BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka adalah penelitian-penelitian yang telah dilakukan sebelumnya oleh orang lain dan dapat digunakan untuk mendukung penelitian yang sedang dilakukan sekarang. Tinjauan pustaka ini memuat ulasan dan analisis terhadap berbagai literatur terkait yang telah dipublikasi sebelumnya. Penulis telah mengumpulkan beberapa tinjauan pustaka yang dapat dilihat pada tabel dibawah ini

Tabel 2. 1 Tinjauan Pustaka

No	Nama Peneliti Dan Tahun	Judul	Metode Penelitian	Hasil Penelitian
1	(Moch. Rizky Yuliansyah, B and Franz, 2022)	Perbandingan Metode *K-Nearest Neighbors* dan *Naïve Bayes Classifier* Pada Klasifikasi Status Gizi Balita di Puskesmas Muara Jawa Kota Samarinda	*K-Nearest Neighbors dan Naïve Bayes Classifier*	Berdasarkan hasil perbandingan performa antara metode *K-Nearest Neighbors* dan *Naïve Bayes Classifier* menggunakan f1 score sebagai patokan utama performa klasifikasi. Yaitu dengan perolehan nilai metode K-Nearest Neighbor sebesar 45,95% dan Naïve Bayes 32,53%. K-Nearest Neighbors unggul pada f1 score dengan selisih cukup besar yakni 13,42 %. Sehingga pada masalah klasifikasi status gizi balita metode *K-Nearest Neighbors* mengungguli *Naïve Bayes Classifier.*

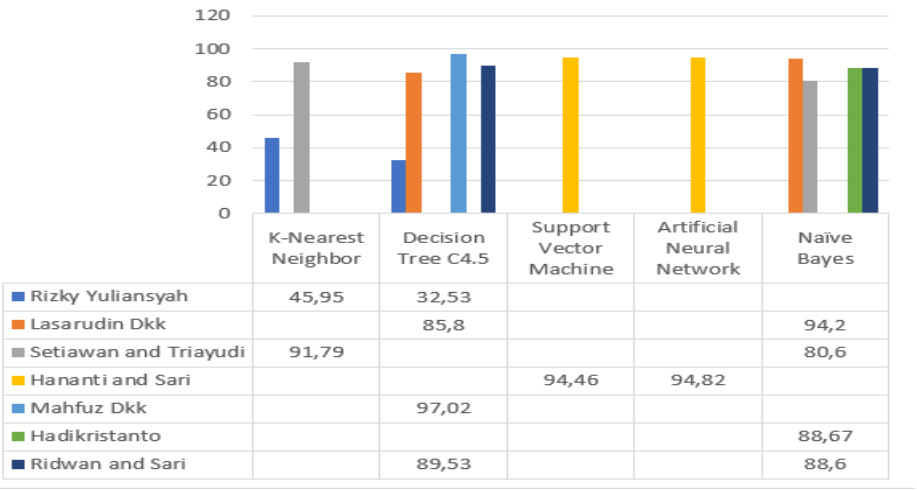
Tabel 2. 2 Tinjauan Pustaka (Lanjutan)

No	Nama Peneliti Dan Tahun	Judul	Metode Penelitian	Hasil Penelitian
2	(Lasarudin, Gani and Tomayahu, 2022)	Perbandingan Metode *Naïve Bayes* dan C4.5 Klasifikasi Status Gizi Bayi Balita	*Metode Naïve Bayes* dan C4.5	Hasil akurasi yang didapatkan pada penelitian ini diperoleh nilai akurasi metode *Naïve Bayes* sebesar 94.20% dan metode C4.5 sebesar 85,80% dengan menggunakan 90% data training dan 10% data testing
3	(Setiawan and Triayudi, 2022)	Klasifikasi Status Gizi Balita Menggunakan *Naive Bayes* dan *K Nearest Neighbor* Berbasis Web	*Metode Naive Bayes* dan *K Nearest Neighbor*	Pada penelitian ini Menggunakan 412 data gizi balita didapatkan hasil dengan Naïve Bayes medapatkan akurasi 80.60% sedangakn dengan *K-Nearest Neighbor* didapatkan akurasi 91.79%. kemudian dari metode yang menghasilkan nilai akurasi tinggi yaitu metode *K-Nearest Neighbor* selanjutnya diterapkan kedalam aplikasi berbasis web.
4	(Hananti and Sari, 2021)	Perbandingan Metode *Support Vector Machine* (SVM) dan *Artificial Neural Network* (ANN) pada Klasifikasi Gizi Balita	*Metode Support Vector Machine* dan *Artificial Neural Network*	Hasil penelitian ini Untuk ukuran ketepatan klasifikasi pada metode ANN, yaitu accuracy sebesar 94,82%, precision sebesar 51.00%, recall sebesar 51.09%, dan AUC sebesar 0.910, sedangkan pada metode SVM, yaitu accuracy sebesar 94,46%, precision sebesar 46.08%, recall sebesar 50.59%, dan AUC sebesar 0.900.

Tabel 2. 3 Tinjauan Pustaka (Lanjutan)

No	Nama Peneliti Dan Tahun	Judul	Metode Penelitian	Hasil Penelitian
5	(Mahfuz, Amri Muliawan Nur, 2022)	Penerapan Algoritma C4.5 Dalam Mengklasifikasi Status Gizi Balita Pada Posyandu Desa Dames Damai Kabupaten Lombok Timur	*Metode C4.5*	Hasil penelitian ini menggunakan metode Decision Tree C4.5 dan memperoleh hasil akurasi yang diperoleh sangat baik atau cukup sempurna yaitu 97.02%.
6	(Hadikristanto and Pungkas, 2019)	Klasifikasi status gizi orang dewasa menggunakan algoritma *Naive Bayes* (studi kasus klinik bhakti mulia cikarang)	*Metode Naive Bayes*	Pada penelitian ini menggunakan algoritma *naïve bayes* digunakan untuk menentukan status gizi orang dewasa. Data diklasifikasikan menjadi tiga, yaitu Normal, Kurang, dan Obesitas. Uji coba ini dilakukan dengan 150 data, dan hasil akurasi yang didapat sebesar 88,67%.
7	(Ridwan and Sari, 2021)	*The comparison of accuracy between naïve bayes clasifier and c4.5 algorithm in classifying toddler nutrition status based on anthropometry index*	*Metode Naïve Bayes* dan C4.5	Penelitian ini membandingkan *Naïve Bayes Classifier* dan Algoritma C4.5 yang diimplementasikan dengan dataset Status Gizi Balita berdasarkan Indeks Antropometri. menghasilkan akurasi sebesar 88,60%. Dan algoritma C4.5 menghasilkan akurasi 89,53%.

Yang membedakan dari beberapa penelitian yang terdahulu yaitu dari metode yang digunakan, jumlah data dan dari atribut yang digunakan untuk melakukan klasifikasi status gizi balita pada penelitian yang dilakukan oleh (Moch. Rizky Yuliansyah, B and Franz, 2022) menggunakan Metode *K-Nearest Neighbors* dan *Naïve Bayes Classifier* Jumlah data yang digunakan dalam penelitian ini adalah 1443 balita dengan proporsi pembagian data yaitu 80:20, dimana 80% untuk training dan 20% untuk testing. Pada penelitian yang dilakukan oleh (Lasarudin, Gani and Tomayahu, 2022) melakukan perbandingan akurasi metode Naïve bayes dan metode C4.5 untuk mengklasifikasi status gizi pada bayi dan balita. Pada penelitian yang dilakukan oleh (Setiawan and Triayudi, 2022) menggunakan dua metode sekaligus untuk membandingkan mana yang lebih optimal yaitu menggunakan *Metode Naive Bayes* dan *K Nearest Neighbor* Hasilnya metode Naïve Bayes medapatkan akurasi 80.60% sedangakan dengan K-Nearest Neighbor didapatkan akurasi 91.79%. kemudian dari hasil penelitiannya metode yang menghasilkan nilai akurasi tinggi diterapkan dengan melakukan pembuatan aplikasi berbasis web. pada penelitian (Hananti and Sari, 2021) pada penelitian ini hanya menggunakan dua kelas untuk status gizi nya yaitu gizi baik dan gizi kurang. Selanjutnya pada penilitian yang dilakukan oleh (Hadikristanto and Pungkas, 2019) penelitian ini menggunakan algoritma *naïve bayes* yang digunakan untuk menentukan status gizi orang dewasa. Data diklasifikasikan menjadi tiga, yaitu Normal, Kurang, dan Obesitas. Uji coba ini dilakukan dengan 150 data, dan hasil akurasi yang didapat sebesar 88,67%.



Gambar 2. 1 Diagram Hasil Akurasi Penelitian Terdahulu

Dari penelitian – penelitian tersebut akan digunakan oleh penulis sebagai bahan referensi dalam menghasilkan tingkat akurasi dalam klasifikasi status gizi balita. Dari penelitian diatas penulis akan mencoba membandingkan nilai akurasi dari algoritma *K-Nearest Neighbor* yang penulis lakukan dan melakukan percobaan dengan menggunakan algoritma *Decision Tree C4.5* dan *Support Vector Machine* untuk membandingkan nilai akurasi dari ketiga algoritma. Tools yang akan digunakan oleh penulis sendiri adalah aplikasi RapidMiner.

2.2 Status Gizi

2.2.1 Pengertian Status Gizi

Status gizi merupakan keadaan kesehatan tubuh seseorang atau sekelompok orang yang diakibatkan oleh konsumsi, penyerapan, dan penggunaan zat gizi makanan. Status gizi seseorang atau sekelompok orang dapat digunakan untuk mengetahui apakah seseorang atau sekelompok orang tersebut keadaan gizinya baik atau sebaliknya. Masalah gizi dipengaruhi banyak faktor dan saling mempengaruhi. Salah satunya adalah faktor genetik dari orang tua, yaitu faktor tinggi dan berat badan orang tua. Selain itu, faktor pendidikan, ketersediaan pangan di tingkat rumah tangga, pola asuh konsumsi makanan, pola makanan, kepercayaan, tradisi atau budaya, dan lain sebagainya. Beberapa hasil penelitian lain yang menyatakan bahwa status gizi disebabkan oleh karakteristik orang tua seperti ukuran antropometri ibu dan bapak, seperti tinggi badan orang tua memungkinkan anak memiliki risiko gagal pertumbuhan serta mengalami *underweight* (Hadikristanto and Pungkas, 2019).

2.2.2 Penilaian Status Gizi

Pada dasarnya penilaian status gizi dibagi menjadi dua yaitu penilaian secara langsung dan tidak langsung.

a. Penilaian Status Gizi Secara Langsung

Penilaian status gizi secara langsung dibedakan menjadi empat penilaian yaitu:

1. Antropometri

Metode ini dilakukan beberapa macam pengukuran antara lain pengukuran berat badan (BB), tinggi badan (TB), dan lingkar lengan atas. Pengukuran tersebut

di atas paling sering dilakukan dalam survei gizi terhadap balita berdasarkan kelompok umurnya. Dalam ilmu gizi, maka status gizi tidak hanya diketahui dengan mengukur berat badan (BB) atau tinggi badan (TB) berdasarkan umur secara sendiri-sendiri, tetapi juga dalam bentuk indikator yang dapat merupakan kombinasi dari ketiganya. Indikator yang dapat mempengaruhi status gizi antara lain penyebab langsung yaitu makanan dan penyakit infeksi yang mungkin diderita. Timbulnya gizi kurang bukan saja karena makanan yang kurang tetapi juga karena penyakit dan penyebab tidak langsung yaitu ketahanan pangan di keluarga, pola pengasuhan anak, pelayanan kesehatan dan kesehatan lingkungan merupakan faktor yang saling berhubungan.

a. Indikator berat- badan/usia (BB/U) menunjukkan secara sensitif status gizi saat ini (saat diukur) karena mudah berubah, namun tidak spesifik karena berat badan selain dipengaruhi oleh umur juga dipengaruhi oleh tinggi badan. Indikator ini dapat dengan mudah dan cepat dimengerti oleh masyarakat umum, dan cukup sensitif untuk melihat perubahan status gizi dalam jangka waktu pendek. Selain itu pengukuran antropometrik dapat mendeteksi kegemukan. Indikator BB/U digunakan untuk menentukan kategori berat badan sangat kurang (*severely underweight*), berat badan kurang (*underweight*), berat badan normal dan risiko berat badan lebih.
b. Indikator TB/U dapat menggambarkan status gizi masa lampau atau masalah gizi seseorang yang pendek kemungkinan keadaan gizi masa lalu tidak baik. Berbeda dengan berat badan yang dapat diperbaiki dalam waktu singkat, baik pada anak maupun dewasa, maka tinggi badan pada usia dewasa tidak dapat lagi dinormalkan. Kemungkinan untuk mengejar pertumbuhan tinggi badan optimal pada anak balita masih bisa sedangkan anak usia sekolah sampai remaja kemungkinan untuk mengejar pertumbuhan tinggi badan masih bisa tetapi kecil kemungkinan untuk mengejar pertumbuhan optimal. Secara normal tinggi badan tumbuh bersamaan dengan bertambahnya umur. Pertambahan TB relatif kurang sensitif terhadap kurang gizi dalam waktu singkat. Pengaruh kurang gizi terhadap pertumbuhan TB baru terlihat dalam waktu yang cukup lama. Indikator ini juga dapat dijadikan indikator keadaan sosial ekonomi

penduduk. Indikator TB/U digunakan untuk menentukan kategori sangat pendek (*severely stunted*), pendek (*stunted*), normal dan tinggi.

c. Indikator BB/TB merupakan pengukuran antropometri yang terbaik karena dapat menggambarkan secara sensitif dan spesifik status gizi saat ini atau masalah gizi akut. Berat badan berkorelasi linier dengan tinggi badan, artinya dalam keadaan normal perkembangan berat badan akan mengikuti pertambahan tinggi badan pada percepatan tertentu. Hal ini berarti berat badan yang normal akan proporsional dengan tinggi badannya. Ini merupakan indikator yang baik untuk menilai status gizi saat ini terutama bila data umur yang akurat sering sulit diperoleh. WHO dan UNICEF merekomendasikan menggunakan indikator BB/TB dengan cut of point < -3 Standar Deviasi (SD) dalam kegiatan identifikasi dan manajemen penanganan bayi dan anak balita gizi buruk akut(Arlius, 2017). Indikator BB/TB digunakan untuk menentukan kategori gizi buruk (*severely wasted*), gizi kurang (*wasted*), gizi baik (normal), berisiko gizi lebih (*possible risk of overweight*), gizi lebih (*overweight*) dan obesitas (*obese*).

2. Klinis

Penilaian secara klinis merupakan metode yang sangat penting untuk menilai status gizi masyarakat. Metode ini didasarkan pada perubahan-perubahan yang terjadi yang dihubungkan dengan ketidakcukupan zat gizi. Hal ini dapat dilihat pada jaringan epitel (*superficial epithelial tissues*) seperti kulit, mata, rambut, atau pada organ-organ yang dekat dengan permukaan tubuh seperti kelenjar tiroid.

3. Biokimia

Penilain dengan biokimia adalah pemeriksaan spesimen yang diuji secara laboratoris yang dilakukan pada berbagai macam jaringan tubuh. Jaringan tubuh yang digunakan yaitu: darah, urine, tinja, hati dan otot. Metode ini digunakan untuk suatu peringatan bahwa kemungkinan akan terjadi keadaan malnutrisi yang lebih parah lagi. Banyak gejala klinis yang kurang spesifik, maka dari itu penentuan kimia faali dapat lebih menolong untuk menentukan kekurangan gizi yang spesifik.

4. Biofisik

Penentuan status gizi secara biofisik merupakan metode penentuan status gizi dengan melihat kemampuan fungsi (khususnya jaringan) dan melihat perubahan

struktur dari jaringan. Umumnya dapat digunakan dalam situasi tertentu seperti kejadian buta senja epidemik (*epidemic of night blindness*). Cara yang digunakan untuk mengatasinya adalah tes adaptasi gelap.

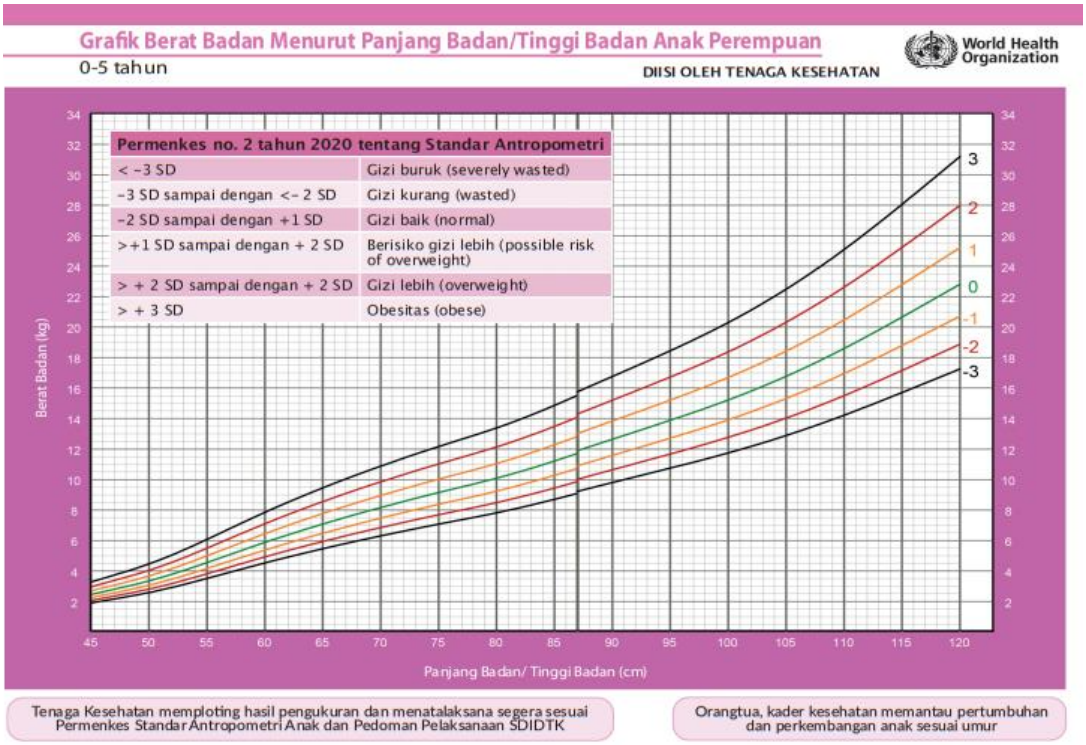
5. Menggunakan Buku Kesehtan Ibu dan Anak (KIA)

a. Pengertian Buku Kesehtan Ibu dan Anak

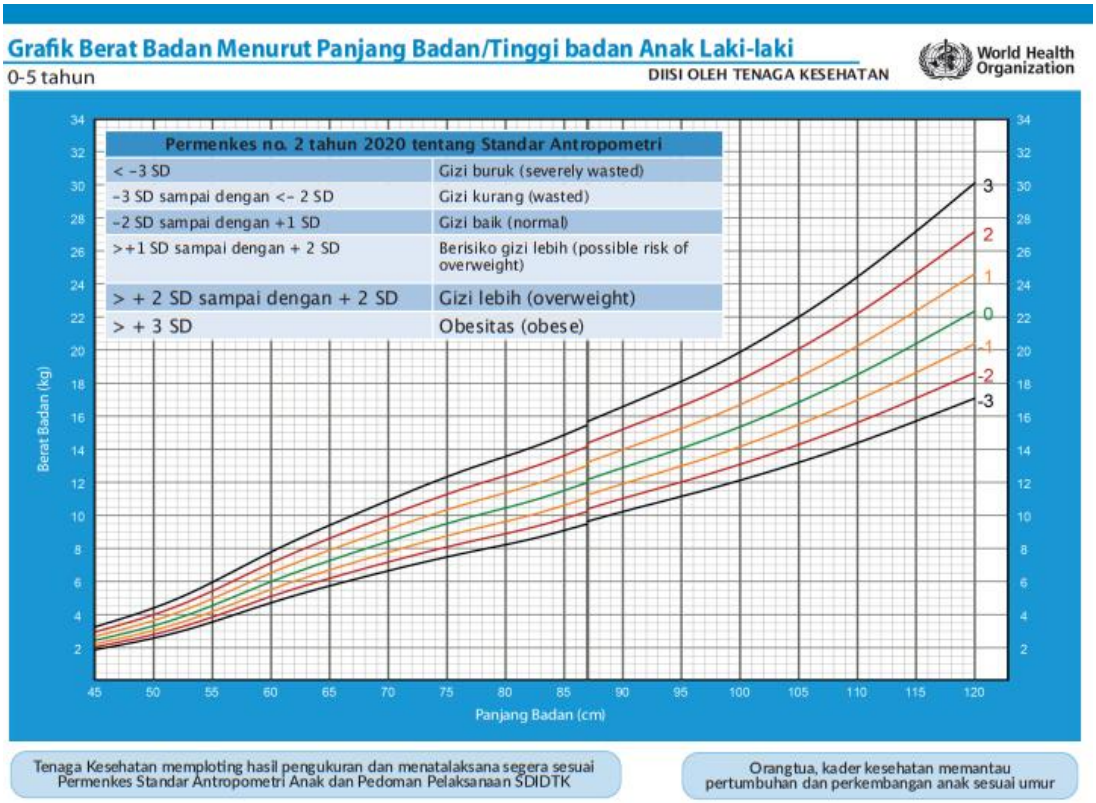
adalah buku yang berisi catatan kesehatan ibu mulai dari hamil, bersalin, nifas, dan catatan kesehatan anak mulai dari bayi baru lahir hingga balita, serta berbagai informasi cara merawat kesehatan ibu dan anak. Pengukuran status gizi berdasarkan Berat Badan per Tinggi Badan. Berikut adalah tampilan gambar grafik berat badan menurut panjang badan/tinggi badan untuk anak perempuan dan anak laki-laki.

b. Tujuan Buku KIA

Tujuan buku KIA adalah untuk memudahkan keluarga dalam memahami informasi kesehatan tentang ibu dan anak yang terdapat pada buku KIA, memudahkan tugas Ibu untuk memahami kondisi kesehatan Ibu dan bayi secara mandiri, serta untuk meningkatkan praktik keluarga dan masyarakat dalam memelihara/merawat kesehatan ibu dan anak



Gambar 2. 2 Grafik Penentuan Status Gizi Balita Perempuan



Gambar 2. 3 Grafik Penentuan Status Gizi Balita Laki-laki

c. Komponen Grafik

- Berat badan (kg): dengan skala 1 sampai dengan 18
- Umur (bulan): dari umur 0 hingga 60 bulan
- Grafik warna ungu untuk pengukuran balita perempuan
- Grafik warna biru untuk pengukuran balita laki-laki

d. Cara Membaca Buku KIA

Setelah anak diukur panjang badan dan ditimbang berat badannya selanjutnya diukur menggunakan grafik sesuai dengan jenis kelaminnya. Kemudian ditarik garis lurus antara berat badan dan panjang badan sampai bertemu dititik tengah. Setalah itu dicocokan dengan tabel standar Antropometri yang tertera digrafik masing-masing.

e. Penilaian Status Gizi Secara Tidak Langsung

Penilaian status gizi secara tidak langsung dibedakan menjadi tiga penilaian yaitu:

1. Survei Konsumsi Makanan

Metode ini melihat jumlah dan jenis zat gizi yang dikonsumsi. Survei ini

dapat mengidentifikasi kelebihan dan kekurangan zat gizi. Metode penelitian asupan makanan dibagi menjadi dua kelompok yaitu metode kuantitatif meliputi food recall, estimated food record dan food weighting serta metode kualitatif seperti *dietary history* dan *food frequency*.

2. Statistik Vital

Penilaian status gizi menggunakan statistik vital yaitu dengan menganalisis data statistik kesehatan seperti angka kematian berdasarkan umur, angka kesakitan, kematian akibat penyebab tertentu dan data lainnya yang berhubungan dengan gizi. Penggunaannya dipertimbangkan sebagai bagian dari indikator tidak langsung pengukuran status gizi.

3. Faktor Ekologi

Faktor ekologi digunakan untuk mengungkapkan bahwa malnutrisi sebagai hasil interaksi beberapa faktor fisik, biologis dan lingkungan budaya. Jumlah makanan yang tersedia sangat tergantung dari keadaan ekologi seperti iklim, tanah, irigasi, dan lain-lain. Pengukuran faktor ekologi dipandang sangat penting untuk mengetahui penyebab malnutrisi di suatu masyarakat sebagai dasar untuk melakukan program intervensi gizi.

2.2.3 Metode Penilaian Antropometri Status Gizi

Metode penilaian status gizi merupakan cara untuk menilai keadaan gizi pada seseorang. Maka dari itu untuk dapat mengetahui keadaan gizi pada seseorang dapat dilihat dari status gizinya. Metode penilaian status gizi menggunakan metode antropometri. Antropometri berasal dari kata anthropos yang berarti manusia, dan metri adalah ukuran. Metode antropometri dapat diartikan sebagai mengukur fisik dan bagian tubuh manusia. Jadi antropometri adalah metode penilaian status gizi dengan menggunakan pengukuran melalui ukuran fisik dan bagian tubuh manusia untuk menentukan status gizi pada seseorang.Konsep dasar antropometri yakni konsep dasar pertumbuhan. Pertumbuhan adalah terjadinya perubahan sel-sel tubuh, terdapat dua bentuk yaitu bertambahnya jumlah sel dan atau terjadinya pembelahan sel, secara akumulasi menyebabkan terjadinya perubahan ukuran tubuh. Agar pertumbuhan seorang anak dapat berkembang dengan pesat yakni dengan memenuhi asupan gizi yang seimbang antara kebutuhan gizi dengan asupan gizinya.

Gizi yang tidak seimbang dapat mengakibatkan terjadinya gangguan pertumbuhan. Kekurangan gizi dapat menghambat pertumbuhan anak. Oleh karena itu antropometri dapat dijadikan salah satu metode penilaian terhadap status gizi pada anak dengan cara mengukur pertumbuhan dari pada ukuran fisik dan bentuk tubuhnya. Parameter yang digunakan untuk pengukuran dengan metode antropometri yang sering digunakan untuk menentukan status gizi misalnya berat badan, tinggi badan, ukuran lingkar kepala, ukuran lingkar dada, ukuran lingkar lengan atas dan lain-lain. Hasil ukuran antropometri tersebut kemudian dirujukkan pada standar atau rujukan pertumbuhan manusia.

1. Berat Badan

Berat badan menggambarkan jumlah protein, lemak, air dan mineral yang terdapat dalam tubuh.berat badan merupakan komposit pengukuran ukuran total tubuh. Beberapa alasan mengapa berat badan digunakan sebagai parameter antropometri. Alasan tersebut diantaranya adalah perubahan berat badan mudah terlihat dalam waktu singkat dan menggambarkan status gizi saat ini. Pengukuran berat badan mudah dilakukan dan alat ukur untuk menimbang berat badan mudah untuk diperoleh.

2. Tinggi Badan

Tinggi badan atau panjang badan menggambarkan ukuran pertumbuhan massa tulang yang terjadi akibat dari asupan gizi. Oleh karena itu tinggi badan digunakan sebagai parameter antropometri untuk menggambarkan pertumbuhan linier.pertambahan tinggi badan atau panjang badan terjadi dalam waktu yang ama sehingga sering disebut akibat masalah gizi kronis.

3. Lingkar Kepala

Lingkar kepala dapat digunakan sebagai pengukuran ukuran pertumbuhan lingkar kepala dan pertumbuhan otak, walaupun tidak sepenuhnya berkorelasi dengan volume otak. Pengukuran lingkar kepala merupakan predikator terbaik dalam melihat perkembangan saraf anak dan pertumbuhan global otak dan struktur internal.

4. Lingkar Lengan Atas (LILA)

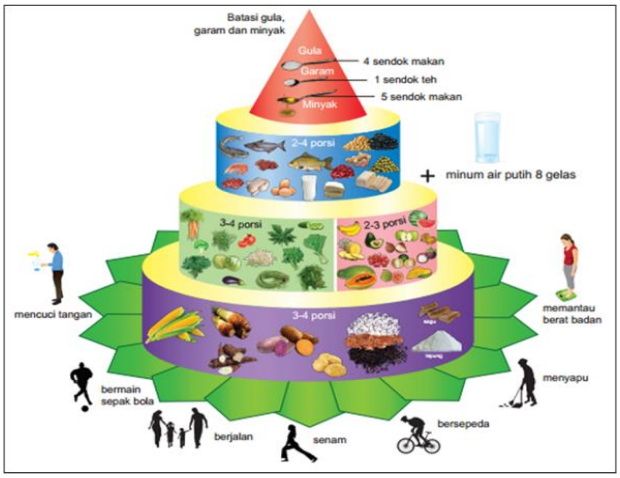
Lingkar lengan Atas (LILA) merupakan gambaran keadaan jaringan otot dan lapisan lemak bawah kulit.

2.2.4 Pedoman Gizi Seimbang di Indonesia

Pedoman gizi di Indonesia mulanya yakni ”Empat Sehat Lima Sempurna”(ESLS). ESLS dicetuskan pada tahun 1952 yang dimotori oleh Prof. Poorwo Soedarmodi. ESLS disempurnakan menjadi gizi seimbang pada tahun 2014, tepatnya di tahun itu adalah penyempurnaan yang kedua penyempurnaan yang pertama dilakukan pada tahun 1995.

Gizi seimbang adalah makanan sehat untuk pemenuhan kebutuhan gizi sehari-hari sesuai dengan jenis dan jumlah yang dibutuhkan oleh tubuh dengan memperhatikan keanekaragaman makanan, aktivitas fisik, kebersihan dan berat badan ideal. Pedoman gizi seimbang ini divisualisasikan dengan tumpeng karena bentuknya yang mengerucut dan disesuaikan dengan kebudayaan Indonesia maka jadilah tumpeng gizi seimbang (TGS). TGS tersebut terdiri dari:

1. potongan besar: golongan makanan karbohidrat.
2. potongan sedang dan 2 potongan kecil yang merupakan golongan sayuran dan buah.
3. 2 potongan kecil diatasnya yang merupakan golongan protein hewani dan nabati.
4. 1 potongan terkecil di puncak yaitu gula, garam dan minyak yang dikonsumsi seperlunya.
5. Potongan TGS juga dilapisi dengan air putih yang idealnya dikonsumsi 2 liter atau 8 gelas sehari.
6. Luasnya potongan TGS ini menunjukkan porsi makanan setiap orang per hari. Karbohidrat 3-8 porsi, sayuran 3-5 porsi, buah 2-3 porsi serta protein hewani dan nabati.
7. Konsumsi ini dibagi untuk makan pagi, siang dan malam. Kombinasi makanan per harinya perlu dilakukan.
8. Di bagian bawah TGS terdapat prinsip gizi seimbang yang lain yakni: pola hidup aktif dengan berolahraga, menjaga kebersihan dan pantau berat badan.



Gambar 2. 4 Piramida Pedoman Gizi

2.3 Konsep Balita

2.3.1 Pengertian Balita

Balita adalah individu atau sekelompok individu dari suatu penduduk yang berada dalam rentan usia tertentu. Usia balita yaitu usia 1-3 tahun(batita) dan anak prasekolah (3-5 tahun). Saat usia batita anak masih tergantung kepada orang tua dalam melakukan kegiatan seperti mandi, makan, buang air. Masa pertumbuhan dan perkembangan anak saat balita memiliki perbedaan sendiri karena mengalami pola pertumbuhan dan perkembangan fisik seperti koordinasi antara motorik halus dan motorik kasar, selain itu juga kecerdasan anak sesuai dengan masa pertumbuhan dan perkembangannya (Mutiara, 2018).

Balita adalah anak yang berumur 0-59 bulan,pada masa ini ditandai dengan proses pertumbuhan dan perkembangan yang sangat pesat dan disertai dengan perubahan yang memerlukan zat-zat gizi yang jumlahnya lebih banyak dengan kualitas yang tinggi. Akan tetapi, balita termasuk kelompok yang rawan gizi serta mudah menderita kelainan gizi karena kekurangan makanan yang dibutuhkan. Konsumsi makanan memegang peranan penting dalam pertumbuhan fisik dan kecerdasan anak sehingga konsumsi makanan berpengaruh besar terhadap status gizi anak untuk mencapai pertumbuhan fisik dan kecerdasan anak (Hondro, 2021). Saat usia batita, anak masih bergantung penuh kepada orang tua untuk melakukan kegiatan penting, seperti mandi, buang air dan makan. Perkembangan berbicara dan berjalan sudah bertambah baik, namun kemampuan lain masih terbatas. Masa balita merupakan periode penting dalam proses tumbuh kembang manusia.

Perkembangan dan pertumbuhan pada masa itu menjadi penentu keberhasilan pertumbuhan dan perkembangan anak pada periode selanjutnya. Masa tumbuh kembang di usia ini merupakan masa yang berlangsung cepat dan tidak akan pernah terulang kembali, karena itu sering disebut *golden age* atau masa keemasan.

2.3.2 Karakteristik Balita

Menurut (Mutiara, 2018) Karakteristik balita dibagi menjadi dua yaitu umur dan jenis

Kelamin :

1. Anak usia 1-3 tahun

Usia 1-3 tahun merupakan konsumen pasif artinya anak menerima makanan yang disediakan orang tuanya. Laju pertumbuhan usia balita lebih besar dari usia prasekolah, sehingga diperlukan jumlah makanan yang relatif besar. Perut yang lebih kecil menyebabkan jumlah makanan yang mampu diterimanya dalam sekali makan lebih kecil bila dibandingkan dengan anak yang usianya lebih besar oleh sebab itu, pola makan yang diberikan adalah porsi kecil dengan frekuensi sering. Jenis kelamin Anak perempuan yang memiliki riwayat BBLR beresiko menjadi ibu stunting. Ibu stunting juga berisiko melahirkan anak stunting.

2. Anak usia prasekolah (3-5 tahun)

Usia 3-5 tahun anak menjadi konsumen aktif. Anak sudah mulai memilih makanan yang disukainya. Pada usia ini berat badan anak cenderung mengalami penurunan, disebabkan karena anak beraktivitas lebih banyak dan mulai memilih maupun menolak makanan yang disediakan orang tuanya.

2.4 Pengertian Puskesmas

Pusat Kesehatan Masyarakat (Puskesmas) adalah salah satu sarana pelayanan kesehatan masyarakat yang amat penting di Indonesia. Puskesmas adalah unit pelaksana teknis dinas kabupaten/kota yang bertanggung jawab menyelenggarakan pembangunan kesehatan di suatu wilayah kerja (Depkes, 2011). Pengertian puskesmas adalah suatu unit pelaksana fungsional yang berfungsi sebagai pusat pembangunan kesehatan, pusat pembinaan peran serta masyarakat dalam bidang kesehatan serta pusat pelayanan kesehatan tingkat pertama yang menyelenggarakan kegiatannya secara menyeluruh, terpadu yang berkesinambungan pada suatu masyarakat yang bertempat tinggal dalam suatu wilayah tertentu (Sari, 2019).

Upaya peningkatan pelayanan kesehatan di puskesmas dirangkum dalam kegiatan pokok puskesmas, antara lain, upaya kesehatan ibu dan anak, upaya keluarga berencana, upaya perbaikan gizi, upaya kesehatan lingkungan upaya pencegahan dan pemberantasan penyakit menular, upaya pengobatan, dan upaya penyuluhan kesehatan masyarakat. Puskesmas merupakan ujung tombak dari penyelenggaraan upaya kesehatan masyarakat maupun perorangan di tingkat pertama. Strategi puskesmas adalah untuk mewujudkan pembangunan kesehatan melalui pelayanan kesehatan yang bersifat menyeluruh (*comprehensive health care service*) serta pelayanan kesehatan yang menerapkan pendekatan yang menyeluruh.

2.5 Data Mining

Data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Darmawan, Kustian and Rahayu, 2018). Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar.

Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Ryanwar, 2020). Data Mining adalah suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (database) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui (Purba, Siawin and ., 2019).

Berdasarkan pengertian di atas dapat disimpulkan data mining adalah suatu pencarian di dalam suatu database besar untuk membantu mengambil keputusan berdasarkan pola yang diinginkan di waktu yang akan datang. Dengan menggunakan teknik tertentu dari suatu set data berukuran besar data mining dapat mengekstrak informasi atau pengetahuan penting. Untuk memperbaiki pengambilan keputusan dapat diperoleh dari informasi yang dihasilkan data mining tersebut.

2.5.1 Tahapan Proses Data Mining

Menurut Santoso dalam (Saiyar, 2018) data mining sering disebut juga *Knowledge Discovery in Database* (KDD), merupakan suatu pengumpulan 13 pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Proses tahapan data mining terdiri dari langkah utama yaitu:

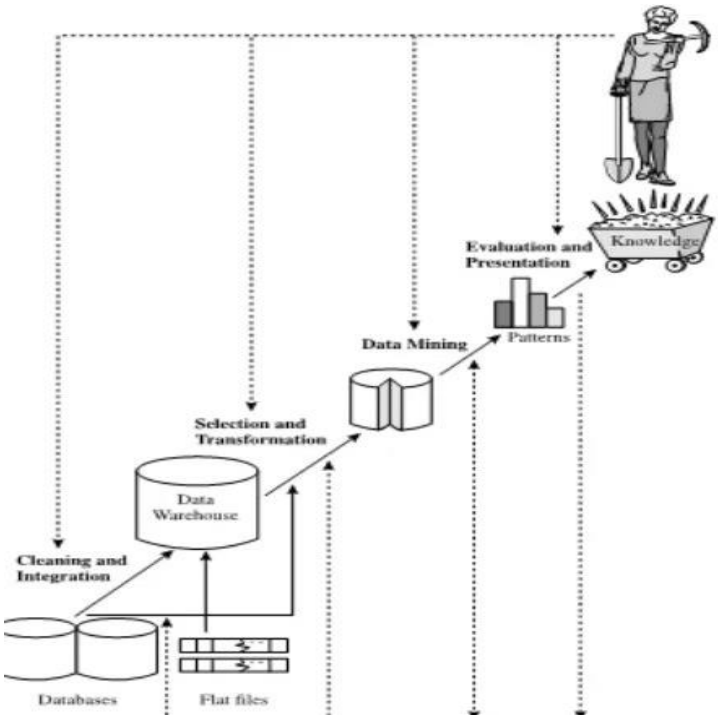
1. *Cleaning and Integration*

   Pada tahap *cleaning*, yaitu data yang tidak konsisten dan "*noise*" dihilangkan. Dan pada integration merupakan proses penggabungan sumber-sumber data.
2. *Selection and Transformation*

   Tahap selection adalah proses pengambilan data-data yang relevan untuk dianalisis. Sedangkan konsolidasi data ke bentuk yang sesuai untuk di "mining" dan menghasilkan ringkasan atau penggabungan merupakan proses transformasi.
3. Data Mining

   Pada tahap ini, proses awal dimana metode pengkajian diterapkan untuk mengekstraksi pola data.
4. *Evaluation and Presentation*

   Tahap evaluation adalah yang mewakili basis pengetahuan berdasarkan ukuran tertentu dengan mengidentifikasi pola yang menarik. Sedangkan, yang digunakan untuk menampilkan pengetahuan kepada pengguna yaitu pada tahap presentation.



Gambar 2. 5 Tahapan Data Mining

2.5.2 Pengelompokan Data Mining

Menurut Buulolo dalam (Ulkhairi, 2020) menerangkan bahwa data mining terbagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu:

1. Deskripsi: bertujuan mengidentifikasi/untuk menemukan pola yang sering muncul dan mengubah pola tersebut menjadi aturan yang dapat digunakan untuk mempermudah suatu aktivitas. Seperti pada sebuah Supermarket, pelanggan sering membeli produk A dan produk B secara bersamaan dan berulang, sehingga manajemen supermarket mengubah katalog barang dengan meletakkan produk A dan pada tempat yang berdekatan. Agar pelanggan tidak kesulitan dalam membeli produk tersebut.
2. Klasifikasi: berdasarkan hubungan antara variabel kriteria dengan variabel target yang dikelompokkan. Seperti pengelompokkan bencana alam yaitu rusak berat, rusak sedang, atau tidak berdampak.
3. Prediksi: prediksi mirip dengan klasifikasi. Memprediksi merupakan salah satu fungsi data mining yang biasa digunakan. Berdasarkan data-data sebelumnya nilai dari hasil prediksi akan digunakan di masa yang akan datang.

4. Estimasi: definisi dari estimasi adalah perkiraan/prediksi, sehingga hampir sama dengan klasifikasi, perbedaannya terletak pada bentuk pengelompokkan, dimana estimasi ke arah numerik dan bukan ke arah kategori.
5. Pengklasteran: pengklasteran adalah pengelompokkan data yang memiliki persamaan nilai (homogen). Bentuk data yang dapat dikelompokkan dalam pengklasteran yaitu hasil pengamatan, record data, atau kelas-kelas dan objek-objek yang memiliki kesamaan.
6. Asosiasi: asosiasi merupakan kumpulan, himpunan, persatuan, atau persekutuan. Pencarian attribute yang muncul/selalu muncul dalam waktu bersamaan merupakan proses pada data mining, misalnya seperti ketika dibeli produk A maka dibeli produk B, ketika dibeli produk B maka dibeli produk A, ketika dibeli produk A,B, maka dibeli produk C, dan seterusnya.

2.5.3 Langkah-langkah Data Mining

Menurut (Rahmawati and Merlina, 2018) Langkah-langkah data mining terdiri dari beberapa tahap sebagai berikut:

1. Tahap pertama: *Precise statement of the problem* (pernyataan tepat terhadap permasalahan) sebelum mengakses perangkat lunak data mining. Pertanyaan yang akan dijawab oleh seorang analis harus memiliki kejelasan tentang pertanyaan tersebut. Harus memiliki formulasi yang tepat untuk problematika yang ada dalam membuat solusinya.Menurut Thomas Timmreck, penyakit dapat diartikan sebagai sebuah keadaan dimana terdapat gangguan terhadap bentuk ataupun fungsi salah satu bagian tubuh yang menyebabkan tubuh menjadi tidak dapat bekerja dengan normal.
2. Menurut Elizabeth J. Crown, penyakit merupakan perihal hadirnya sekumpulan respons tubuh yang tidak normal terhadap agen, yang mana manusia memiliki toleransi yang sangat terbatas atau bahkan tidak memiliki toleransi sama sekali.
3. Tahap dua: Initial exploration. Kunci dari tahap ini yaitu diawali dengan mengidentifikasi dan menghilangkan data yang dikodekan salah (*cleaning*), transformasi data, memilih subset record, dataset, langkah awal seleksi , mendeskripsikan dan memvisualisasikan data.

4. Tahap tiga: Model building and validation. Tahap ini menentukan yang terbaik bagi performa prediktif dan melibatkan pertimbangan terhadap jenis pemodelan.
5. Tahap keempat: Deployment. Pada tahap ini untuk membuat (*generate*) prediksi yaitu dengan memilih aplikasi yang tepat beserta pemodelan.

2.5.4 Persiapan Data Mining

(Sulastri and Gufroni, 2017) mendefinisikan bahwa *Preprocessing* Data mining bisa meningkatkan kualitas data, sehingga data yang diperoleh langsung dari Toko Yati Kosmetik Sagulung, Batam diolah terlebih dahulu melalui tahap-tahap *data cleaning, data integration, data selection*, dan data transformation. Hal tersebut dilakukan agar data yang diolah lebih berkualitas, maksudnya data-data tersebut bersifat objektif, representatif, memiliki sampling error yang kecil, terbaharui dan relevan. Persiapan tersebut antara lain

1. *Data Cleaning*

Data cleaning adalah proses untuk mengatasi nilai yang hilang seperti noise dan data yang tidak konsisten.

2. *Data Integration*

Data integration merupakan proses menggabungkan data dari banyak database.

3. *Data Selection*

Data selection tetap merepresentasikan data aslinya dengan proses meminimalkan jumlah data yang digunakan untuk proses mining. Data selection dapat berupa *sampling, denoising, dan feature extraction*.

4. *Data Transformation*

Data transformation merupakan proses untuk mengubah bentuk dan format data. Hal tersebut dapat membantu memahami hasil yang didapat dan meringankan pengguna dalam proses mining.

2.5.5 Data Set

Set data (data set) merupakan data yang diolah pada data mining, dataset merupakan kumpulan data, dimana satu dataset merepresentasikan satu tabel pada database, atau bisa juga suatu matriks data dimana setiap kolom mewakili variabel tertentu, tiap baris merepresentasikan banyaknya data (Ulkhairi, 2020).

Beberapa faktor yang menjadi pertimbangan karakteristik set data seperti atribut, class, tipe data dan jumlah instan, merupakan beberapa faktor yang menjadi pertimbangan. Parameter atau faktor yang menyebabkan class/label/target terjadi yaitu atribut. Class adalah atribut yang dijadikan target, sering juga disebut dengan label. Tipe data untuk variabel pada statistik terbagi menjadi empat: nominal, ordinal, interval, ratio, tetapi secara praktis tipe data untuk atribut pada data mining hanya menggunakan dua: Nominal (Diskrit) dan Numerik (Kontinyu atau Ordinal) (Machmudi, 2018).

2.5.6 Teknik Data Mining

Menurut (Amalia, 2018) Beberapa teknik dan sifat data mining adalah sebagai berikut:

a. Klasterisasi. Adalah mempartisi dataset menjadi beberapa subnet atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set property yang di share bersama, dengan tingkat similaritas yang tinggi dalam suatu kelompok yang rendah. Disebut juga dengan "*unsupervised learning*".
b. Regresi. Adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
c. Klasifikasi. Adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut juga dengan "*supervised learning*".
d. Kaidah Asosiasi (*association rule*). Adalah mendeteksi kumpulan atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut.

2.6 Klasifikasi

Klasifikasi berasal dari bahasa latin yaitu classis yang artinya pengelompokan benda yang sama serta memisahkan benda yang tidak sama. Secara harfiah arti klasifikasi adalah penggolongan, pengelompokan. Dalam kaitannya di dunia perpustakaan klasifikasi diartikan sebagai kegiatan pengelompokan bahan pustaka berdasarkan ciri-ciri yang sama , misalnya pengarang, fisik, isi dan sebagainya.

Klasifikasi adalah salah satu pembelajaran yang paling umum di dalam data mining. Klasifikasi dapat didefinisikan sebagai bentuk dari analisis data yang digunakan untuk mengekstrak model yang akan digunakan untuk memprediksi label kelas. Kelas yang terdapat dalam klasifikasi merupakan atribut dalam satu set data yang paling unik yang merupakan variabel bebas yang terdapat dalam statistik. Klasifikasi data terdiri dari dua proses yaitu tahap pembelajaran dan tahap pengklasifikasian. Tahap Pembelajaran merupakan tahapan dalam pembentukan model klasifikasi, sedangkan tahap pengklasifikasian merupakan tahapan dalam penggunaan model klasifikasi yang digunakan untuk memprediksi label kelas pada data (Sartika *et al.*, 2017).

Algoritma yang terdapat dalam klasifikasi *berupa Decision Tree, Nearest Neighbor, Naïve Bayesian Classification, Neural Network dan Support Vector Machines*. Algoritma diatas merupakan beberapa cara untuk melakukan proses klasifikasi dimana memiliki fungsi sebagai pembelajaran yang mengklasifikasi sebuah unsur data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Klasifikasi melihat dari pola-pola historis suatu data untuk dipelajari dan bertujuan untuk menempatkan objek-objek baru kedalam kelompok kelasnya tersendir.

2.7 Algoritma

Menurut (Maulana, 2017) Algoritma adalah metode efektif yang diekspresikan sebagai rangkaian terbatas. Algoritma juga merupakan kumpulan perintahuntuk menyelesaikan suatu masalah. Perintah-perintah ini dapat diterjemahkan secara bertahapdari awal hingga akhir. Masalah tersebut dapat berupa apa saja, dengan syarat untuk setiap permasalahan memiliki kriteria kondisi awal yangharus dipenuhi sebelum menjalankan sebuah algoritma. Algoritma juga memiliki pengulangan proses (iterasi), dan juga memiliki keputusan hingga keputusan selesai.

Dalam cabang disiplin ini, algoritma dipelajari secara abstrak, terlepas dari system komputer atau bahasa pemrograman yang dipergunakan. Algoritma yang berbeda dapat diterapkan untuksuatu permasalahan dengan kriteria yang sama. Kompleksitas dari suatu algoritma merupakan ukuran seberapa banyak komputasi yangditerapkan pada algoritma tersebut untuk menyelesaikan permasalahannya.

Secarainformal, algoritma yang dapat menyelesaikan permasalahan dalam waktu yang relative singkat memiliki tingkat kompleksitas yang rendah, sementara untuk algoritma yang menyelesaikan permasalahan dalam waktu yang lebih lama memiliki tingkat kompleksitas yang lebih tinggi pula.

Pertimbangan dalam pemilihan algoritma adalah, pertama, algoritma haruslah benar. Artinya algoritma akan memberikan luaran yang dikehendaki dari sejumlah masukan yang diberikan. Tidak peduli sebagus apapun algoritma, kalau memberikan luaran yang salah, pastilah algoritma tersebut bukanlah algoritma yang baik. Kedua, algoritma yang baik harus mampu memberikan hasil yang sedekat mungkin dengan nilai yang sebenarnya. kita harus mengetahui seberapa baik hasil yang dicapai oleh algoritma tersebut. Hal ini penting terutama pada algoritma untuk menyelesaikan masalah yang memerlukan aproksimasi hasil (hasil yang hanya berupa pendekatan). Ketiga, efisiensi algoritma, semisal algoritma itu benar (mendekati kebenaran), tetapi memakan waktu yang lama dalam mendapatkan kebenaran algoritma, untuk apa algoritma tersebut dipakai? Karena inti dari algoritma yang baik adalah mendapatkan jawaban kebenaran (mendekati kebenaran) dengan cepat.

2.8 Algoritma *K-Nearest Neighbor* (K-NN)

Algoritma *K-Nearest Neighbor* (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan obyek. Prinsip kerja dari K-Nearest Neighbor (K-NN) adalah untuk mencari jarak terdekat antara data yang akan dievaluasi dengan k tetangga (*neighbor*) terdekatnya dalam data pelatihan. Ketepatan algoritma K-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi (Fitrianingsih, Bettiza and Uperiati, 2021).

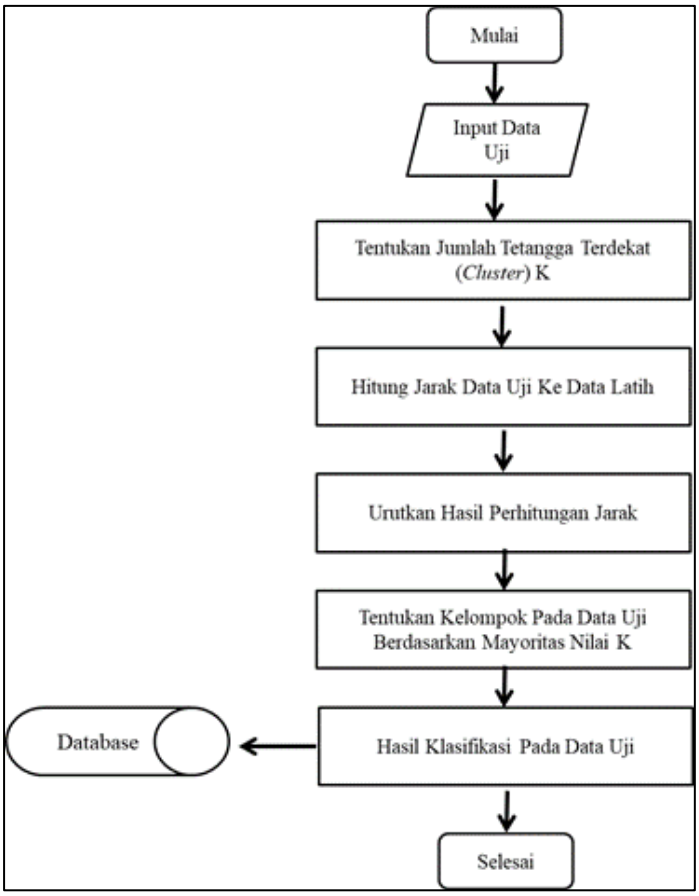
Secara umum data mining memiliki teknik-teknik yang digunakan dalam melakukan pengklasifikasian. Pendekatan dari teknik-teknik itu terdiri dari 2 teknik, yaitu teknik *Supervised learning* dan *unsupervised learning*. Supervised learning adalah sebuah pendekatan dimana Algoritma *K-Nearest Neighbor* merupakan suatu metode yang menggunakan algoritma supervised. Perbedaan

antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada *unsupervised learning* data belum memiliki pola apapun, dan tujuan unsupervised learning untuk menemukan pola dalam sebuah data. Dalam penelitian penerimaan peserta didik baru ini menggunakan algoritma *supervised learning*. Metode *K-Nearest Neighbor* bekerja berdasarkan asumsi bahwa suatu data akan memiliki kelas atau kategori yang sama dengan data yang berada disekitarnya. Konsep ini dikenal dengan konsep ketetanggaan (Kurniawan and Saputra, 2019).

K-Nearest Neighbor merupakan sebuah metode untuk melakukan klasifikasi terhadap objek baru berdasarkan (K) tetangga terdekat. Untuk mendukung pengambilan keputusan tersebut, akan terlihat mayoritas dari keputusan teman atau tetangga. Teman atau tetangga tersebut dapat dipilih berdasarkan dengan kedekatannya. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan *Euclidean Distance*, atau dapat juga menggunakan rumus jarak yang lain. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori *K-Nearest Neighbor*. Dengan demikian *K-Nearest Neighbor* dari sebuah *instance* x didefinisikan sebagai K *instance* yang memiliki jarak terkecil (kedekatan terbesar, nearest) dengan x. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan *Euclidean Distance* yang direpresentasikan pada persamaan sebagai berikut(Kurniawan and Saputra, 2019):

$$D\ (a,b) = \sqrt{\sum_k^d (a_k - b_k)^2}$$

dimana matriks D (a,b) adalah jarak skalar dari kedua vektor a dan b dari matriks. Dalam menentukan nilai K sebaiknya dilihat dari jumlah klasifikasi bila jumlahnya genap maka sebaiknya menggunakan nilai K yang ganjil, dan sebaliknya jika jumlah klasifikasi jumlahnya ganjil maka sebaiknya dalam menggunakan nilai K yang genap, karena jika tidak begitu maka sistem kemungkinan tidak akan mendapatkan jawaban. Berikut ini adalah proses dari metode *K-Nearest Neighbor* yang ditunjukkan pada Gambar di bawah ini(Kurniawan and Saputra, 2019):



Gambar 2. 6 Proses Metode *K-Nearest Neighbor*

Langkah-langkah untuk menghitung metode *K-Nearest Neighbor* antara lain (Sutoyo, Sembilanbelas and Kolaka, 2018):

1. Tentukan parameter k = jumlah tetangga terdekat

   Hitung jarak antara data yang akan dievaluasi dengan semua data pelatihan. Secara umum untuk mendefinisikan jarak antara dua objek x dan y, digunakan rumus jarak Euclidean pada persamaan berikut.

$$d\ (x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi\ )^{\ 2}}$$

Dimana :

- *d (x,y) = dissimilarity*/jarak
- *n* = dimensi data
- *i* = variabel data

- *Xi* = sampel data
- *Yi* = data uji atau testing

2. Mengurutkan jarak yang terbentuk
3. Menentukan jarak terdekat sampai urutan k
4. Memasangkan kelas yang bersesuaian
5. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi.

Algoritma metode K-NN sangatlah seder-hana, bekerja berdasarkan jarak terpendek dari query instance ke training sample untuk menentukan K-NN-nya.Training sample diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresen-tasikan fitur dari data.Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi training sample. Sebuah titik pada ruang ini ditandai kelas c jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat dari titik tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan Euclidean Distance (Kartika, Santoso and Sutrisno, 2017).

2.8.1 Konsep *K-Nearest Neighbour*

Algoritma k-NN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (Sutoyo, Sembilanbelas and Kolaka, 2018). Metode K-NN murni termasuk dalam klasifikasi pelajar malas karena menunda proses pelatihan (atau bahkan tidak ada pelatihan sama sekali) hingga ada data pengujian yang ingin Anda ketahui label kategorinya, kemudian metode baru akan dijalankan. Algoritma K-NN mengklasifikasikan berdasarkan kesamaan satu jenis data dengan jenis data lainnya (Kurniawan and Saputra, 2019).

2.8.2 Fungsi *K-Nearest Neighbour*

Secara umum, Data Mining memiliki keterampilan untuk klasifikasi. Metode teknis ini mencakup dua teknik: pembelajaran yang diawasi daripada pembelajaran yang diawasi; pembelajaran yang diawasi bertujuan untuk menemukan pola-pola baru dalam data dengan menghubungkan pola data yang ada dan data baru; pada saat yang sama, dalam pembelajaran tanpa pengawasan, data tidak memiliki pola

dan gol. Belajar tanpa pengawasan untuk menemukan pola dalam data. Dalam studi pendaftaran ini, algoritma pembelajaran yang diawasi digunakan. Metode *K-Nearest Neighbour* didasarkan pada asumsi berikut: data memiliki kategori yang sama dengan data sekitarnya. Konsep ini disebut dengan konsep K-NN.

2.8.3 Kelebihan *K-Nearest Neighbour*

Setiap algoritma pasti mempunyai kelebihan dan kekurangannya sendiri begitu pula dengan algoritma *K-Neraest Neighbour* berikut adalah kekurangan dari algoritma *K-Neraest Neighbour*(Trivusi, 2022) :

a. Mudah diterapkan

Mengingat kesederhanaan dan akurasi algoritma, K-NN merupakan salah satu pengklasifikasi pertama yang sebaiknya dipelajari oleh data scientist pemula.

b. Mudah beradaptasi

Saat sampel training baru ditambahkan, algoritma K-NN menyesuaikan untuk ikut memperhitungkan data baru karena semua data pelatihan disimpan ke dalam memori.

c. Memiliki sedikit *hyperparameter*

K-NN hanya membutuhkan nilai k dan metrik jarak, yang relatif lebih sedikit jika dibandingkan dengan algoritma machine learning lainnya.

2.8.4 Kekurangan *K-Nearest Neighbour*

Berikut adalah kekurangan dari algoritma *K-Neraest Neighbour*(Trivusi, 2022) :

1. Tidak berfungsi dengan baik pada dataset berukuran besar

Untuk dataset berukuran besar, *cost* untuk menghitung jarak antara titik baru dan setiap titik yang ada sangat besar dan cenderung menurunkan kinerja algoritma.

2. Kurang cocok untuk dimensi tinggi

Algoritma K-NN tidak bekerja dengan baik pada data berdimensi tinggi karena dengan jumlah dimensi yang besar, menjadi sulit bagi algoritma untuk menghitung jarak di setiap dimensi.

3. Perlu penskalaan fitur

Kita perlu melakukan penskalaan fitur (standarisasi dan normalisasi) sebelum menerapkan algoritma K-NN ke kumpulan data apa pun. Jika kita tidak melakukannya, K-NN dapat menghasilkan prediksi yang salah.

4. Sensitif terhadap noise data, missing values dan outliers

K-NN sensitif terhadap noise dalam dataset. Kita perlu secara manual memasukkan nilai yang hilang dan menghapus outlier.

2.9 *Distance Metric*

Distance metric atau metrik jarak adalah metode yang digunakan untuk mengukur kesamaan dan kedekatan antara dua titik data. Saat menilai seberapa mirip dua titik data, kita perlu acuan untuk dapat membandingkannya. *Distance metric* memungkinkan kita untuk menghitung secara numerik seberapa mirip dua titik dengan menghitung jarak di antara keduanya. Adapun jenis-jenis *distance metric* sebagai berikut(M. Nishom, 2019) :

1. Euclidean Distance

Euclidean Distance Euclidean distance merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam Euclidean space (meliputi bidang *euclidean* dua dimensi, tiga dimensi, atau bahkan lebih). Untuk mengukur tingkat kemiripan data dengan rumus *euclidean* distance digunakan rumus berikut :

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Dimana :

d = jarak antara x dan y

x = data pusat klaster

y = data pada atribut

i = setiap data

n = jumlah data

xi = data pada pusat klaster ke i

yi = data pada setiap data ke i

2. Manhattan Distance

Manhattan distance digunakan untuk menghitung perbedaan absolut (mutlak) antara koordinat sepasang objek. Rumus yang digunakan adalah sebagai berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} | x_i - y_i |}$$

d = jarak antara x dan y

x = data pusat klaster

y = data pada atribut

i = setiap data

n = jumlah data

xi = data pada pusat klaster ke i

yi = data pada setiap data ke i

3. Minkowski Distance

Minkowski distance merupakan sebuah metrik dalam ruang vektor di mana suatu norma didefinisikan (*normed vector space*) sekaligus dianggap sebagai generalisasi dari Euclidean distance dan Manhattan distance. Dalam pengukuran jarak objek menggunakan *minkowski distance* biasanya digunakan nilai p adalah 1 atau 2. Rumus yang digunakan adalah sebagai berikut:

$$(x,y) = \sqrt{(\sum_{i=1}^{n} | x_i - y_i |^p)(\frac{1}{p})}$$

d = jarak antara x dan y

x = data pusat klaster

y = data pada atribut

i = setiap data

n = jumlah data,

xi = data pada pusat klaster ke i

yi = data pada setiap data ke i

p = power

2.10 Algoritma *Decision Tree* C4.5

Algoritma C4.5 merupakan salah satu teknik klasifikasi pada *machine learning* yang digunakan pada proses data mining dengan membentuk sebuah pohon keputusan (*decision tree*) yang direpresentasikan dalam bentuk aturan. Algoritma

C4.5 merupakan kelompok algoritma dengan menggunakan pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Semakin kaya informasi atau pengetahuan yang dikandung oleh data training, maka akurasi akan semakin meningkat. Pohon dalam analisis pemecahan masalah pengambilan keputusan adalah pemetaan mengenai alternatif-alternatif pemecahan masalah yang dapat diambil dari masalah tersebut. Pohon tersebut juga memperlihatkan faktor-faktor kemungkinan/probabilitas yang akan mempengaruhi alternatif-alternatif keputusan tersebut, disertai dengan estimasi hasil akhir yang akan didapat bila kita mengambil alternatif b keputusan tersebut.Pengambilan keputusan merupakan masalah penting bagi organisasi untuk menemukan alternatif terbaik dari alternatif yang ada (Mada, 2018).

C4.5 adalah algoritma yang dibuat oleh Ross Quinlan dan digunakan untuk membuat pohon keputusan. Algoritma ini sering dikategorikan sebagai pengklasifikasi statistik. C4.5 merupakan pengembangan dari algoritma ID3 yang menggunakan entropi informasi, atribut kontinyu dan diskret, atribut kategorial dan numerik, dan missing values. Algoritma ini membutuhkan set data latih karena termasuk algoritma pembelajaran yang terawasi (*supervised learning algorithm*). Set data latih berupa sampel yang sudah terklasifikasi. C4.5 menganalisa set data latih dan membangun pengklasifikasian yang harus secara tepat mampu mengklasifikasikan data latih maupun data uji. Formulasi Matematis dari algoritma C 4.5 dapat dilihat pada formula dibawah ini

$$Entrophy(s) = \sum_{i=1}^{n} pi * log_2\ pi$$

Dengan penjelas $\log_2$:

$$log_2(x) = \frac{In(x)}{In(2)}$$

Dengan :

S = ruang (data) sampel yang digunakan untuk training

A = atribut

V = suatu nilai yang mungkin untuk atribut A

Nilai(A) = himpunan yang mungkin untuk atribut A

|Si| = jumlah sampel untuk nlai i

|S| = jumlah seluruh sampel data

Entropy(Si) = entropiuntuk sampel-sampelyang memiliki nilai i

Menghitung Nilai Gain :

$$Gain(A) = entropy(s) - \sum_{i=}^{k} \frac{|S_i|}{|S|} x\ entropy(S_i)$$

Dengan :

S = ruang (data) sampel yang digunakan untuk training

A = atribut

|Si| = jumlah sampel untuk nlai V

|S| = jumlah seluruh sampel data

Entropy(Si) = entropiuntuk sampel-sampelyang memiliki nilai i

$$SplitInfo(S, A) = -\sum_{i=1} \frac{S^i}{S} log_2 \frac{S_i}{S}$$

Dengan :

S = ruang (data) sampel yang digunakan untuk training

A = atribut

S_i = jumlah sampel untuk atribut i

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

Dengan :

S = ruang (data) sampel yang digunakan untuk training

A = atribut

Gain(S , A) = *nformation* gain pada atribut A

SplitInfo(S, A) = split *nformation* pada atribut A

Menurut (Ryanwar, 2020) *Entropy* adalah ukuran dari teori informasi yang dapat mengetahui karakteristik dari impurty dan homogenity dari kumpulan data. Dari nilai Entropy tersebut kemudian dihitung nilai information gain masing-masing atribut. *Information Gain* adalah informasi yang didapatkan dari perubahan entropy pada suatu kumpulan data, baik melalui observasi atau

bisa juga disimpulkan dengan cara melakukan partisipasi terhadap suatu set data.

Secara umum langkah dalam membuat pohon keputusan menggunakan algoritma C4.5 adalah sebagai berikut:

1. Menghitung entropitotal dari datasetdilanjutkan denganentropi masing-masing atribut.
2. Setelah diperoleh entropi masing-masing atribut, menghitung information gain masing-masing.
3. Memilih atribut yang memiliki information gainpaling besar sebagai akar.
4. Mengulangi perhitungan entropi dan gain untuk menentukan atribut berikutnya sebagai daun.

2.10.1 Keuntungan Algoritma *Decision Tree* C4.5

Berikut beberapa keuntungan dari penggunaan algoritma Decision Tree C4.5(Joko Minardi, 2016):

1. Mudah untuk dipahami dan ditafsirkan.
2. Memiliki nilai walau hanya dengan data yang sedikit.
3. Dapat dipadukan dengan teknik pengambilan keputusan lainnya.
4. Membentangkan semua masalah sehingga semua kemungkinan dapat diklasifikasikan.
5. Memungkinkan untuk menganalisa dalam mengambil keputusan mengenai kemungkinan dari alternatif.
6. Menyediakan suatu kerangka kerja untuk mengukur hasil dari nilai dan kemungkinan untuk mencapai keputusan.
7. Membantu untuk membuat keputusan yang terbaik berdasarkan informasi yang ada.

2.10.2 Keuntungan Algoritma *Decision Tree* C4.5

Adapun kekurangan yang ada menurut algoritma Decision Tree C4.5 yaitu:

1. Algoritma C4.5 dapat membangun cabang kosong yang nilainya tidak berkontribusi untuk menghasilkan aturan.
2. Terjadi overfitting akibat dari noise data.
3. Kebisingan (noise data) yang rentan terjadi pada C4.5

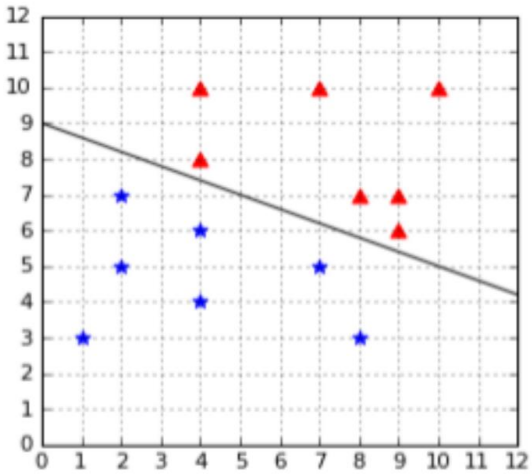
2.11 Algoritma *Support Vector Machine* (SVM)

Support Vector Machine (SVM) dikenalkan pertama kali oleh Vapnik tahun 1992 sebagai salah satu metode *learning machine* yang bekerja dengan prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada input *space*. Metode ini menggunakan hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur yang berdimensi tinggi, dengan mengimplementasikan learning bias yang berasl dari teori pembelajaran statistik. Tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan dengan SVM sangat bergantung terhadap fungsi kernel dan parameter yang digunakan (Parapat and Furqon, 2018).

Berdasarkan dengan karakteristiknya metode SVM dibagi menjadi dua yaitu linear dan non linear, SVM linear merupakan data yang dipisahkan secara linear yaitu memsiahkan dua kelas pada *hyperplane* dengan *soft margin*. Sedangkan non linear yaitu merupakan fungsi dari kernel trick terhadap ruang yang berdimensi tinggi. SVM sangat cepat dan efektif untuk masalah klasifikasi teks, dalam istilah geometris sebuah klasifikasi biner dapat dilihat sebagai *hyperplane* dalam ruang fitur yang memisahkan tiitk-titik yang mewakili contoh positif dari kategori yang mewakili keadaan negative (Alita, Fernando and Sulistiani, 2020).

SVM digunakan untuk mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas. Hyperplane adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai line whereas, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut plane similarly, sedangan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi di sebut hyperplane. Pada perkembangannya, SVM dapat diperluas untuk klasifikasi multi kelas. SVM multi kelas diperlukan pendekatan yang berbeda dengan kasus dua kelas. Ada beberapa metode SVM Multi Kelas yaitu salah satunya metode SVM Muti Kelas *One-AgainstOne*. Pada metode *One-Again-One*, dengan cara membangun sejumlah model SVM biner yang nantinya akan dibandingkan satu kelas dengan kelas lainnya. Untuk mengklasifikasikan data ke k-kelas, maka harus membangun sejumlah k(k-1)/2 model SVM biner (Hendrastuty *et al.*, 2021).

Support Vector Machine menggunakan 2 titik (vector) yang selanjutnya dua titik tersebut akan membentuk garis pembatas (sisi pembatas jika 3 dimensi atau lebih) garis pembatas yang dibentuk dari dau buah vector ini disebut hyperplane.



Gambar 2. 7 *Hyperplane* Dua Titik

Dua titik yang menjadi patokan hyperplane disebut dengan *support vector*. Dapat dilihat bahwa memiliki dua kelompok data yang disebut klasifikasi, kemudian tugas SVM adalah membagi dua kelompok ini sebaik mungkin atau menentukan hyperplane terbaik, pembagian dimana garis batasnya dapat memisahkan dua kelompok dengan jarak terjauh antara titik terluar di masing masing kelompok dengan garis pembatas itu sendiri. Permasalahan non linear dapat diatasi dengan memodifikasi trick kernel ke dalam SVM yang akan menjadi pemisah kelas atau *hyperplane* menjadi dua kelas didalam ruang vector dalam penelitian ini kernel yang akan digunakan adalah kernel linear. Seperti yang dapat dilihat persamaannya pada Tabel 2.2 di bawah ini.

Tabel 2. 4 Rumus Kernel

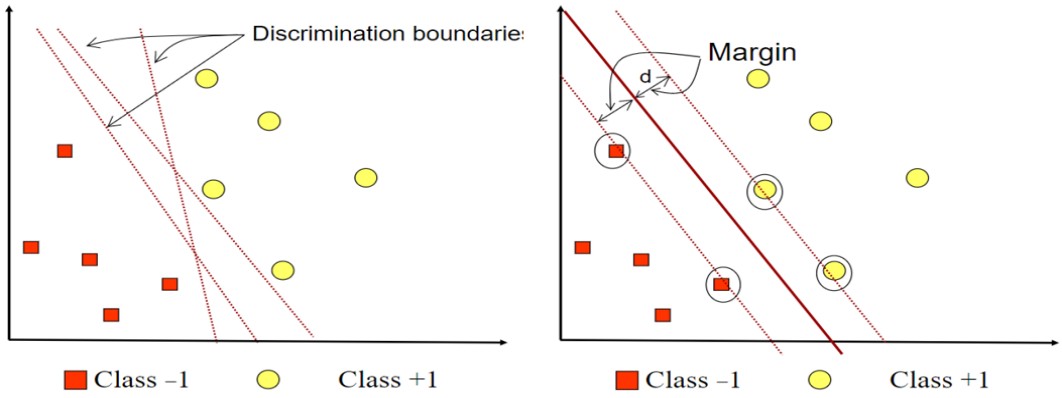
Jenis Kernel	Model
Linear	$K(x.x') = x.x'$
Polynomial	$K(x.x') = x.x' + c)'$
RBF *Gaussian*	$K(x.x') = \exp(-y\|\|x - x'\|\|^2)$
Siqmoid	$K(x.x') = \tanh(\alpha x.x' + \beta)$

Penggunaan kernel dapat dibedakan sesuai dengan data yang digunakan. Kernel linier digunakan pada saat data yang akan diklasifikasikan dapat dipisahkan oleh

hyperplane berbentuk garis. Dalam artian lain, kernel linier digunakan pada data berdimensi dua. Sedangkan kernel non-linier digunakan pada data yang dipisahkan oleh hyperplane berbentuk bidang pada ruang berdimensi tinggi (Puspitasari dkk., 2018).

2.11.1 Linier *Support Vector Machine*

Linier *Support Vector Machine* dapat diterapkan pada data yang dapat dipisahkan secara linier. Konsepnya ialah mencari *hyperplane* atau garis pemisah antara dua kelas yang paling baik. Misalkan $x_i = \{x_1, \dots, x_n\}, x_i \in R^n$ merupakan data set. Positive class dinotasikan dengan 1, dan negative class dinotasikan dengan -1. Maka label kelas dinotasikan sebagai $y_i \in \{+1, -1\}$, dimana $i = 1,2, \dots\dots, l \; dengan \; l$ menunjukan banyak data.



Gambar 2. 8 *Linear Support Vector Machine*

Pada Gambar 2.8 memperlihatkan beberapa pattern yang menggambarkan anggota dari dua kelas dimana data set dapat dipisahkan sesuai dengan kelasnya dengan menggunakan beberapa *hyperplane* (*discrimination boundaries*). Selanjutnya, Gambar 2.6 juga menunjukkan sepasang *hyperplane* sejajar yang memisahkan dua kelas, sedangkan pattern yang berada pada hyperplane merupakan support vector. Nilai margin terbesar menunjukkan hyperplane terbaik. Margin adalah jarak terdekat antara *hyperplane* dengan pattern masing-masing kelas. Persamaan umum hyperplane yang memisahkan dua kelas dapat didefinisikan sebagai berikut

$$w_i x_i + b = 0 \qquad (2,1)$$

Berikut merupakan pertidaksamaan dari 2 hyperplane dimana *hyperplane* pertama membatasi kelas pertama dan *hyperplane* kedua membatasi kelas kedua, yaitu:

$$w_i x_i + b \geq +1, y_i = +1$$

$$w_i x_i + b \leq -1, y_i = -1$$

dimana *w* merupakan vektor pembobot dan *b* merupakan bias. Pemaksimalan jarak terdekat antara *hyperplane* dengan pattern dilakukan untuk menghitung margin yang mana dirumuskan dengan $\frac{1}{||w||}$. Hal ini dapat diformulasikan dalam *Quadratic Programming* (QP) problem, yaitu dengan meminimalkan persamaan

$$\frac{min}{w}\tau(w) = \frac{1}{2}||w||^2$$

Dengan syarat

$$y_i(W_i x_i + b - 1 \geq 0 \forall i$$

Optimasi dapat dilakukan dengan menggunakan Lagrange Multiplier seperti berikut:

$$L = \frac{1}{2}||W||^2 - \sum_{i=1}^{l} \alpha_i [y_i(w_i x_i + b) - 1]$$

$$L = \frac{1}{2}||W||^2 - \sum_{i=1}^{l} \alpha_i y_i [y_i(w_i x_i + b) - \sum_{i=1}^{l} \alpha_i]$$

αi merupakan Lagrange Multiplier dengan nilai nol atau positif ($\alpha i \geq 0$). Optimasi dilakukan dengan meminimalkan *L* terhadap w dan b sebagai berikut

$$\frac{\partial L}{\partial b} = 0$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} = 0$$

$$w_i - \sum_{i=1}^{l} \alpha_i y_i x_i = 0$$

$$w_i = \sum_{i=1}^{l} \alpha_i y_i x_i$$

$$\begin{aligned} L &= 1\,2\,||w||\,2 - \sum \alpha i\, yi(\, l\, i = 1\, wixi + b) - \sum \alpha i\, l\, i = 1\, L \\ &= 1\,2\,(wi\,.\,wi\,) - (\sum \alpha i\, yi\, l\, i = 1\, wixi + \sum \alpha i\, yi\, l\, i = 1\, b - \sum \alpha i\, l\, i \\ &= 1\,)\, L = 1\,2\,(\sum \alpha i\, yi\, l\, i = 1\, xi\,.\,\sum \alpha j\, yj\, l\, j = 1\, xj) - (\sum \alpha i\, yi\, l\, i \\ &= 1\, xi\,.\,\sum \alpha j\, yj\, l\, j = 1\, xj + 0 - \sum \alpha i\, l\, i = 1\,)\, L \\ &= 1\,2 \sum \alpha i \alpha j yiyjxixj\, l\, i, j = 1 - (\sum \alpha i \alpha j yiyjxixj\, l\, i, j = 1 - \sum \alpha i\, l\, i \\ &= 1\,)\, L = \sum \alpha i - l\, i = 1\, 1\,2 \sum \alpha i \alpha j yiyjxixj\, l\, i, j = 1\,(\end{aligned}$$

$$\text{dimana } \alpha i \geq 0, \sum \alpha i\, l\, i = 1\, yi = 0.$$

Pada umumnya, dua kelas tidak dapat dipisahkan secara sempurna oleh *hyperplane* sehingga syarat dalam persamaan (2.5) tidak dapat terpenuhi. Hal itu menyebabkan tidak dapat dijalankannya optimasi. Teknik softmargin dapat digunakan untuk menangani permasalahan tersebut. Softmargin memodifikasi persamaan (2.5) dengan menambahkan slack variabel ξi dengan $\xi i > 0$ sebagai berikut

$$y_i\,(w_i x_i + b) \geq 1 - \xi i\,, \forall i$$

Sehingga didapatkan

$$\min 1\,2\,||w||\,2 + C \sum_{i=1}^{l} \xi i$$

dengan syarat

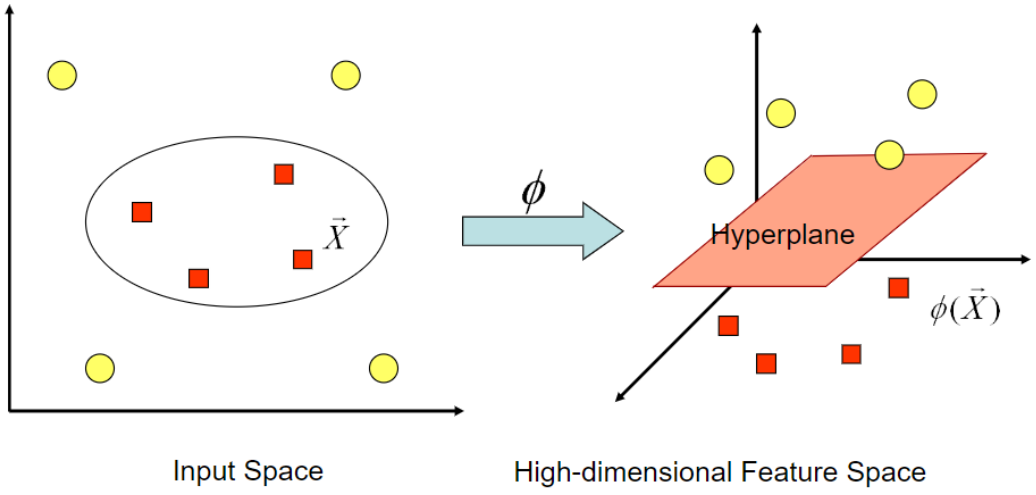
$$y_i\,(w_i x_i + b) - 1 + \xi i \geq 0, \forall i$$

Meminimalkan persamaan (2.10) akan memaksimalkan jarak margin antar kelas. Adanya penambahan slack variabel ξi akan meminimalisir *misclassification error* (jumlah kesalahan pada klasifikasi). Parameter C digunakan untuk mengontrol tradeoff antara margin dengan error klasifikasi ξ. Nilai C yang besar menunjukkan penalti yang besar terhadap error klasifikasi tersebut.

2.11.2 Non-Linier *Support Vector Machine*

Non-Linier *Support Vector Machine* digunakan pada permasalahan data yang tidak dapat dipisahkan secara linier. Metode SVM dapat digunakan pada kasus non-linier melalui pendekatan kernel. Agar data yang digunakan dapat dipisahkan secara linier maka dapat diatasi dengan menggunakan kernel. Konsep kerja kernel adalah dengan mentransformasi data ke dalam dimensi ruang fitur (feature space). Penyelesaian kasus non-linier dapat diatasi dengan SVM yang telah dikembangkan yaitu dengan menggunakan kernel trick yang dapat mengubah data menjadi linier (Hamel, 2009). Adapun kernel trick dirumuskan dengan:

$$K(Xi, Xj) = \Phi(Xi).\Phi(Xj)$$



Gambar 2. 9 Kernel Trick SVM

Pada Gambar 2.9 mengilustrasikan data berdimensi dua yang tidak dapat dipisahkan secara linier oleh hyperplane. Selanjutnya, pada Gambar 2.9 juga mengilustrasikan pemetaan data ke dalam ruang yang lebih tinggi dimensinya (dimensi tiga) sehingga dua kelas dapat dipisahkan secara linier oleh hyperplane. Berikut merupakan notasi matematika dari mapping tersebut:

$$\Phi : Rd \rightarrow Rq, d < q$$

Umumnya, transformasi Φ tidak diketahui sehingga dapat diganti dengan fungsi kernel $K(Xi , Xj)$ yang mendefinisikan transformasi Φ secara implisit. Hasil klasifikasi dapat diperoleh dari persamaan:

$$\alpha_i = \frac{N}{\sum_{i=1}^{N}(K(X_i, X_j)y_i y_j}$$

$$b = -\frac{1}{2}(wx^{+} + wx^{-})$$

$$f(\Phi(X)) = \text{sign}\,(\,wi\,.\,\Phi(X) + b)$$

$$f(\Phi(X)) = \text{sign}\,(\sum aiyi\ n\ i = 1\ \Phi(Xi\,).\,\Phi(Xj) + b)$$

$$f(\Phi(X)) = \text{sign}\,(\sum aiyi\ n\ i = 1\ K(Xi\,, Xj) + b)$$

dimana $\alpha i \geq 0$. Adapun untuk nilai αi dan b dapat diperoleh dengan persamaan sebagai berikut (Saputra & Ary, 2020):

2.12 Aplikasi Rapidminer

RapidMiner adalah aplikasi atau perangkat lunak yang berfungsi sebagai alat pembelajaran dalam ilmu data mining. Platfrom dikembangkan oleh perusahaan yang didedikasikan untuk semua langkah yang melibatkan sejumlah besar data dalam bisnis komersial, penelitian, pendidikan, pelatihan, dan pembelajaran. RapidMiner memiliki sekitar 100 solusi pembelajaran untuk pengelompokan, klasifikasi dan analisis regresi. RapidMiner juga mendukung sekitar 22 format file, seperti .xls, .csv, dan sebagainya(Riandaru *et al.,* 2021).

Rapidminer merupakan salah satu tools yang dipakai dalam data mining. Rapidminer adalah platform perangkat lunak ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama. Rapidminer menyediakan lingkungan terpadu untu pembelajaran *machine learning, deep learning, text mining, dan predictive analytics.* Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan prototype, dan pengembangan aplikasi dengan mendukung semua langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan.

Rapid Miner juga menjadi sofware yang dapat berdiri dengan sendirinya agar analisis data pada mesin Data Mining dapat diintegritskan pada produknya. Rapid Miner telah menyediakan GUI (*Grapic User Interface*) sbagai perancang sebuah pipeline analisis. GUI (*Graphic User Interface*) sebagai perancang pipeline Analisis. GUI menghasilkan file XML (Extensible Markup Language) yang di

definisikan sebagai proses analisis sebagai keinginan pengguna di terapkan ke suatu data, lalu file tersebut dibaca oleh Rapid Miner agar dapat menjalankan analisis dengan cara otomatis.

RapidMiner sebelumnya bernama YALE (*Yet Another Learning Environment*), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit dari University of Dortmund*. RapidMiner didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan menggunakan RapidMiner di lebih dari 40 negara. RapidMiner sebagai software open source untuk data mining tidak perlu diragukan lagi karena software ini sudah terkemuka di dunia. RapidMiner menempati 23 peringkat pertama sebagai Software data mining pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011.

RapidMiner memiliki beberapa sifat sebagai berikut:

- Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.
- Proses penemuan pengetahuan dimodelkan sebagai operator trees
- Representasi XML internal untuk memastikan format standar pertukaran data.
- Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
- Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
- Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.

Beberapa Fitur dari RapidMiner, antara lain:

- Banyaknya algoritma data mining, seperti decision treee dan selforganization map.
- Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.

2.12 *Confusion Matrix*

a. Pengujian *Confusion Matrix* Dua Kelas

Merupakan tabel yang menggambarkan performa dari sebuah model atau algoritma secara spesifik. Setiap baris dari matrix tersebut, merepresentasikan kelas aktual dari data, dan setiap kolom merepresentasikan kelas prediksi dari data (atau sebaliknya)(Saputro and Sari, 2020). *Confusion matrix* memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah. Matrix tersebut dijelaskan pada Tabel 2.5

Tabel 2. 5 Tabel *Confusion Matrix* Dua Kelas

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Negative	False Negative (FN)	True Positive (TP)

1. *True Positive* = Berarti seberapa banyak data yang aktual kelasnya positif, dan model juga memprediksi positif.
2. *True Negative* = Berarti seberapa banyak data yang aktual kelasnya negatif, dan model memprediksi negatif.
3. *False Positive* = Berarti seberapa banyak data yang aktual kelasnya negatif, namun model memprediksi positif.
4. *False Negative* = Berarti seberapa banyak data yang aktual kelasnya positif, namun model memprediksi negatif.

   Melalui 4 data tersebut, dapat diperoleh data lain yang sangat berguna untuk mengukur perfoma sebuah model, diantaranya:

   1. *Accuracy* merupakan perhitungan terhadap proporsi dari jumlah total prediksi yang benar

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

   2. *Precision* merupakan perhitungan terhadap perkiraan proporsi kasus positif yang benar dandirumuskan dalam persamaan dibawah ini

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

3. *Recall* merupakan perhitungan terhadap perkiraan proporsi kasus positif yang diidentifikasi benar

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

4. Merupakan rata-rata harmonik dari Precision dan Recall

$$F1 - Score = 2 * \frac{Precission * Recall}{Precision + Reacall}$$

b. Pengujian Confusion Matrix *Multiclass*

Pengujian kualitas dilakukan untuk mengetahui kinerja dari algoritma klasifikasi yang telah diterapkan. Ada beberapa cara untuk mengukur kinerja algoritma klasifikasi tiga diantaranya adalah precision, recall dan f-measure (Fauziah *et al.,* 2018). Untuk mengukur kinerja algoritma dapat menggunakan Tabel confusion matrix multiclass yang dapat dilihat pada Tabel 2.

Tabel 2. 6 *Confusion Matrix Multiclass*

		Aktual		
Prediksi		*Class 1*	*Class 2*	*Class 3*
	Class 1	TP	E_{12}	E_{13}
	Class 2	E_{21}	TP	E_{23}
	Class 3	E_{31}	E_{32}	TP

1. TP (*True Positive*) menunjukkan jumlah data testing yang diklasifikasikan sistem sesuai dengan kategori yang sesungguhnya.
2. FP (*False Positive*) menunjukkan jumlah data testing pada kolom yang sesuai kelasnya namun tidak termasuk TP. Contoh untuk FP Class 1 = E_{12} + E_{13}
3. FN (*False Negative*) menunjukkan jumlah data testing pada baris yang sesuai kelasnya namun tidak termasuk TP. Contoh untuk FN Class 1 = E_{21} + E_{31}

4. TN (True Negative) menunjukkan jumlah data testing pada semua kolom dan baris namun tidak termasuk kolom dan baris kelas itu. Contoh untuk TN Class 1 = TP Class 2 + E_{23} + E_{32} + TP Class 3.

   Kemudian untuk mencari nilai Precision, Recall dan F1-Score adalah dengan cara menghitung semua Precision, Recall dan F1-Score dari masing masing kelas dengan cara dijumlahkan lalu dibagi dengan banyaknya kelas.

$$Accuracy = \frac{True\ Positif}{Jumlah\ Data}$$

$$All\ Precision = \frac{Precision\ Class\ 1 + Precision\ Class\ 2 + Precision\ Class\ 3}{Jumlah\ Kelas}$$

$$All\ Recal = \frac{Recall\ Class\ 1 + Recall\ Class\ 2 + Recall\ Class\ 3}{Jumlah\ Kelas}$$