BAB 2
LANDASAN TEORI

2.1 *Data*

Data adalah representasi fakta dunia nyata yang mewakili objek seperti orang, barang, hewan, peristiwa, konsep, dan situasi. Semua ini direkam dalam bentuk angka, huruf, simbol, teks, gambar, suara, atau kombinasi semuanya [11].Data dapat dikatan tidak ada artinya, sehingga kita masih perlu mengolah data tersebut. Data memiliki set, yang merupakan kumpulan objek dan atributnya. Atribut-atribut tersebut adalah properti atau karakteristik dari objek yang disebut variabel, field, properti.

Sebuah data membentuk suatu himpunan yang dapat diartikan sebagai *record* data, di mana masing-masing terdiri dari suatu atribut yang konstan. Data transaksi merupakan tipe yang termasuk ke dalam data *record* di mana setiap *record* meliputi satu set *item*.

Data terbagi atas dua golongan berdasarkan sifat-sifatnya, sebagai berikut :

2.1.1 Data Kualitatif

Data kualitatif adalah data yang digunakan untuk menjelaskan karakteristik atau sifat, selain hal tersebut tipe data ini bertujuan untuk menggolongkan. Tipe data ini termasuk ke dalam klasifikasi data yang berskala ukur nominal dan ordinal. Contohnya adalah review produk, misalnya: memuaskan, normal, biasa saja.

2.1.2 Data Kuantitatif

Data kuantitatif adalah data direpresentasikan dalam bentuka angka, tipe data ini termasuk ke dalam kategori yang berskala interval dan rasio. Seperti contohnya adalah data massa suatu benda : 1kg, 10kg dan lainnya.

2.2 *Data Mining*

Data mining merupakan kegiatan mencari dan menggali informasi yang belum diketahui secara manual dari data[12]. Informasi yang dihasilkan diperoleh dengan mengestraksi dan mengenali pola menarik dari data yang terkandung pada database. *Data mining* merupakan kegiatan dengan menggunakan beberapa teknik yang bertujuan untuk mengestrak informasi dan pengetahuan yang belum diketahui dari data yang ukurannya besar, kemudian dari data tersebut dilakukan pencarian pola

atau trend sesuai dengan tujuan dari penerapan *data mining,* selanjutnya hasil dari pengolahan *data mining* tersebut digunakan untuk pengambilan keputusan maupun hasil prediksi analisis yang dibutuhkan.

Data mining merupakan suatu kegiatan menganalisis data dengan memanfaatkan perangkat lunak untuk menemukan pola atau tren dengan mengidentifikasi aturan dan karakteristik pada data. Pada dasarnya, data mining dapat dibagi menjadi dua kategori utama, yaitu :

2.2.1 *Descriptive*

Data mining descriptive merupakan kegiatan untuk menggali nilai penting dari sebuah database yang tersembunyi dan menemukan pola data baru yang belum diketahui sebelumnya.

2.2.2 *Predictive*

Data mining predictive merupakan proses pencaran pola dari data dengan menggunakan beberapa atribut lain untuk di masa akan datang. Metode *data mining* regresi dan klasifikasi termasuk yang terdapat dalam kategori ini.

2.3 Metodologi *Data Mining* CRISP-DM

CRISP-DM adalah proses standar untuk *data mining* yang dikembangkan pada tahun 1996 oleh Daimler-Chrysler, SPSS, dan NCR [13]. CRISP-DM sendiri merupakan singkatan dari *Cross-Industry Standard Process for Data Mining*. Metodologi ini mencakup 6 fase seperti yang bisa dilihat di Gambar 1.5-1 [14] :

1. *Business Understanding* – fase pertama yang difokuskan untuk memahami objektif dari projek dan keperluan dari perspektif bisnis, lalu merubah dari pengetahuan tersebut ke definisi masalah data mining dan merencanakan desain alur untuk memecahkan masalah demi mencapai objektif.
2. *Data Understanding* – Fase ini diawali dengan pengkoleksian data dan diproses dengan aktivitas demi memahami lebih jauh dengan data yang dihadapi, untuk mengidentifikasi kualitas masalah data, untuk mencari *insight* kepada data sampai ke mendeteksi pola menarik untuk membentuk hipotesis untuk informasi tersembunyi.
3. *Data Preparation* – Semua hal aktivitas yang berhubungan dengan membangun dataset baru yang lebih bersih seperti tidak ada baris kosong, dan sebagainya.

4. *Modelling* – Di fase ini, berbagai macam tekni model akan digunakan dan diterapkan dan di tahap ini juga dilakukan kalibrasi parameter model.
5. *Evaluation* – di tahap ini model akan dievaluasi dengan diukur kualitas performanya dan dilakukan *review* apakah sudah memenuhi dari objektif bisnis di fase pertama.
6. *Deployment* – Dilakukannya implementasi model terhadap suatu sistem dan direpresentasikan sedemikian rupa sehingga pengguna dapat menggunakannya.

2.4 Metode Klasifikasi

Pada *data mining* terdapat beberapa metode yang bisa digunakan, salah satunya ialah klasifikasi. Klasifikasi merupakan suatu metode *data mining* yang memberi label kepada suatu data dari kategori target atau kelas [11]. Adapun dalam proses pembelajaran pengklasifikasian data dibagi menjadi dua tahap proses.

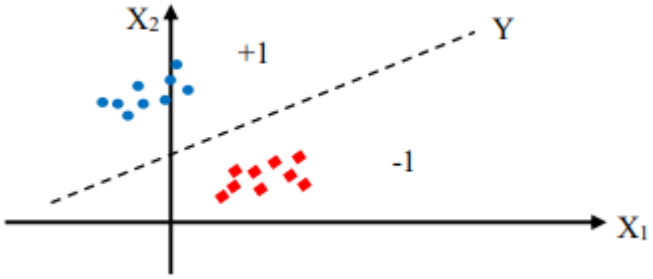
1. *Learning step*

   Merupakan suatu proses untuk mencari pemodelan dengan menggunakan *data training* dengan labelnya yang sesuai. Di tahap ini model klasifikasi mempelajari dari atribut dataset lalu memahami *classification rule nya.* Pada penelitian ini menggunakan algoritma SVM untuk metode klasifikasi.
2. *Classification step*

   Merupakan suatu proses untuk menguji suatu model yang sudah dilakukan pembelajaran di tahap sebelumnya yaitu di *learning step.* Penggunaan model untuk mengklasifikasikan data yang baru. Data yang diberikan akan diprediksi oleh model dengan memberikan label ke kelas tertentu dari hasil perhitungan.

Pada klasifikasi dibagi menjadi dua macam yakni klasifikasi dua kelas dan multi kelas [12]. Yang pertama, klasifikasi dua kelas dapat dijabarkan sebagai berikut. Dimisalkan terdapat set data *training* (x_i , y_i), $i = 1, \dots , l$ dengan data input $X = \{x_1, x_2, \dots , x_l\} \subseteq \Re^N$ dan output yang bersangkutan $Y = \{y_1, \dots , y_l\} \subseteq \{\pm\}\ l$. Tujuan dari klasifikasi dua kelas adalah menemukan suatu fungsi keputusan (decision fuction) f(x) yang secara akurat memprediksi kelas dari data test (x,y) yang berasal dari fungsi distribusi yang sama dengan data untuk training, lihat

Gambar 2.4-1. Set data (x_i , y_i), i = 1, … , l biasa dinamakan set training, dimana xi berkaitan dengan parameter input dan yi menunjukkan parameter output.



Gambar 2.4-1 Ilustrasi Klasifikasi Dua Kelas

Untuk klasifikasi Multi Kelas, misalkan kita memiliki set data untuk training (x_i , y_i), i = 1, … , l dengan data input $X = \{x_1, x_2, . . ,l\} \subseteq \Re^N$ dan output yang bersangkutan $Y = \{y1, . . , y_l\} \subseteq \{1,2, \dots , k\}_1$. Terlihat bahwa output Y tidak lagi terbatas ±1 seperti dalam kasus dua kelas. Output dari data kita bisa 1,2,3,4 atau bahkan 10. Ilustrasi ditampakkan dalam Gambar 2.2. Dalam hal ini kita harus mengelompokkan obyek yang kita pelajari ke dalam lebih dari dua kelas atau ke dalam k kelas dimana nilai k lebih dari 2. Dalam kasus demikian tugas klasifikasi menjadi lebih rumit dan perlu teknik khusus untuk mengatasinya.

Gambar 2.4-2 Ilustrasi Klasifikasi Multikelas

2.5 Metode Regresi

Selain klasifikasi terdapat metode lain juga seperti metode regresi. Regresi adalah teknik *data mining* untuk memprediksi nilai ril (*continous value*) dari dataset yang diberikan. Seperti contoh aplikasinya ialah untuk memprediksi biaya produk atau layanan dengan diberikan variabel lain.

Biasanya proses dari metode regresi meliputi variabel *predictor* (nilai yang diidentifikasi oleh user) dan variabel respon (nilai yang akan diprediksi). Secara umum metode regresi terdapat dua jenis :

2.5.1 *Linear Regression Model*

Linear Regression biasa digunakan untuk mencari hubungan korelasi antara dua variabel. Biasa dilakukan dengan menggunakan persamaan linear terhadap dataset untuk mendapat interpretasi.

Selain itu juga, jenis regresi ini dapat digunakan untuk mencari persamaan matematika dari hubungan antar variabel. Jenis ini merupakan paling sederhana dari metode regresi.

Dalam kasus tertentu, ketika hasilnya berupa garis kurva, maka modelnya dianggap sebagai *non-linear* dan ketika yang dioberservasi berupa model *linear* maka hasilnya akan berupa garis lurus. Formula untuk *linear regression model* ditulis sebagai berikut :

$$y = bx + a$$

Dimana y merupakan fungsi model *linear* dari variabel x, b merupakan gradien garis *linear* dan a merupakan garis potong.

2.5.2 *Multiple Regression Model*

Multiple regression model secara umum digunakan untuk menjelaskan hubungan korelasi antara beberapa variabel independen atau variabel *predictor*. Model ini bisa dibilang model paling popular untuk prediksi regresi pada *data mining.*

Secara umum, diperlukan dua atau lebih variabel independen untuk memprediksi hasil untuk pengguna. Persamaan dari *multiple regression model* ialah :

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \ldots + a_kx_k + e$$

Dimana y merupakan variabel respon (nilai yang akan diprediksi) dan x_i, i = 1,2, …, k merupakan variabel independen. variabel e merupakan nilai error dari persamaan model diatas dan a_i, i = 0,1,2, …, k merupakan koefisien regresi.

2.6 *Support Vector Machine* (SVM)

Di kajian bagian ini akan dibahas dasar teori mengenai *Support Vector Machine* (SVM). Adapun penjelasannya terdiri dari definisi, non-linear SVM, multi-kelas SVM, kelebihan dan kekurang algoritma SVM.

2.6.1 Definisi

Suppport Vector Machine (SVM) merupakan algoritma *supervised machine learning* yang bisa digunakan untuk masalah regresi ataupun klasifikasi, akan tetapi secara umum biasanya digunakan untuk klasifikasi. [8]

Properti khusus dari algoritma SVM ialah algoritma ini secara bersamaan meminimalkan *empirical classification error* dan memaksimumkan *geometric margin*. Sehingga terkadang algoritma SVM disebut sebagai *Maximum Margin Classifier*. Algoritma SVM didasarkan pada *Structural Risk Minimization* (SRM). SVM menerima input vektor ke dimensi tipe data yang lebih tinggi yang dimana *maximal separating hyperplane* dibuat yaitu bisa sebut sebagai garis pemisah antara kelas yang berbeda. Asumsi dari *hyperplane* ialah bahwa semakin besar jarak margin atau jarak antara dua paralel hyperplane maka semakin baik *generalization errornya*. Dimisalkan terdapat beberapa poin sebagai berikut :

$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), ..., (x_n, y_n)\}$

Dimana $y_n = 1 / -1$, suatu konstan yang mendenotasikan suatu kelas dimana titik x_n berada dan n merupaka total sample yang ada. Tiap x_n adalah p-dimensi vektor. Skala suatu value sangatlah penting apabila variabel (atribut) memiliki variansi berbeda – beda, persamaan garis hyperplane bisa dituliskan sebagai berikut :

$$w.x + b = 0$$

Dimana b merupaka satuan skalar dan w merupakan vektor dengan p-dimensi. Vektor w memiliki sudut tegak lurus terhadap garis hyperplane. Ditambah dengan offset b menaikkan jarak margin. Apabila b tidak ada (nol), garis hyperplane melewati titik pusat. Persamaan parallel hyperplane dapat dituliskan sebagai berikut :

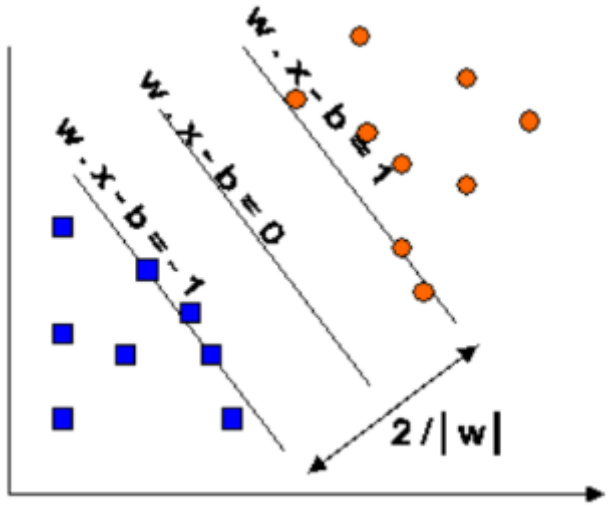
$$w.x + b = 1$$

$$w.x + b = -1$$

Jika *training dataset*-nya dapat dipisah secara linear, maka hyperplane dapat dibuat sehingga tidak ada titik diantara kedua garis hyperplane diatas dan bisa dimaksimumkan jarak antaranya. Dengan geometri, dapat dicari jarak antara dua garis paralel tersebut ialah $2 / |w|$. Sehingga $|w|$ harus diminimalkan untuk mendapatkan jarak maksimum. Maka dari itu persamaan dapat ditulis sebagai berikut :

$$y_i(w.x_i-b) \geq 1,\ 1 \leq i \leq n$$

Sehingga dua paralel garis hyperplane pada visulisasi dapat digambarkan sebagai berikut :

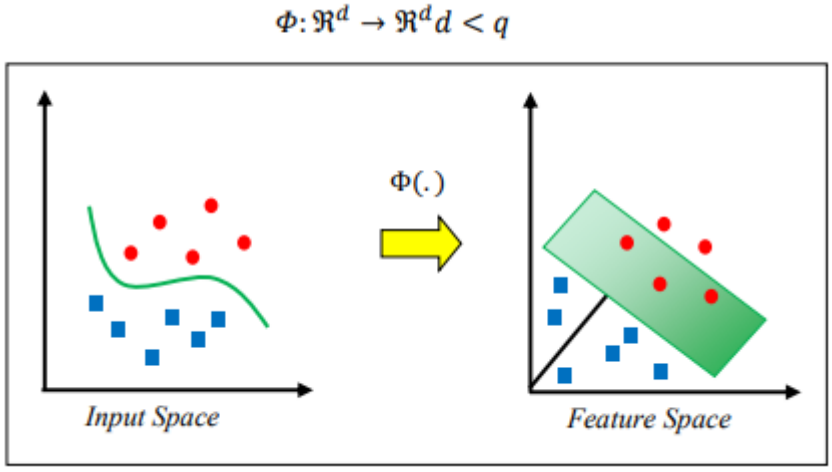


Gambar 2.6-1 Visualiasi Algortima SVM

Dan yang dimaksud dari *support vector* sederhananya ialah koordinat dari individual observasi data poin.

2.6.2 Non-Linear *Classification*

Pada umumnya masalah di dunia nyata jarang yang bersifat linear *separable* tetapi bersifat non-linear. Untuk menyelesaikan masalah tersebut (non-linear), SVM dimodifikasi dengan memasukkan fungsi kernel [12]. Dalam non-linear SM, pertama-tama data x dipetakan oleh fungsi $\Phi(x)$ ke ruang vektor yang berdimensi lebih tinggi. Hyperplane yang memisahkan kedua kelas tersebut dapat dikontruksikan. Selanjutnya gambar 2.5-2 menunjukkan bahwa fungsi Φ memetakan tiap data pada input space tersebut ke ruang vektor baru yang berdimensi lebih tinggi yaitu dimensi 3, sehingga kedua kelas dapat dipisahkan secara linear oleh sebuah hyperplane. Notasi matematika dari mapping ini adalah sebagai berikut:



Gambar 2.6-2 Fungsi Φ memetakan data ke ruang vector yang berdimensi lebih tinggi

Selanjutnya proses pembelajaran pada SVM dalam menemukan titik-titik *support vector*, hanya bergantung pada dot product dari data yang sudah ditransformasikan pada ruang baru yang berdimensi lebih tinggi, yaitu : $\Phi(x_i)$. $\Phi(x_j)$.

Karena umunya transformasi Φ ini tidak dapat diketahui dengan mudah maka perhitungan *dot product* dapat digantikan dengan fungsi Kernel sehingga persamaan akhir menjadi *support vector*.

2.6.3 Kelebihan dan Kekurang SVM

Berikut adalah kelebihan – kelebihan dari SVM :

1. SVM efektif untuk dataset yang berada pada *high-dimensional space* karena memiliki fungsi kernel yang didesain untuk masalah tersebut.
2. SVM efektif ketika ukuran dimensi data lebih besar daripada jumlah total dari sample, yang biasa disebut *Curse of Dimensionality*.
3. SVM menggunakan *subset training point* di fungsi pengambilan keputusannya (biasa dipanggil *support vector*), sehingga memori efisien

Berikut adalah kekurangan dari SVM :

1. Performa SVM menurut apabila ukuran dataset terlalu besar dikarenakan waktu yang diperlukan untuk *training* akan sangat lama.
2. SVM tidak berperforma dengan baik apabila datanya memiliki *noise* banyak, seperti contohnya target kelas yang saling *overlap*.

2.7 *Support Vector Regression* (SVR)

Support Vector Regression merupakan metode untuk masalah regresi dan algoritma ini hasil pengembangan dari *Support Vector Machine* (SVM). Tujuan dari algoritma SVR adalah menemukan fungsi *f(x)* sebagai suatu *hyperplance* (garis pemisah) berupa fungsi regresi yang sesuai dengan semua input data dengan sebuah error ε dan membuat ε sekecil mungkin [15].

Tujuan dari SVR ialah untuk memetakan *vector* input ke dalam dimensi yang lebih tinggi. Dimisalkan terdapat $\boldsymbol{l}$ data latihan, $(\boldsymbol{x_I}, \boldsymbol{y_i})$, i=1, …, $\boldsymbol{l}$ dimana $\boldsymbol{x_i}$ merupakan *vector input* $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\} \subseteq \boldsymbol{R}$ dan $\boldsymbol{l}$ adalah banyaknya data training. Dengan algoritma SVR ditentukan suatu fungsi *f(x)* yang mempunyai deviasi paling besar dari target sebenarnya $(\boldsymbol{y_i})$, untuk semua data training. Jika nilai ε = 0 maka diperloeh suatu persamaan regresi yang sempurna dengan metode SVR seperti berikut :

$$\boldsymbol{f(x) = w^T \varphi(x) + b}$$

Dengan keterangan :

$\boldsymbol{w}$ = *vector* berbobot berdimensi *l*

$\boldsymbol{\varphi(x)}$ = fungsi yang menentukan x pada ruang dengan *l* dimensi

$\boldsymbol{b}$ = bias

2.8 Fungsi Kernel

Pada masalah klasifikasi biasanya diasumsikan kelinierannya sehingga algoritma yang dihasilkan terbatas untuk kasus linier saja. Akan tetapi untuk masalah selain linier bisa menggunakan bantuan dari fungsi kernel. Kernel *trick* memberikan berbagai kemudahan, karena dalam proses pembelajaran SVM. Untuk menentuka *support vector*, maka cukup dengan mengetahui fungsi kernel yang dipakai dan tidak perlu mengetahui wujud dari fungsi non-linear.

Ada beberapa fungsi kernel yang sering digunakan dalam literatur SVM antara lain sebagai beriut :

1. Kernel linear adalah kernel yang paling sederhana dari semua fungsi kernel. Kernel ini biasa digunakan untuk masalah kasus klasifikasi teks dan sejenisnya.
2. Kernel *polynomial* adalah kernel yang sering digunakan untuk klasifikasi gambar dan masalah lain dengan data dua dimensi.

3. Kernel *Radial Basis Gaussian* adalah kernel yang umum digunakan untuk data yang sudah valid (*available)* dan merupakan default dalam tools SVM.
4. Kernel *Tangent Hyperbolic* adalah kerel yang sering digunakan untuk *neural network.*

Pemilihan fungsi kernel yang tepat adalah hal penting karena akan menentukan *feature space*, dimana fungsi *classifier* akan dicari. Sepanjang fungsi kernelnya cocok, SVM akan beroperasi secara benar meskipun tidak tau pemetaan yang digunakan. Fungsi kernel *Gaussian RBF* memiliki kelebih yaitu otomatis menentukan nilai rentang tak terhingga. Kernel tersebut juga efektif menhindari overfitting.

2.9 Evaluasi Ukuran Performa Metode Klasifikasi

Akurasi klasifikasi adalah ukuran ketepatan klasifikasi dalam performa teknik klasifikasi secara keseluruhan. Biasanya untuk masalah klasifikasi digunakan *confusion matrix.*

Confusion Matrix adalah pengukuran performa untuk masalah *mahine learning* dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix* berguna untuk evaluasi dari performa model yang telah dirancang apakah sudah mencapai tujuan objektifnya. Tabel *confusion matrix* diperlihatkan pada tabel Tabel 2.6-1.

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Tabel 2.9-1 *Confusion Matrix*

Keterangan dari tabel tersebut :

TP : *True Positive* (Jumlah prediksi benar pada kelas positif)

FP : *False Positive* (Jumlah prediksi salah pada kelas positif)

FN : *False Negative* (Jumlah prediksi salah pada kelas negatif)

TN : *True Negative* (Jumlah prediksi benar pada kelas negatif)

Nilai TP dan TN menunjukkan tingkat ketepatan klasifikasi. Biasanya semakin tinggi nilai tersebut maka semakin baik pula tingkat klasifikasi dari akurasi, recall, dan presisi. Sebagai contoh apabila ada data yang nilai sebenarnya bernilai *false* tetapi diprediksi dengan nilai *true* disebut sebagai FP. Sedangkan sebaliknya, jika ada data dengan nilai sebenarnya *true* tetapi diprediksi dengan nilai *false* maka disebut FN. Dan dari Tabel 2.6-1 bisa didapatkan perhitungan evaluasi performansi metode klasifikasi dengan formula berikut :

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$