang sangat menarik yang mewakili pengetahuan diidentifikasi berdasarkan tindakan yang diberikan.

g. Representasi pengetahuan (*Knowledge representation*): adalah fase terakhir di mana pengetahuan yang ditemukan secara visual diberikan kepada pengguna. Langkah ini menggunakan teknik visualisasi untuk membantu pengguna memahami dan menafsirkan hasil penambangan data (*data mining*).

Konsep pembelajaran dalam *Data mining* terbagi menjadi 2 jenis (Maimon & lior 2010), yaitu pertama *Supervised Learning* yang merupakan algoritma yang membuat suatu fungsi berdasarkan data latihan yang sudah ada, dalam hal ini dapat dikatakan untuk algoritma ini sudah tersedia data latihan secara lengkap dan detil dan terklasifikasi dengan baik yang akan dijadikan sebuah model data saat dilakukan proses ujicoba dengan data tes yang baru dan menghasilkan hasil keluaran yang sesuai diharapkan sebelumnya berdasarkan data latihan yang ada. Kedua adalah *Unsupervised Learning* yang merupakan algoritma yang melakukan representasi atau mewakili pola sebuah input yang berasal dari data latihan dan yang menjadi salah satu perbedaan dengan *Supervised Learning* adalah tidak adanya pengklasifikasian dari input data

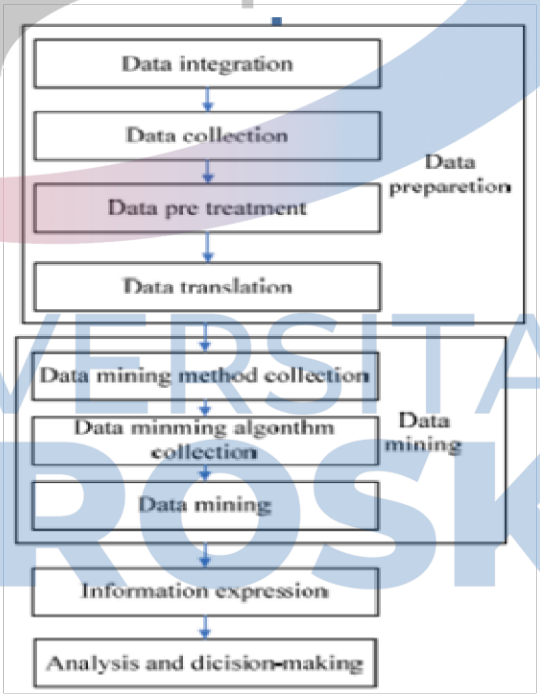
Dalam (Fayyad et al, 1996) Data mining terdiri dari 4 kelompok :

1. ***Clustering*** adalah tugas menemukan kelompok dan struktur dalam data yang dalam beberapa cara "serupa", tanpa menggunakan struktur yang diketahui dalam data. *Clustering* adalah penambangan data (pembelajaran mesin) teknik yang digunakan untuk menempatkan elemen data ke dalam kelompok terkait tanpa pengetahuan sebelumnya tentang definisi kelompok. Teknik pengelompokan yang populer termasuk pengelompokan *K-Means* dan pengelompokan maksimalisasi (EM).
2. **Klasifikasi** adalah tugas untuk menggeneralisasi struktur yang diketahui dan diterapkan pada data baru. Misalnya, program email mungkin mencoba untuk mengklasifikasikan email sebagai sah atau spam. Algoritma yang umum digunakan adalah *decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.*
3. **Regresi** adalah teknik data mining (pembelajaran mesin) yang digunakan untuk mencocokkan suatu persamaan dengan dataset. Bentuk paling sederhana dari regresi, regresi linier, menggunakan rumus garis lurus (y = mx + b) dan menentukan nilai yang sesuai untuk m dan b untuk memprediksi nilai y berdasarkan nilai x yang diberikan. Teknik-teknik canggih, seperti regresi

berganda, memungkinkan penggunaan lebih dari satu variabel input dan memungkinkan pemasangan model yang lebih kompleks, seperti persamaan kuadratik. Keterbatasan dari teknik ini adalah bahwa teknik ini hanya berfungsi dengan baik dengan data kuantitatif terus menerus (seperti berat, kecepatan atau usia).

4. ***Association rule learning*** yaitu mencari hubungan antara variabel. Misalnya supermarket mengumpulkan data tentang kebiasaan pembelian pelanggan. Menggunakan pembelajaran aturan asosiasi, supermarket dapat menentukan produk mana yang sering dibeli bersama dan informasi ini digunakan untuk tujuan pemasaran yang disebut sebagai analisis keranjang pasar.

Secara umum, proses dari data mining adalah : persiapan data, penambangan data, dan ekspresi informasi dan analisis tahap pengambilan keputusan, proses secara umum dapat dilihat pada gambar berikut :



Gambar 2.2 Proses data mining secara umum (*IJCTEE,* 2015)

2.3 Klasifikasi

Klasifikasi adalah salah satu teknik *Data Mining* yang digunakan untuk menganalisis dataset yang diberikan dan mengambil setiap instance dari itu dan menetapkan instance ini ke kelas tertentu sehingga kesalahan klasifikasi akan

menjadi paling sedikit. Klasifikasi ini digunakan untuk mengekstrak model yang secara akurat mendefinisikan kelas data penting dalam dataset yang diberikan. Ada 2 tahap dalam proses klasifikasi, langkah pertama model dibuat dengan menerapkan algoritma klasifikasi pada set data pelatihan dan pada langkah kedua model yang diekstraksi diuji terhadap dataset uji yang telah ditentukan sebelumnya untuk mengukur kinerja dan akurasi model yang dilatih oleh model. Jadi klasifikasi adalah proses untuk menetapkan label kelas dari dataset yang label kelasnya tidak diketahui (Nikam, *Orient. J. Comp. Sci. & Technol*, 2015). Ada banyak jenis algoritma yang digunakan untuk melakukan klasifikasi data,diantaranya : ID3 *Algorithm*, C4.5 *Algorithm*, *K Nearest Neighbors Algorithm*, *Naïve Bayes Algorithm*, ANN *Algorithm*, SVM *Algorithm*.

2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah sistem pembelajaran untuk mengklasifikasikan data menjadi dua kelompok data yang menggunakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang fitur (*feature space*) berdimensi tinggi. SVM memiliki sifat yang tidak dimiliki oleh mesin pembelajaran pada umumnya yaitu dalam proses menemukan garis pemisah (*hyperplane*) terbaik sehingga diperoleh ukuran margin yang maksimal antara ruang *input* bukanlinear dengan ruang ciri menggunakan kaidah *kernel* (Cortes & Vapnik 1995). *Margin* adalah dua kali jarak antara *hyperplane* dengan *support vector*. Titik yang terdekat dengan *hyperplane* disebut *support vector*.

Dalam apa yang diasumsikan maka diberikan satu set S dari poin $x_i \epsilon \square^n$ IRn dengan i = 1, 2, ..., N. Setiap poin xi milik salah satu dari dua kelas dan dengan demikian diberi label $Y_i \epsilon\{1, -1\}$. Tujuannya adalah untuk menetapkan persamaan *hyperplane* yang membagi S meninggalkan semua titik dari kelas yang sama di sisi yang sama sambil memaksimalkan jarak minimum antara salah satu dari dua kelas dan *hyperplane*. Untuk tujuan ini diperlukan beberapa defenisi awal (Pontil&Verri 1997).

Pertama,himpunan S secara linier dapat dipisahkan jika ada $w \epsilon \square^n$dan $b \epsilon \square$ sedemikian rupa

$$x_i.w + b \geq +1 \; if \; y_i = +1 \quad (1)$$

$$x_i.w + b \leq -1 \; if \; y_i = -1 \quad (2)$$

dengan

w = vektor bobot yang tegak lurus terhadap *hyperplane* (bidang normal)

b = posisi bidang relatif terhadap pusat koordinat

Dalam notasi yang lebih sederhana, kedua pertidaksamaan di atas dapat ditulis ulang

$$Y_i(w.x_i + b) \geq 1, \quad (3)$$

Untuk*i*=1,2,…., N. Pasangan (w, b) menunjukkan hyperplane persamaan

$$w.x + b = 0 \quad (4)$$

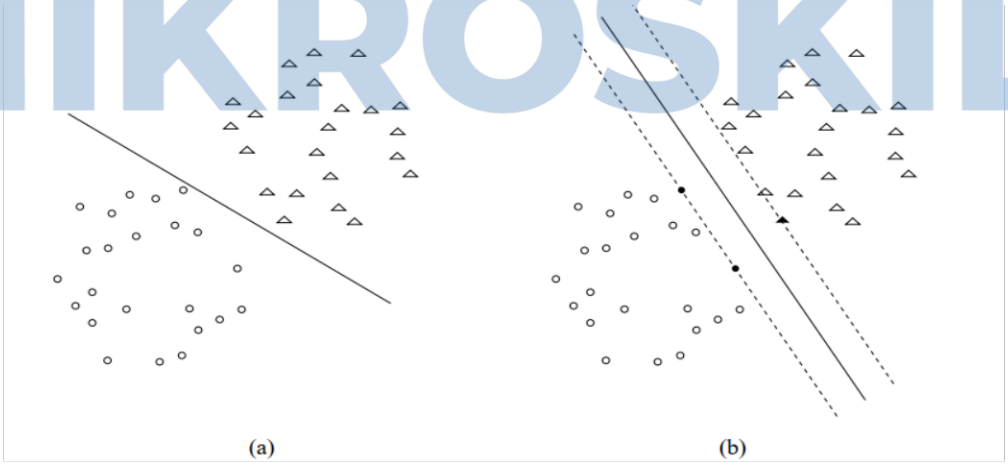
Dinamakan dengan *separating hyperplane.* Jika dilambangkan dengan ω yang berarti w, jarak yang ditandai $d_i$ditikikx_i dari hyperplane pemisah (w, b) diberikan oleh

$$d_i = \frac{w.x_i+b}{\omega} \quad (5)$$

Kombinasi dari pertidaksamaan dan persamaan diatas untuk seluruh $x_i \in S$, maka

$$y_i d_i \geq \frac{1}{\omega} \quad (6)$$

Oleh karena itu,1/ ω adalah batas bawah pada jarak antara titik x_i dan hyperplane pemisah (w, b).



Gambar 2.3.Memisahkan hyperplane (a) dan hyperplane pemisah yang optimal(b).

Kedua, diberikan hyperplane pemisah (w, b) untuk himpunan yang dapat dipisahkan secara linier S, *Canonical Representation* dari hyperplane pemisah diperoleh dengan mengubah ukuran pasangan (w, b) kedalam pasangan (w ', b') sedemikian rupa sehingga jarak titik terdekat sama dengan 1 / w '.

Melalui defenisi ini diperoleh

$$min_{x_i \epsilon S}\{y_i(w'.x_i + b)\} = 1 \quad (7)$$

Ketiga, dengan set S yang dapat dipisahkan secara linear, *optimal separating hyperplane* (OSH) adalah hyperplane pemisah yang memaksimalkan jarak titik terdekat S. Karena jarak dari titik terdeka t sama dengan 1 / w OSH dapat dianggap sebagai solusi dari masalah memaksimalkan 1 / w berdasarkan

Problem**P1**

Minimize $\frac{1}{2}w.w$

Subject to$y_i(w.x_i + b) \geq, i = 1, 2, \ldots, N$ (8)

Masalah P1 dapat diselesaikan dengan menggunakan metode klasik pengganda *Lagrange* (Bazaraa*et al.* 2006). Jika dinyatakan dengan α = (α1, α2,……,αn) pengganda N *Lagrange*non negatif yang terkait dengan kendala, solusi untuk masalah P1 setara dengan menentukan titik pelana fungsi.

$$L = \frac{1}{2}w.w - \sum_{i=1}^{N} \alpha_i\{y_i((w.x_i + b) - 1)\} \quad (9)$$

Dengan L= L(w, b, α). Pada titik pelana, L memiliki nilai minimum untuk w = $\bar{w}$dan b = $\bar{b}$ dan maksimum untuk α=$\bar{\alpha}$, dan dengan demikian dapat dituliskan sebagai berikut:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{N} y_i\alpha_i = 0 \quad (10)$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \quad (11)$$

dengan

$$\frac{\partial L}{\partial w} = (\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \ldots, \frac{\partial L}{\partial w_N}) \quad (12)$$

Persamaan diatas dapat dimodifikasi sebagai maksimalisasi L yang hanya mengandung α_i sebagai persamaan dibawah ini.

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i . x_j, \qquad (13)$$

Berdasarkan persamaan di atas dengan $\alpha >= 0$[2]. Ini adalah permasalahan yang baru yang disebut dengan *dual problem.* Dan dapat dituliskan sebagai berikut:

Problem **P2**

$$maximize -\frac{1}{2}\alpha . D\alpha + \sum \alpha$$

$$s.t. \sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geq 0 \qquad (14)$$

Dimana kedua persamaan di atas untuk i=1, 2, …, N,dan D adalah matriks N x N seperti berikut:

$$D_{ij} = y_i y_j x_i . x_j \qquad (15)$$

Dan untuk pasangan $(\bar{w}, \bar{b})$, dari persamaan *dua lproblem* mengikuti persamaan

$$\bar{w} = \sum_{i=1}^{N} \bar{\alpha} y_i x_i, \qquad (16)$$

selama$\bar{b}$ dapat ditentukan dari kondisi Kuhn-Tucker

$$\bar{\alpha}\left(y_i\left(\bar{w} . x_i + \bar{b}\right) - 1\right) = 0, i = 1, 2, \ldots, N. \quad (17)$$

α yang dihasilkan digunakan untuk mencari w. Data yang memiliki nilai $\alpha_i \geq 0$ merupakan *support vector* sedangkan sisanya memiliki nilai $\alpha_i = 0$.

Setelah nilai α_i ditemukan, maka kelas dari data pengujian x dapat ditentukan berdasarkan nilai fungsi keputusan:

$$f(x_d) = \sum_{i=1}^{N.S} \alpha_i y_i x_i . x_d + b \qquad (18)$$

dengan

x_i = *support vector*

n_s = jumlah *support vector*

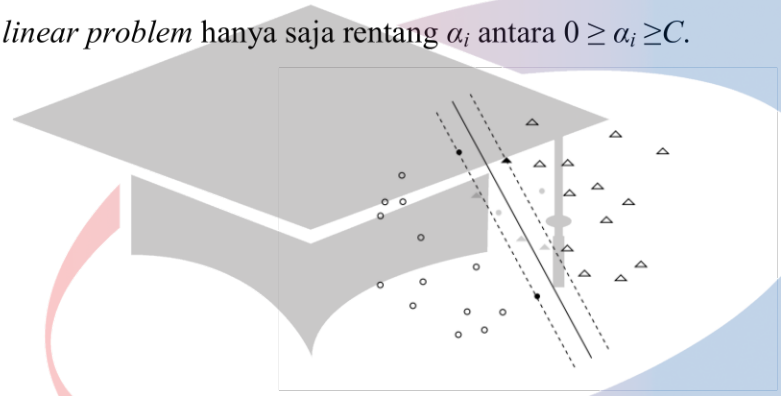
x_d = data yang akan diklasifikasikan.

Jika data tidak dapat dipisahkan secara sempurna dengan pemisahan secara linear, SVM dimodifikasi dengan menambahkan variabel ξ_i ($\xi_i \geq 0$, $\forall_I$;$\xi_i = 0$ jika x_i diklasifikasikan dengan benar) sehingga formula pencarian bidang pemisah terbaik menjadi:

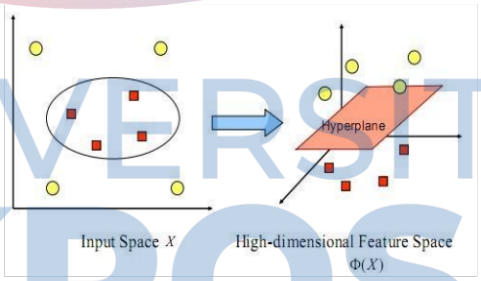
$$min\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right)$$

$$s.t.\, y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \qquad (19)$$

Pencarian bidang pemisah terbaik dengan penambahan variabel ξ_i disebut *soft margin hyperplane*. *C* adalah parameter yang menentukan besar penalti akibat kesalahan dalam klasifikasi. Dengan demikian, *dual problem* yang dihasilkan pada *non linear problem* sama dengan *dual problem* yang dihasilkan dengan *linear problem* hanya saja rentang α_i antara $0 \geq \alpha_i \geq C$.



Gambar 2.4 Hyperplane pemisah optimal yang digeneralisasi



Gambar 2.5 Fungsi Φ memetakan data ke ruang vektor yang berdimensi lebih tinggi

Cara lain untuk data yang tidak dapat dipisahkan secara linear, *Support Vector Machine* dimodifikasi dengan memasukkan fungsi ϕ(x). Dalam *Support Vector Machine* (SVM) untuk data yang tidak dapat dipisahkan secara linear, pertama-tama data dipetakan oleh fungsi ϕ(x) ke ruang vector baru yang berdimensi lebih tinggi, seperti pada Gambar 2.3. Selanjutnya di ruang vector yang baru itu, *Support Vector Machine* (SVM) mencari *hyperplane* yang memisahkan kedua

kelas secara linear. Pencarian ini hanya bergantung pada *dot product* dari data yang sudah dipetakan pada ruang baru yang berdimensi lebih tinggi, yaitu $\phi(x_i)\phi(x_d)$. Karena umumnya transformasi $\phi(x)$ ini tidak diketahui dan sangat sulit untuk diketahui, maka perhitungan *dot product* dapat digantikan dengan fungsi *Kernel* yang dirumuskan sebagai berikut:

$$K(x_i, x_d) = \Phi(x_i).\Phi(x_d) \quad (20)$$

sehingga persamaan diatas menjadi seperti berikut:

$$L = \sum_{i=1}^{i} \alpha_i - \frac{1}{2}\sum_{i=1}^{i}\sum_{j=1}^{i} \alpha_i\alpha_j y_i y_i K(x_i.\alpha_d) + b \quad (21)$$

Dengan demikian fungsi yang dihasilkan adalah:

$$f(x_d) = \sum_{i=1}^{N.S} \propto_i y_i K(x_i.x_d) + b \quad (22)$$

x_i= *support vector* dan NS adalah jumlah *support vector*.

Beberapa fungsi *Kerne l* yang umum digunakan adalah (Lin,2010):

1. *Linear kernel*

$$K(x_i, x) = x_i^T.x \quad (23)$$

2. *Polynomial kernel*

$$K(x_i, x) = \left(\gamma.x_i^T.x + r\right)^d, \gamma > 0 \quad (24)$$

3. *Radial Basic Function*

$$K(x_i, x) = exp(-\gamma|x_i - x|^2), \gamma > 0 \quad (25)$$

4. *Sigmoid kernel*

$$K(x_i, x) = tan(\gamma.x_i^T.x + r)^d, \gamma > 0 \quad (26)$$

Dalam hal ini *γ, r,* dan *d* merupakan parameter *kernel*, serta parameter C sebagai penalti akibat kesalahan dalam klasifikasi untuk masing-masing *kernel*.

2.5 Pre processing data

Sebelum data diolah, maka tahapan pertama adalah pengolahan data (*Data Pre-processing)* yaitu untuk membersihkan data dari *noise* dan *outlier* atau proses *cleaning* dan dinormalisasi dengan metode *min max* untuk mendapatkan atribut yang relevan, ringkas dan sesuai dengan format input algoritma *Support Vector Machine*.

2.6 Normalisasi

Normalisasi merupakan sebuah teknik dalam logical desain sebuah basis data yang mengelompokkan atribut dari suatu relasi sehingga membentuk struktur relasi yang baik (tanpa redudansi).

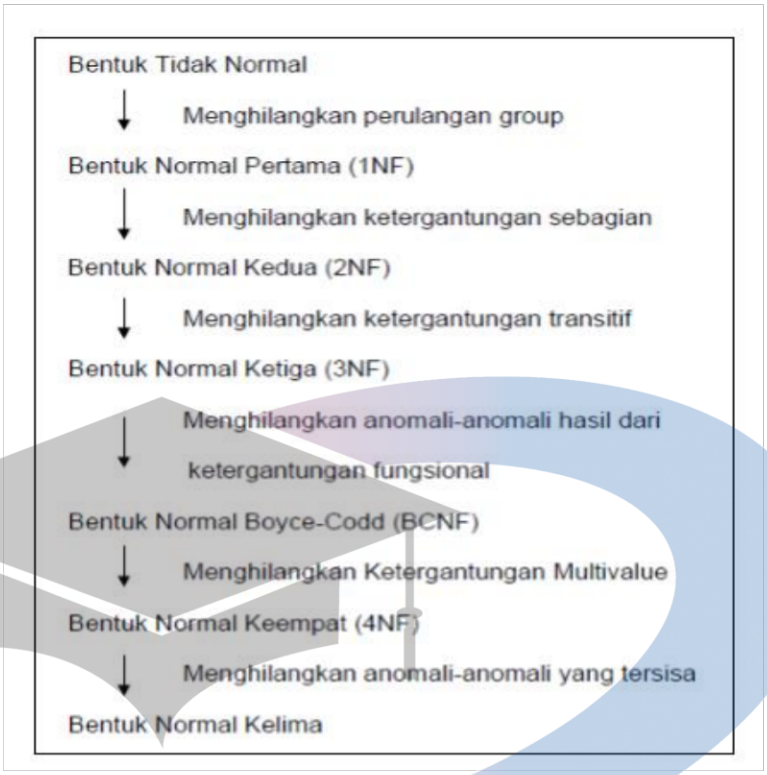
Adapun tujuan normalisasi adalah

a. Untuk menghilangkan dan mengurangi redudansi data.
b. Untuk memastikan dependensi data (Data berada pada tabel yang tepat).
c. Untuk mempermudah pemodifikasian data

Proses normalisasi pertama kali diperkenalkan oleh E.F. Codd pada tahun 1972. Normalisasi sering dilakukan sebagai serangkaian tes pada relasi untuk menentukan apakah suatu relasi sudah memenuhi atau masih melanggar persyaratan bentuk normal tertentu. Pada awalnya terdapat 3 jenis bentuk normal yang diusulkan, yaitu bentuk normal ke satu (1NF), bentuk normal kedua (2NF), dan bentuk normal ketiga (3NF). Setelah itu R. Boyce dan E.F. Codd memperkenalkan Boyce Codd Normal Form (BCNF), bentuk normal yang lebih tinggi dari bentuk normal ketiga pada tahun 1974. Pada perkembangan selanjutnya muncul pula bentuk normal ke-4 dan ke-5.

2.6.1 Tahapan Normalisasi

Berikut ini adalah tahapan normalisasi. Penjelasan lebih rinci akan diberikan pada modul 12. Bentuk normal kedua (2NF) adalah lebih baik dari bentuk normal kesatu (1NF); bentuk normal ketiga (3NF) adalah lebih baik dari Bentuk normal kedua 2NF. Untuk kepentingan rancangan database bisnis, bentuk normal ketiga (3NF) adalah bentuk terbaik dalam proses normalisasi (sudah mencukupi). Normalisasi dengan level paling tinggi tidak selalu diharapkan. Jadi normalisasi dilakukan, sepanjang dirasa sudah cukup normal (dengan mengikuti prasyarat normalisasi diatas). Tahapan Normalisasi dapat dilihat pada gambar 2.6.



Gambar 2.6 Tahapan Normalisasi

Penjelasan Gambar 2.6

1. Bentuk Normal Kesatu (1NF) Suatu relasi dikatakan sudah memenuhi bentuk normal kesatu bila setiap data bersifat atomik yaitu setiap irisan baris dan kolom hanya mempunyai satu nilai data Tujuan 1NF adalah:
   a. Membuang adanya pengulangan (Redudansi) data,
   b. Menghindari adanya pencatatan Null Value, dan
   c. Menjaga setiap entry data dari relasi (perpotongan baris kolom) memiliki maksimal satu nilai tunggal.
2. Bentuk Normal Kedua (2NF) Suatu relasi dikatakan sudah memenuhi bentuk normal kedua bila relasi tersebut sudah memenuhi bentuk normal kesatu, dan atribut yang bukan key sudah tergantung penuh terhadap keynya.
3. Bentuk Normal Ketiga (3NF) Suatu relasi dikatakan sudah memenuhi bentuk normal ketiga bila relasi tersebut sudah memenuhi bentuk normal kedua dan atribut yang bukan key tidak tergantung transitif terhadap keynya.
4. Boyce Codd Normal Form (BCNF) Suatu relasi R dikatakan dalam bentuk BCNF jika dan hanya jikasetiap atribut kunci (Key) pada suatu relasi adalah kunci kandidat (candidate key).

2.6.2 Konsep Normalisasi

Suatu rancangan database disebut buruk jika terdapat beberapa fakta berikut ini: data yang sama tersimpan di beberapa tempat yang berbeda, tidak mampu untuk menghasilkan informasi tertentu, kehilangan informasi, terjadi duplikasi data (pengulangan) yang menyebabkan pemborosan ruang penyimpanan serta timbulnya null value. Fakta ini disebabkan oleh adanya anomali (penyimpangan).

Sebaliknya, tabel-tabel dalam sebuah database yang baik harus memenuhi aturan normalisasi, yaitu bebas dari ketergantungan struktural atau anomali yang disebabkan oleh modifikasi data. Ini disebut dengan modification anomaly. Modification anomaly dibagi menjadi 3, yaitu: deletion anomaly, insertion anomaly, dan update anomaly. Dengan demikian tujuan dari normalisasi adalah untuk menghilangkan duplikasi/kerangkapan data, mengurangi kompleksitas data, dan mempermudah modifikasi data.

Pada bentuk normal, setiap baris dalam suatu tabel harus unik, setidaknya pada satu atribut yang disebut sebagai primary key. Tabel-tabel pada database dihubungkan dengan menanamkan primary key dari suatu tabel ke tabel yang berhubungan sebagai foreign key. Nilai atribut di setiap kolom harus dari kelas atau tipe data yang sama. Dan setiap kolom pada tabel harus memiliki nama yang unik.

2.7 K-Fold Cross Validation

Cross validation adalah teknik statistik yang umumnya digunakan untuk memeriksa dan mengevaluasi algoritma atau model pembelajaran dengan mempartisi data ke dalam set pembelajaran untuk melatih model dan set pengujian untuk mengevaluasinya. Set pelatihan dan pengujian dalam validasi silang secara acak dibagi menjadi beberapa partisi (60% data dalam set pelatihan dan 40% data dalam set pengujian) dan melalui putaran *crossover* berturut-turut sehingga setiap *instance* diuji. *K-fold cross validation* adalah bentuk dasar dari salah satu partisi K yang digunakan sebagai set validasi (A. Lavecchi, 2005).

2.8 Feature Selection

Feature selection atau seleksi fitur adalah salah satu teknik terpenting dan sering digunakan dalam *pre-processing*. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi. Tujuan utama dari seleksi fitur ialah memilih fitur terbaik dari suatu kumpulan fitur data.

Feature selection adalah suatu proses menghapus *features* yang berlebihan dan tidak relevan dari dataset yang sebenarnya. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena *features* yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif (S. Doraisami dan S. Golzari,2008). Dengan *feature selection* dapat meningkatkan pemahaman dan biaya penanganan data menjadi lebih kecil (A. Arauzo-Azofra,2011). Algoritma *Feature selection* dibagi menjadi tiga kelompok: *filters, wrappers*, dan *embedded selectors*. *Filters* mengevaluasi setiap *feature* secara bebas dari pengklasifikasi, memberikan peringkat pada feature setelah mengevaluasi dan mengambil yang unggul (Guyon Isabelle dan A. Elisseef, *Journal of Machine learning Research*). Wrappers mengambil subset dari feature set, mengevaluasi kinerja pengklasifikasian pada *subset* ini, dan kemudian *subset* lainnya dievaluasi oleh pengklasifikasi. Subset yang memiliki kinerja paling maksimum pada pengklasifikasian yang akan dipilih. Jadi *wrappers* bergantung pada pengklasifikasi yang dipilih. Bahkan wrappers lebih dapat diandalkan karena algoritma klasifikasi mempengaruhi tingkat akurasi (J. Novakovic,2010). Teknik *Embedded* disisi lain melakukan feature selection selama proses mempelajari data sama seperti yang dilakukan jaringan syaraf tiruan.

Terdapat banyak algoritma *features selection*, yang akan digunakan dalam penelitian ini yaitu algoritma *features selection* yang bersifat *univariate* yang disebut *select K best*. Untuk *Univariate features selection* secara umum bekerja dengan cara memilih *features* terbaik berdasarkan *test statistic univariate*. Hal ini

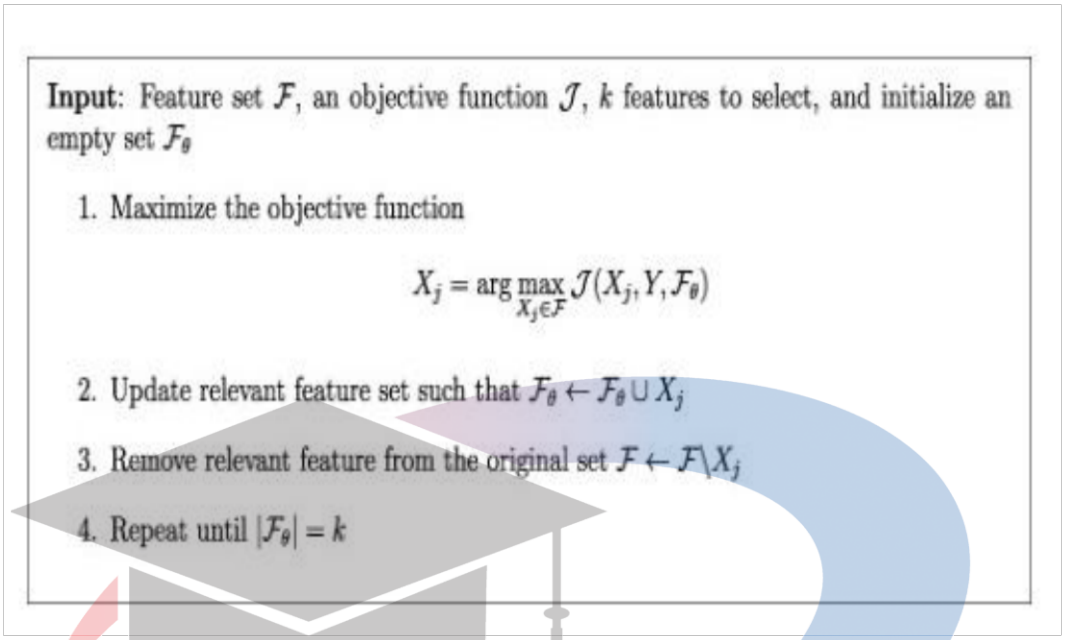
dapat dilihat sebagai langkah *preprocess* sebuah estimator. *Select K best* secara khusus bekerja dengan cara memilih sejumlah *K features* terbaik berdasarkan pengujian statistic (Pedregosa et al,2011).

2.8 Forward Selection

Forward selection adalah jenis regresi bertahap yang dimulai dengan model kosong. Dalam seleksi *forward*, variabel pertama yang dipilih untuk entri ke dalam model yang sudah dibangun adalah yang memiliki korelasi terbesar dengan variabel dependen. Setelah variabel telah dipilih kemudian dievaluasi berdasarkan kriteria tertentu. Jika variabel yang dipilih pertama memenuhi kriteria untuk dimasukkan, maka pemilihan maju berlanjut. Prosedur berhenti, ketika tidak ada variabel lain yang tersisa yang memenuhi kriteria entri dan menambahkan variabel satu per satu.

Algoritma *Forward Selection*

Dalam pemilihan fitur, *forward Selection* memiliki fungsi objektif J yang sudah dimaksimalkan, dan fungsi ini bergantung pada subset fitur Fθ. Tujuan dari algoritma *Forward Selection* adalah untuk menemukan fitur k dalam F yang memaksimalkan fungsi objektif. Algoritma Forward Selection dapat dilihat pada gambar 2.6.



Gambar 2.7 Algoritma *Forward Selection*

2.9 Confusion Matrix

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Pada pengukuran kinerja menggunakan confusion matrix, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, *True Positive* (TP) merupakan data positif yang terdeteksi benar. *False Negative* (FN) merupakan kebalikan dari *True Positive*, sehingga data positif, namun terdeteksi sebagai data negatif.

Akurasi adalah persentase dari total data yang diidentifikasi dan dinilai. Perhitungan akurasi diperlukan untuk mengetahui baik atau tidaknya performa suatu model klasifikasi. Pada penelitian ini dilakukan perbandingan akurasi hasil antara model klasifikasi tanpa atau menggunakan *Support Vector Machine*. Untuk menghitung akurasi dapat digunakan rumus *confusion matrix* pada Tabel 2.1 berikut ini (Sokolova & Lapalme, 2009):

Tabel 2.1 Confusion matrix

		Correct result / classification	
		E1	E2
Obtained result / Classification	E1	TP (True Positive)	FP (False Positive)
	E2	FN (False Negative)	TN (True Negative)

□□□□□□ = (TP + TN) / (TP + TN + FP + FN)) * 100% (27)

Dimana :

1. TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.

2. TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.

3. FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

4. FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

Presisi adalah data yang diambil berdasarkan informasi yang kurang. Dalam klasifikasi biner, presisi dapat dibuat sama dengan nilai prediksi positif. Berikut ini adalah aturan presisi.

Precision = (TP / (TP + FP)) * 100% (28)

Recall adalah data penghapusan yang berhasil diambil dari data yang relevan dengan kueri. Dalam klasifikasi biner, *recall* dikenal sebagai sensitivitas.

Munculnya data relevan yang diambil adalah menyetujui dengan query dapat dilihat dengan *recall.* Berikut ini adalah peran *recall*.

Recall = (TP / (TP + FN)) * 100% (29)