BAB II
TINJAUAN PUSTAKA

1.1. Penelitian Terkait

Penelitian sebelumnya yang menjadi latar belakang penelitian ini dijabarkan pada tabel

Tabel 2.1. Penelitian Terkait

	Peneliti	Judul	Tahun	Dataset	Metode	Hasil
[9]	Agustin Trihartati S.1, et al.	*An Identification of Tuberculosis (TB) Disease in Humans using Naïve Bayesian Method*	2016	237 data sample	*Naïve Bayesian*	*accuracy* 85,95%
[10]	I Md. Dendi Maysanjaya	Klasifikasi Pneumonia pada Citra X-rays Paru-paru dengan *Convolutional Neural Network*	2020	5.840 citra	*Convolutional Neural Network*	Akurasi 89,58%
[11]	Aida Muhdina, et al.	Klasifikasi Tuberkulosis Dengan *Compressive Sensing* Dan *Support Vector Machine*	2021	90 Buah Citra Yang Terbagi Menjadi 63 Data Latih Dan 27 Data Uji	*Compressive Sensing* Dan *Support Vector Machine*	Akurasi 92,593%
[12]	Ovy Rochmawanti, et al	Analisis Performa *Pre-Trained Model Convolutional Neural Network* Dalam Mendeteksi Penyakit Tuberkulosis	2021	662 gambar yang terbagi atas 326 kasus normal dan 336 kasus TB	*Convolutional Neural Network*	Akurasi 91,57%
[13]	V Bharath et al	*Recognition of Tuberculosis Through Image Modalities*	2021	3500 normal chest X-ray and 3500 tuberculosis chest x-rays.	*Support Vector Machine*	90%
[14]	Chang Liu	*Tx-Cnn: Detecting Tuberculosis In Chest X-Ray Images Using Convolutional Neural Network Chang*	2017	4701 *images*, 453 *normal and* 4248 TB	*Convolutional Neural Network*	*Accuracy* 85.68%

	Peneliti	Judul	Tahun	Dataset	Metode	Hasil
[15]	Risha Ambar Wati, et al	Klasifikasi Pneumonia Menggunakan Metode *Support Vector Machine*	2020	5.853 citra rontgen paru-paru	*Support Vector Machine* (SVM)*)*	akurasi terbaik sebesar 62.66%
[16]	Vivin Vidia Nurdiansyah, et al.	Klasifikasi Penyakit Tuberkulosis (TB) menggunakan Metode *Extreme Learning Machine (ELM)*	2020	data penyakit TB Puskesmas Dinoyo Tahun 2018-2019	*Extreme Learning Machine (ELM)*	Akurasi 99,33%
[17]	Yudhi Agussationo. Et.al.	Klasifikasi Citra X-Ray Diagnosis Tuberkulosis Berbasis Fitur Statistis	2018	Data penelitian diperoleh dari RS Dr. Sardjito Yogyakarta sebanyak 33 citra digital x-ray pasien diagnosis tuberkulosis	metode Histogram , GLCM,PCA	metode Histogram (81,81%), metode GLCM (96,96%), metode PCA (81,82%)
[18]	Reni Rahmadewi*, Rahmadi Kurnia	Klasifikasi Penyakit Paru Berdasarkan Citra Rontgen Dengan Metoda Segmentasi Sobel	2016	41 (empat puluh satu) citra rontgen	Segmentasi Sobel	Hasil 94,85%
[19]	Tawsifur Rahman, et al.	*Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization*	2020	*3500 TB infected and 3500 normal chest X-ray images*	*CNN models*	*accuracy 96.47%,*
[20]	Ophir Gozes and Hayit Greenspan	*Deep Feature Learning from a Hospital-Scale Chest X-ray Dataset with Application to TB Detection on a Small-Scale Dataset*	2019	*112K images from the ChestXray14 dataset*	DenseNet-121 CNN	*Accuracy* 0.965
[21]	Irvi Oktanisa, et al	Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing	2017	dataset Bank Direct Marketing	*Stochastic Gradient Descent*, dan *CN2 Rule*	Akurasi terbesar *Stochastic Gradient Descent* sebesar 0,972
[22]	Siti Nur Asiyah	Klasifikasi Berita *Online* Menggunakan Metode *Support Vector Machine* Dan *K-Nearest Neighbor*	2016	Data Berita *Online* Www.Detik.Com.	*SVM Dan KNN*	Akurasi *93.2%,*

1.2. Teknik Klasifikasi

Teknik Klasifikasi adalah sebuah model dalam data mining dimana, classifier dikontruksi untuk memprediksi categorical label, seperti “aman” atau “ beresiko” untuk data aplikasi

peminjaman uang ; “ya” atau “tidak” untuk data marketing ; atau “treatment A”,”treatment B” atau “trratment C” untuk data medis. Kategori tersebut dapat direpresentasikan dengan nilai yang sesuai dengan kebutuhannya, dimana pengaturannya dai nilai tersebut tidak memiliki arti tertentu.

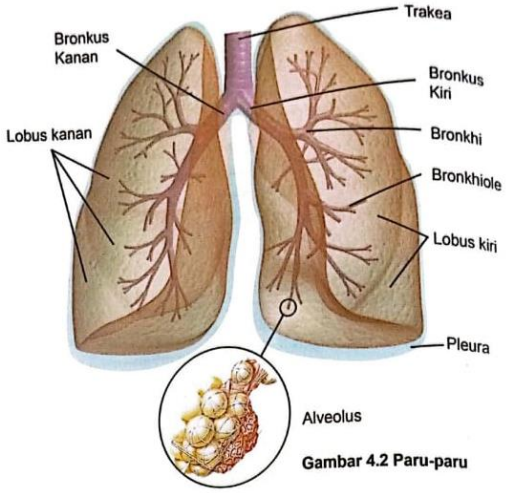
Classification dan association rule discovery merupakan tugas yang sama dengan data mining, dengan pengecualian bahwa tujuan utama dari klasifikasi adalah prediksi label kelas, sedangkan asosisasi atura oenemuan menggambarkan korelasi antara item dalam *database* transaksional.

Proses data klasifikasi memiliki dua tahapan, yang pertama adalah learning; yaitu training data dianalisa dengan menggunakan sebuah algoritma klasifikasi. Dam yang kedua adalah *classification* ; yaitu pada tahao ini test data digunakan untuk mengestimasi ketepatan dari *classification rules*. Jika keakuratan yang dikondisikan dan yang di perkirakan dapat diterima, rule tersebut dapat diaplikasikan pada klasifikasi lainnya dari tuole data yang baru. Lebih spesifik mengatakan bahwa, classification hanya bias diterapkan pada data *training* yang sangat kuat dimana diasumsikan bahwa kelas “positif” sudah mewakili minoritas tanpa kehilangan atribut umum[23].

1.3. Anatomi Paru-Paru Manusia

Paru-paru terletak di dalam rongga mulut dada tepat di atas diafragma. Diafragma adalah sekat berotot yang membatasi rongga dada dan rongga perut. Paru-paru terdiri atas 2 bagian, kiri dan kanan, yang terletak di rongga dada. Sedangkan jantung terletak hamper di tengah rongga dada, diantara kedua paru-paru, dengan posisi yang lebih ke kiri sedikit. Paru-paru kanan tersusun atas 3 gelambir, sedangkan paru-paru kiri 2 gelambir.

Paru-paru dibungkus oleh selaput paru-paru yang disebut pleura. Didepannya terdapat batang tenggorok dan saluran pernafasan (*bronchi*). Oleh sebab jantung agak mengambil tempat ke kiri, bagian paru-paru sebelah kiri lebih kecil sedikit dari paru-paru kanan. dengan demikian dapat dimengerti paru-paru kiri hanya terdiri atas 2 bagian (lobus), sedangkan paru-paru kanan 3 bagian [24].



Gambar 2 1 Anatomi Paru-Paru Manusia

1.4. Tuberkulosis (TBC)

Tuberkulosis (TBC) saat ini masih merupakan masalah kesehatan masyarakat baik di Indonesia maupun internasional sehingga menjadi salah satu tujuan pembangunan kesehatan berkelanjutan (SDGs). Tuberkulosis adalah penyakit menular yang disebabkan oleh kuman Mycobacterium tuberculosis dan merupakan salah satu dari 10 penyebab utama kematian di seluruh dunia.

Indonesia berada pada peringkat ke-2 dengan penderita TB tertinggi di Dunia setelah India. Secara global, diperkirakan 10 juta orang menderita TB pada tahun 2019. Meskipun terjadi penurunan kasus baru TB, tetapi tidak cukup cepat untuk mencapai target Strategi END TB tahun 2020, yaitu pengurangan kasus TB sebesar 20% antara tahun 2015 – 2020. Pada tahun 2015 – 2019 penurunan kumulatif kasus TB hanya sebesar 9%

Begitu juga dengan kematian akibat TB, jumlah kematian pada tahun 2019 sebesar 1,4 juta. Secara global kematian akibat TB per tahun menurun secara *global*, tetapi tidak mencapai target Strategi END TB tahun 2020 sebesar 35% antara tahun 2015 – 2020. Jumlah kematian kumulatif antara tahun 2015 – 2019 sebesar 14%, yaitu kurang dari setengah dari target yang ditentukan [25]

1.5. Data Mining

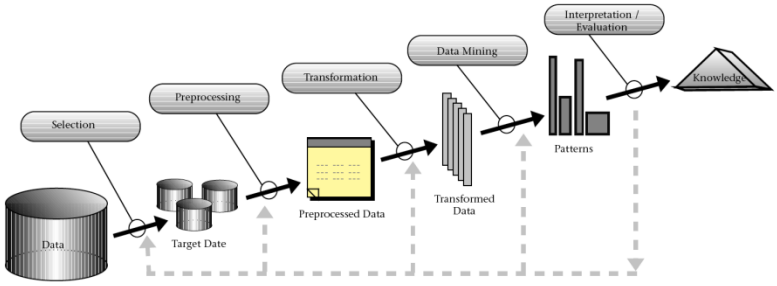
Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat pada basis data. Data mining terutama digunakan untuk mencari pengetahuan

yang terdapat dalam basis data yang besar sehingga sering disebut *knowledge discovery databases* (KDD)

Data mining merupakan salah satu dari rangkaian *knowledge discovery in database* (KDD). KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dan penemuan ilmiah, interpretasi dan visualisasi dan pola-pola sejunlah data. Serangkaian proses tersebut memiliki tahap sebagai berikut.

1. Pembersihan data (untuk membuang data yang tidak konsisten dan noise)
2. Integrasi data (penggabungan data dari beberapa sumber)
3. Transformasi data (data diubah menjadi bentuk yang sesuai untuk di-mining)
4. Aplikasi teknik *Data Mining*, proses ekstraksi pola dari data yang ada
5. Evaluasi pola yang ditemukan (proses interprestasi pola menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan)
6. Presentasi pengetahuan (dengan teknik visualisasi).

Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna. Gambar 2.2 menunjukan proses atau tahapan KDD [26]



Gambar 2.2 Tahapan Knowledge In Database (KDD)

Adapun tahapan KDD pada gambar 2.2 adalah :

1. Data

   Membuat himpunan data target, penetapan himpunan data dan memfokuskan pada subset variabel atau sampel data, dimana penelitiaan akan dilakukan.
2. Pemilihan data (*Selection*)

Langkah pertama pemprosesan data dan pembersihan data adalah tindakan dasar seperti penghapusan noise. Sebelum melakukan proses data mining, maka diperlukan proses cleaning pada data yang menjadi fokus dalam KDD

3. Transformasi (*Transformation*)

   Pada tahap ini merupakan tahapan proses kreatif dan sangat tergantung pada pola informasi yang akan dicari dalam basis data.

4. Data Mining

   Dalam pemilihan algoritma data mining untuk melakukan pencarian proses data mining yaitu antara lain teknik, metode atau algoritma dalam data mining sangat bervariasi. Penetapan metode atau algoritma yang tepat tergantung pada tujuan dan proses KDD secara keseluruhan.

5. Evaluasi (*Evaluation*) Tahap ini merupakan tahapan pemerikasaan, apakah pola yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

1.6. Image Processing

Image processing atau pengolahan citra merupakan suatu metode atau teknik yang dapat digunakan untuk memproses citra atau gambar dengan jalan mjemanipulasinjya menjadi suatu data gambar yang diisikan untuk mendapatkan suatu informasi tertentu mengenai obyek yang sedang diamati [27]

1.7. Citra Digital

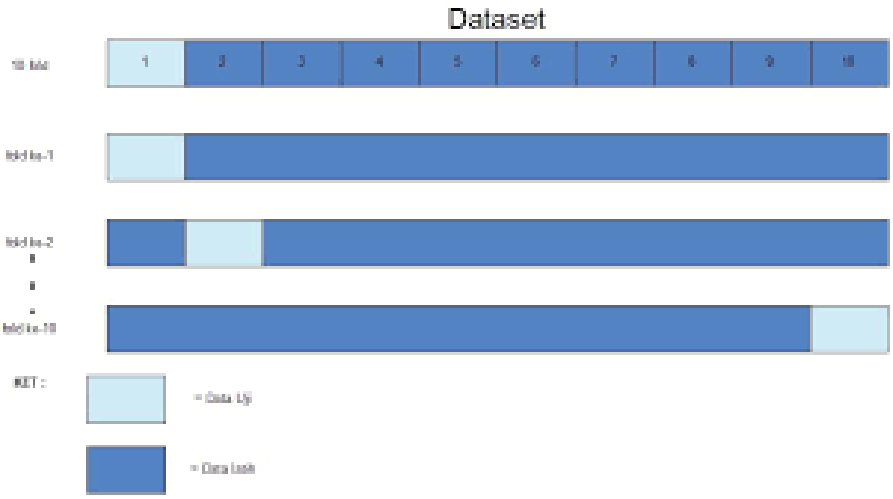
Citra adalah representasi dua dimentasi untuk bentuk-bentuk fisik nyata tiga dimensi.Citra dalam perwujutan dapat bermacam-macam,mulai dari gambar perwujudan nya dapat bermacam –macam,mulai dari gambar putih pada sebuah foto (yang tidak bergerak)sampai pada gambar warna yang bergerak pada televisi .proses transformasi dari bentuk tiga dimensi ke bentuk dua dimensi untuk menghasilkan citra akan dipengaruhi oleh bermacammacam factor yang mengakibatkan citra penampilan citra suatu benda tidak sama persis dengan bentuk fisik nyatanya .faktor-faiktor tersebut merupakan efek degradasi atau penurunan kualitas yang dapat berupa rentang kontras benda yang terlalu sempit atau terlalu lebar ,distorsi geometric kekaburan(blur),kekaburan akibat objek citra yang bergerak 9 motion blur,noisw atau gangguan yang disebabkan oleh interferensi pembuat citra ,baik itu pembuat tranduser ,peralatan elektronik maupun peralatan optik .karena pengolahan citra digital dilakukan dengan computer digital ,maka citra yang akan diolah terlebih dahulu

ditranformasikan kedalam bentuk besaran – besaran diskrit dari niloai tingkat keabuan pada titik element citra .bentuk dari citra ini disebut citra digital .

element-element citra digital apabila ditampilkan dalam layer monitor akan menempati sebuah ruang yang disebut Pixel(*picture element)* .Teknik dan proses untuk mengurangi atau menghilangkan efek degradasi pada citra meliputi teknik perbaikan atau peningkatan citra (image enchancement) ,restorasi citra (image restoration) dan tranformasi spesial (*special transformation*),subyek lain dari pengolahan citra digital diantaranya adalah pengkodean citra ,segmentasi citra(image segmentation),representasi edan diskripsi citra (image representation and diskription) [27].

1.8. Cross Validation

Cross Validation adalah metode untuk memperkirakan kesalaah prediksi untuk evaluasi kinerja model. Dalam cross validation dikenal sebagai estimasi rotasi, dengan membagi data menjadi himpunan bagian k dengan ukuran yang hampir sama, model dalam klasifikasi dilatih dan diuji sebanyak k. Disetiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data penguji dan sub kelompok data k lainnya berfungsi sebagai data pelatihan. K-fold *cross validation* merupakan metode untuk mengevaluasi kinerja classifier, metode ini dapat digunakan apabila memiliki jumlah data yang terbatas (jumlah instance tidak banyak). K-fold cross validation adalah suatu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan redudansi dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. K-fold *cross validation* diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses *cross validation* data akan dibagi dalam n buah partisi dengan ukuran yang sama variabel Data ke 1, variabel Data ke 2, variabel Data ke 3 .. Dn selanjutnya proses uji dan latih dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data uji dan sisanya akan menjadi data latih. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model [28]. Contoh pembagian dataset dalam proses 10-fold *cross validation* bisa dilihat pada gambar 2.3**.**



Gambar 2.3 Contoh iterasi data dengan cross validation

Kinerja dari K-fold cross validation yaitu:

1. Total *instance* dibagi menjadi N bagian. Universitas

2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (*testing* data) dan sisanya menjadi data latih (*training* data). Selanjutnya, hitung akurasi atau kesamaan atau kedekatan suatu hasil pengukuran dengan angka atau data yang sebenarnya berdasarkan porsi data tersebut. Perhitungan akurasi tersebut menggunakan persamaan sebagai berikut.

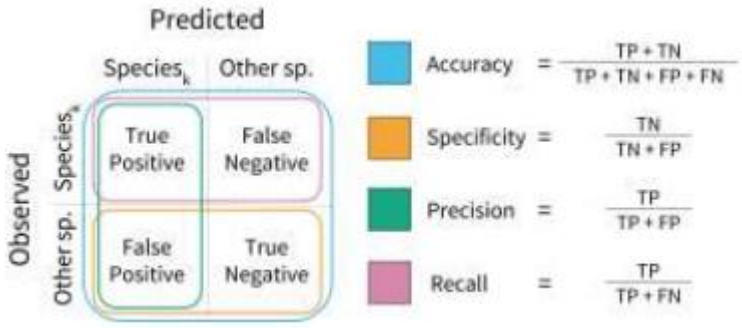
$$Akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji}\ 100X \qquad (1)$$

3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (*testing* data) dan sisanya menjadi data latih (*training* data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.

4. Demikian seterusnya hingga mencapai fold ke-k. Hitung rata-rata akurasi dari k buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

1.9. Confusion Matrix

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi, *recall, precision,* dan *error rate*. Dimana, precision mengevaluasi kemampuan sistem untuk menemukan peringkat yang paling relevan, dan didefinisikan sebagai presentase dokumen yang di retrieve dan benar-benar relevan terhadap *query*. *Recall* mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai presentase dokumen yang relevan terhadap *query*. *Accuracy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan error rate merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus [29].

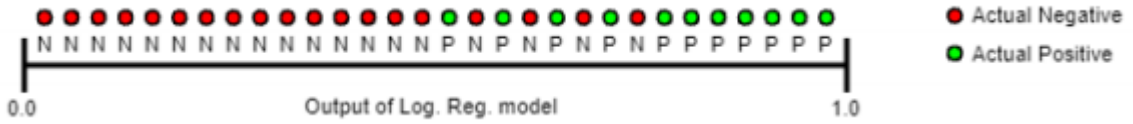


(2)

Gambar 2.4 Tabel Confusion Matrix 2x2

1.10. AUC : Area Under the Curva

AUC memberikan ukuran kinerja agregat di semua ambang klasifikasi yang m ungkin. Salah satu cara untuk menginterpretasikan AUC adalah sebagai probabilitas bahwa model memberi peringkat contoh positif acak lebih tinggi daripada contoh negatif acak. Sebagai contoh, diberikan contoh berikut, yang disusun dari kiri ke kanan dalam urutan menaik dari prediksi regresi logistik:



AUC mewakili probabilitas bahwa contoh acak positif (hijau) diposisikan di sebelah kanan contoh acak negatif (merah). Rentang nilai AUC dari 0 hingga 1. Model yang prediksinya 100% salah memiliki AUC 0,0; yang prediksinya 100% benar memiliki AUC 1,0. AUC diinginkan karena dua alasan berikut:

1. AUC adalah skala-invarian . Ini mengukur seberapa baik prediksi diberi peringkat, bukan nilai absolutnya.
2. AUC adalah klasifikasi-ambang-invarian . Ini mengukur kualitas prediksi model terlepas dari ambang klasifikasi apa yang dipilih.

Namun, kedua alasan ini disertai dengan peringatan, yang dapat membatasi kegunaan AUC dalam kasus penggunaan tertentu :

1. Skala invarians tidak selalu diinginkan. Misalnya, terkadang kami benarbenar membutuhkan keluaran probabilitas yang terkalibrasi dengan baik, dan AUC tidak akan memberi tahu kami tentang itu.
2. Klasifikasi-ambang invarians tidak selalu diinginkan. Dalam kasus di mana ada perbedaan besar dalam biaya negatif palsu vs positif palsu, mungkin penting untuk meminimalkan satu jenis kesalahan klasifikasi. Misalnya, saat melakukan deteksi spam

email, Anda mungkin ingin memprioritaskan meminimalkan positif palsu (bahkan jika itu menghasilkan peningkatan negatif palsu yang signifikan). AUC bukan metrik yang berguna untuk jenis pengoptimalan ini. Performance keakurasian AUC dapat diklasifikasikan menjadi beberapa kelompok yaitu:

a. 0.90 – 1.00 = klasifikasi sangat baik *(excellent classification)*

b. 0.80 – 0.90 = klasifikasi baik *(good classification)*

c. 0.70 – 0.80 = klasifikasi cukup *(fair classification)*

d. 0.60 – 0.70 = klasifikasi buruk *(poor classification)*

e. 0.50 – 0.60 = klasifikasi salah *(failure)*

1.11. Algoritma *Logistic Regression*

Logistic Regression adalah bagian dari metode statistik yang disebut juga dengan generalized linear model. *Logistic Regression* model digunakan saat variabel respon mengacu pada dua nilai. Misal, ketika subjek berupa benda mati atau tidak hidup, punya atau tidak memiliki sebuah karakteristik khusus dan sebagainya. Kita misalkan variabel respon sebagai y dan sebuah subjek / event y=1 ketika subjek itu memiliki karakteristik dan y=0 ketika tidak memilikinya . Berikut persamaan dari Logistic regression [30]

Metode regresi logistik memiliki teknik dan prosedur yang tidak jauh berbeda dengan metode regresi linear. Jika prosedur linear dalam mengestimasi nilai parameter sering menggunakan metode Ordinary Least Squares (OLS), maka untuk mengestimasi nilai parameter dalam regresi logistik adalah dengan menggunakan metode Maximum Likelihood Estimation (MLE). Untuk mencari persamaan (3) logistiknya maka model yang dipakai adalah : [31]

$$\pi\, x = \frac{e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}}{1 + e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}} \qquad (3)$$

Dari persamaan (1) diperoleh 1 – π(x) sebagai berikut:

$$1 - \pi\,(x) = 1 - \frac{e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}}{1 + e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}}$$

$$1 - \pi\,(x) = \frac{1 + e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j} - e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j -}}{1 + e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}}$$

$$= \frac{1}{1 + e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}}$$

Sehingga $\frac{\pi(x)}{1-\pi(x)}$ sebagai berikut

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}$$

Jadi, persamaan logistiknya adalah:

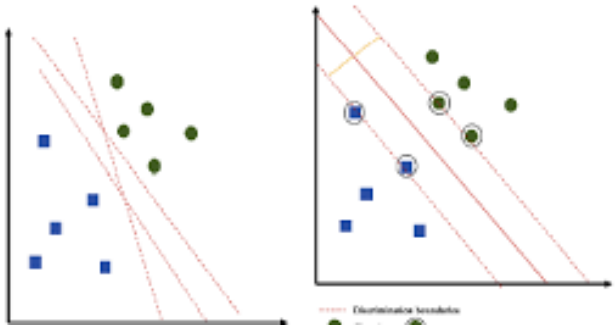
$$g(x) = In\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$= In\left(e^{\beta 0 + \sum_{j=1}^{p} \beta_j x_j}\right)$$

$$= \beta 0 + \sum_{j=1}^{p} \beta_j x_j \quad (4)$$

1.12. Algoritma SVM

Support Vector Machine (SVM) adalah algoritme *supervised* yang berupa klasifikasi dengan cara membagi data menjadi dua kelas menggunakan garis vektor yang disebut hyperplane (Octaviani, et al., 2014). Pada permasalahan yang kompleks atau permasalahan dengan parameter yang banyak, metode ini sangat baik untuk digunakan. Metode ini juga baik digunakan untuk mendiagnosis berbagai macam jenis penyakit. Salah satu kelebihan yang dimiliki metode SVM adalah penanganan error pada set data training yang menggunakan *Structural Risk Minimization* (SRM). SRM dikatakan lebih baik karena tidak hanya meminimalkan error yang terjadi, tetapi meminimalkan faktor-faktor lainnya [32].



Gambar 2.5 Model Support Vector Machine

Pada gambar 2.5 pemahaman sederhana konsep SVM digambarkan sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas. Gambar 5 a dan 5

b memperlihatkan beberapa pola yang merupakan anggota dari dua buah kelas yaitu 1 dan 0. Pola yang tergabung pada kelas 1 digambarkan dengan lingkaran hijau sedangkan pola pada kelas 0 digambarkan dengan kotak biru. Masalah klasifikasi dapat dijabarkan dengan usaha menemukan hyperplane yang memisahkan dua kelompok tersebut. Berbagai alternatif garis pemisah (discrimination boundaries) ditunjukkan pada Gambar 5 a. Hyperplane pemisah yang terbaik diantara kedua kelas ditemukan dengan cara mengukur margin hyperplane dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane dengan pola terdekat dari setiap kelas. Pola yang paling dekat ini disebut sebagai support vector. Garis solid pada Gambar 5 b menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik hijau dan biru yang berada dalam lingkaran hitam adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada SVM. Data dinotasikan sebagai $x_i \vec{} \in R^2$ sedangkan label masing-masing dinotasikan $y_i \in \{1,0\}$ untuk $i = 1,2, \dots, l$ yang mana l adalah banyaknya data. Asumsi kedua kelas 1 dan 0 dapat terpisah secara sempurna oleh hyperplane berdimensi d yang didefinisikan pada Persamaan 1[33].

$$\underset{w}{\rightarrow}.\underset{x}{\rightarrow} + b = 0 \quad (1)$$

Pola $\underset{x}{\rightarrow}$ yang termasuk kelas 1 dapat dirumuskan sebagai pola yang memenuhi pertidaksamaan (2)

$$\underset{w}{\rightarrow}.\underset{x_1}{\rightarrow} + b \geq 1 \quad (2)$$

Sedangkan pola $\underset{x_1}{\rightarrow}$ yang termasuk kelas 0 dirumuskan dengan pertidaksamaan (5)

$$\underset{w}{\rightarrow}.\underset{x_1}{\rightarrow} + b \leq 1 \quad (5)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya dengan persamaan (6)

$$\frac{1}{\left\|\vec{w}\right\|} \quad (6)$$

Hal ini dapat dirumuskan sebagai quadratic programming problem yaitu mencari titik minimal persamaan 7 dengan memperhatikan constraint persamaan 8.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\left\|\underset{w}{\rightarrow}\right\|^2 \quad (7)$$

$$y_1\left(\underset{w}{\rightarrow}.\underset{y_1}{\rightarrow}\right) - 1 \geq 0, \forall i \quad (8)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi di antaranya Langrange Multiplier, seperti ditunjukkan pada Persamaan (9).

$$\boldsymbol{L}\left(\underset{\boldsymbol{W}}{\rightarrow}, \boldsymbol{b}, \boldsymbol{\alpha}\right) = \frac{\mathbf{1}}{\mathbf{2}}\left\|\underset{\boldsymbol{w}}{\rightarrow}\right\|^{\mathbf{2}} - \sum_{\boldsymbol{i=2}}^{\boldsymbol{I}} \boldsymbol{\alpha_1}\left(\boldsymbol{y_1}\left(\left(\underset{\boldsymbol{w}}{\rightarrow} . \underset{\boldsymbol{x_1}}{\rightarrow} + \boldsymbol{b}\right) - \mathbf{1}\right)\right), \boldsymbol{i} = \mathbf{1}.\mathbf{2}.\ldots, \boldsymbol{I} \quad \textbf{(9)}$$

α_i adalah Langrange multipliers, yang bernilai nol atau positif $\alpha_i \geq 0$. Nilai optimal dari persamaan (8) dapat dihitung dengan meminimalkan *L* terhadap $\underset{W}{\rightarrow}$ dan *b* dan memaksimalkan *L* terhadap α_i. Berdasarkan sifat bahwa pada titik optimal *L* = 0, persamaan (9) dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α_i

$$\sum_{i=1}^{I} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{I} \alpha_i \alpha_j y_i y_j \underset{X_I}{\rightarrow} . \underset{x_J}{\rightarrow} \quad (10)$$

dari hasil perhitungan diperoleh *αi*yang kebanyakan bernilai positif. Data yang berkorelasi dengan *αi* yang positif inilah yang disebut *support vector*.

$$\alpha_i \geq 0\ (i = 1,2 \ldots., I) \qquad \sum_{i=1}^{I} \alpha_i y_i = 0 \quad (11)$$