BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.2. Tinjauan Pustaka

Analisis wajah secara antropologis dapat digunakan untuk mengidentifikasi perbedaan ras, etnis, dan seksual (Mane dkk., 2010). Skala yang digunakan untuk mengukur parameter morfologi dapat berupa tinggi wajah, lebar wajah, dan indeks wajah. Tinggi wajah morfologis diukur dengan dengan skala dari nasion (n) hingga gnathion (gn). Lebar wajah diukur sebagai jarak lurus antara zygion kanan dan kiri (Yesmin dkk., 2014). Mengekstrasi ciri bentuk wajah panjang, bulat, atau lebar dapat dilakukan dengan menghitung indeks morfologi wajah. Untuk setiap individu indeks wajah dihitung menggunakan rumus seperti yang diberikan oleh Martin dan Saller dalam tampilan depan untuk menentukan jenis wajah. Aplikasi ilmiah dari proporsi wajah dapat digunakan di berbagai bidang seperti pembedahan, ortodontik, dan estetika wajah (Packiriswamy dkk., 2012).

Metode berdasarkan poin fitur wajah dapat digunakan untuk metode try-on kacamata virtual (Feng dkk., 2018). Metode *snapshot* dapat digunakan untuk sistem komersial pemilihan kacamata. Sistem hanya akan mengambil foto pengguna, kemudian kacamata virtual bisa ditumpangkan ke gambar wajah pengguna dengan benar dan secara stabil (Déniz dkk., 2010).

Identifikasi wajah dapat dilakukan dengan *Max Margin Object Detection* (MMOD). Metode ini tidak melakukan sub-sampling, tetapi sebaliknya mengoptimalkan semua sub-jendela atau gambar (King, 2009). *Max Margin Object Detection* (MMOD) dapat digunakan untuk mendeteksi objek dalam gambar. Max Margin mengadopsi metode *Support Vector Machine* (SVM). Metode ini bertujuan mencari hyperplane pemisah antara kelas negatif dan positif. Metode untuk menemukan batas kelas yang diinginkan mirip dengan satu perumusan SVM (Prasad dkk., 2017). Margin yang lebih besar meningkatkan kemungkinan fitur menjadi vektor dukungan. MMOD dapat digunakan untuk meningkatkan metode deteksi objek apa pun yang linear dalam pembelajaran parameter.

Pohon keputusan dapat digunakan sebagai klasifikasi dalam pembelajaran mesin (Zhang dkk., 2019). Decision tree mampu menangani kumpulan data dan dapat menyesuaikan data pelatihan. Salah satu alat paling populer untuk menambang aliran data adalah pohon keputusan. Dalam pohon keputusan, Algoritma CART (*Classifcation and Regression Tree*) digunakan sebagai dasar menentukan atribut terbaik untuk membuat pemisahan dalam simpul (Rutkowski dkk., 2014).

2.2. Dasar Teori

2.2.1. Produk

Produk merupakan titik pusat dari kegiatan pemasaran karena produk merupakan hasil dari suatu perusahaan yang dapat ditawarkan ke pasar untuk di konsumsi dan merupakan alat dari suatu perusahaan untuk mencapai tujuan dari perusahaannya. Suatu produk harus memiliki keunggulan dari produk-produk yang lain baik dari segi kualitas, desain, bentuk, ukuran, kemasan, pelayanan, garansi, dan rasa agar dapat menarik minat konsumen untuk mencoba dan membeli produk tersebut.

Produk adalah segala sesuatu yang dapat ditawarkan ke pasar untuk mendapatkan perhatian, dibeli, digunakan, atau dikonsumsi yang dapat memuaskan keinginan atau kebutuhan (Kotler dan Armstrong, 2008). Secara konseptual produk adalah pemahaman subyektif dari produsen atas sesuatu yang bisa ditawarkan sebagai usaha untuk mencapai tujuan organisasi melalui pemenuhan kebutuhan dan kegiatan konsumen, sesuai dengan kompetensi dan kapasitas organisasi serta daya beli pasar. Selain itu produk dapat pula didefinisikan sebagai persepsi konsumen yang dijabarkan oleh produsen melalui hasil produksinya. Produk dipandang penting oleh konsumen dan dijadikan dasar pengambilan keputusan pembelian.

2.2.2 Citra Digital

Citra digital merupakan suatu fungsi intensitas cahaya f(x, y). Harga x dan y merupakan koordinat spasial dan harga fungsi tersebut pada setiap titik (x, y) merupakan tingkat kecemerlangan citra pada titik tersebut. Masing-masing

element pada citra digital (berarti elemen matriks) disebut image elemen atau piksel. Jadi, citra yang berukuran N x M mempunyai N.M buah piksel.

Proses digitalisai koordinat (x, y) dikenal sebagai pencuplikan citra (image sampling), sedangkat proses digitalisasi derajat keabuan f(x, y) disebut kuantisasi derajat keabuan (gray-level quantization). Berdasarkan format penyimpanan nilai warnanya, citra terdiri atas empat jenis, yaitu:

a) Citra biner atau monokrom

   Pada citra jenis ini, setiap titik atau piksel hanya bernilai 0 atau 1. Setiap titik membutuhkan media penyimpanan sebesar 1 bit.

b) Citra skala keabuan

   Citra skala keabuan mempunyai kemungkinan warna antara hitam (minimal) dan putih (makasimal).

c) Citra warna

   Setiap titik (piksel) pada citra warna mewakili warna yang merupkan kombinasi dari tiga warna dasar yaitu merah, hijai dan biru. Setiap warna dasar mempunyai intensitas sendiri dengan nilai maksimum 255 (8 bit). Setiap titik pada citra warna membutuhkan data 3 byte.

d) Citra warna berindeks

   Setiap titik (piksel) pada citra warna berindeks mewakili indeks dari suatu tabel warna yang tersedia (biasanya disebut palet warna). Keuntungan pemakaian palet warna adalah kita dapat dengan cepat memanipulasi warna tanpa harus mengubah informasi pada setiap titik dalam citra. Keuntungan lain penyimpanan lebih kecil.

2.2.3 Morfologi Indeks Wajah

Pengamatan sehari-hari membawa pengalaman bahwa manusia ternyata cukup bervariasi. Kenyataan ini mendorong manusia untuk melihat perbedaan-perbedaan dengan lebih teliti dengan mendefinisikan bentuk dan ukuran tubuh. Ukuran hanya memberikan informasi tentang besar-kecilnya model yang diukur (*size*), karenanya untuk mengungkapkan bentuk (*shape*) diciptakan proporsi antara ukuran-ukuran yang dinamakan indeks (Glinka dkk., 2007).

Indeks wajah dihitung menggunakan rumus seperti yang diberikan oleh Martin dan Saller dalam tampilan *Frontal* untuk menentukan jenis wajah (Mane dkk., 2010) dengan formula sebagai berikut :

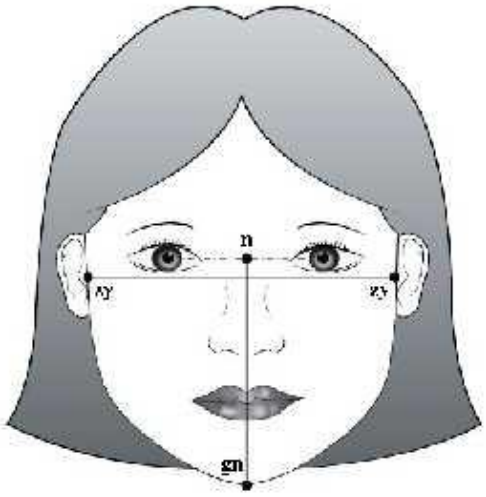
$$I = \frac{P}{l} \times 100 \qquad (2.1)$$

dengan I = Index wajah

P = Panjang wajah

l = lebar wajah

Panjang wajah diambil dari jarak antara *nasion* (n) dan *gnathion* (gn) dan lebar wajah sebagai jarak antara titik *zygoma* kanan dan kiri. Gambar titik wajah dapat dilihat pada gambar 2.1:



Gambar 2.1 Titik Wajah (Yesmin dkk., 2014)

Setelah menghitung indeks wajah, tipe wajah diklasifikasikan seperti yang ditunjukkan pada Tabel 2.1.

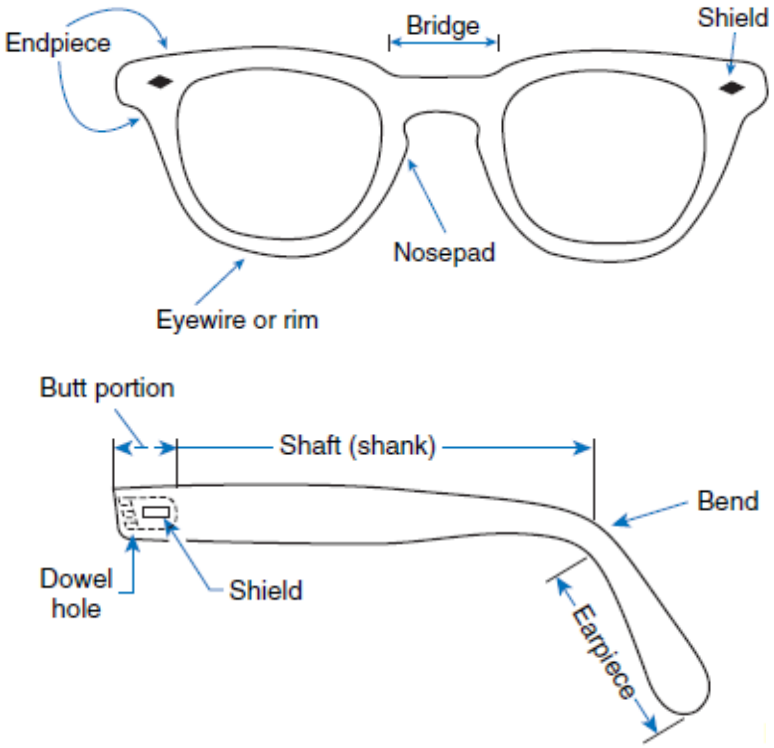
Tabel 2.1. Klasifikasi jenis wajah (Glinka dkk., 2007)

Tipe Wajah	Laki-laki	Perempuan
Hypereuryprosop	≤ 78,9	≤ 76,9
Euryprosop	79,0 – 83,9	77,0 – 80,9
Mesoprosop	84,0 – 87,9	81,0 – 84,9
Leptoprosop	88,0 – 92,9	85,0 – 89,9
Hyperleptoprosop	≥ 93,0	≥ 90,0

Bentuk wajah dalam indeks prosopic subjek *leptoprosop* dan *hyperleptoprosop* memiliki bentuk wajah oval, *mesoprosopic* untuk wajah kotak sedangkan individu dengan bentuk bulat dapat dianggap sebagai *euryprosop* atau *hypereuriprosop*, dengan dimensi horizontal tertinggi (Mane dkk., 2010).

2.2.4 Jenis dan Bagian Bingkai Kacamata

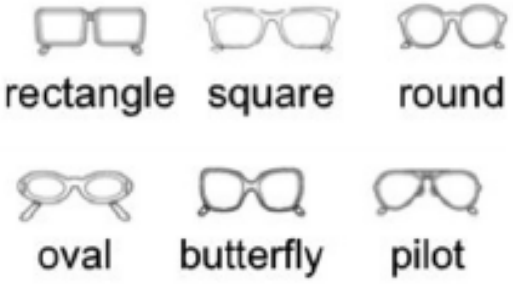
Bingkai adalah bagian dari kacamata yang menopang lensa pada posisi yang tepat di depan mata. Area bingkai depan antara lensa yang terletak di hidung disebut *bridge*. Beberapa bingkai memiliki bantalan hidung (*nosepad*), yang merupakan potongan plastik yang menempel di hidung untuk mendukung bingkai. Bantalan hidung ini mungkin langsung melekat pada bingkai atau untuk menghubungkan potongan logam yang dikenal sebagai lengan pelindung atau lengan bantalan (Clifford., 2007) . Gambar atribut kacamata dapat dilihat pada Gambar 2.2.



Gambar 2.2 Atribut Kacamata (Clifford., 2007)

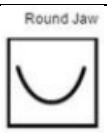
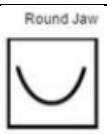
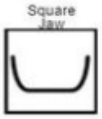
Dari sudut pandang estetika, aspek kacamata tidak kalah pentingnya bagi orang yang memakainya. Setiap individu mengharapkan bantuan untuk penilaian kesesuaian kacamata, tidak hanya dengan ukuran, tetapi juga dengan aspek estetika bingkai (Clifford dkk., 2007). Karena bingkai sangat jelas pada wajah, bentuknya cenderung menekankan atau mengurangi ciri-ciri wajah. Pilihan bingkai yang baik dapat disederhanakan dengan mempertimbangkan terlebih dahulu garis wajah mana yang saling melengkapi untuk orang tersebut. Bentuk wajah pada dasarnya dibagi menjadi 3 yaitu, kotak, bulat dan lonjong (oval). Wajah oval dianggap normal dan dapat memakai hampir semua bingkai, wajah bulat (*round*) dapat memakai bentuk bingkai yang lebih bersudut runcing (rectangle, square), sedangkan wajah kotak (*square*) dapat memakai bingkai yang bulat (*oval, round*) untuk menyamarkan garis wajah. Pengetahuan tentang bentuk wajah dasar merupakan bantuan yang berharga dalam membuat keputusan yang lebih cepat dan lebih akurat tentang bingkai tertentu. Bentuk kacamata (Gu dkk., 2017) dapat dilihat pada Tabel 2.2.

Tabel 2.2 Kategori dan macam-macam bentuk kacamata (Gu dkk., 2017)

Kategori Kacamata	
Frame Shape	

Ringkasan kesesuaian antara bentuk wajah dan bingkai kacamata sesuai prinsip Zen dapat dilihat pada Tabel 2.3.

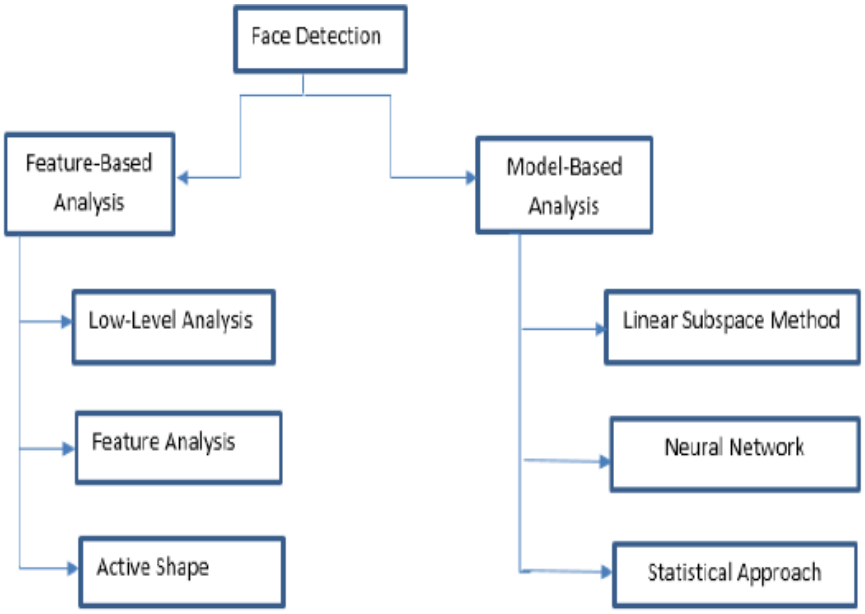
Tabel 2.3 Tabel Bentuk Wajah

No	Wajah	Ciri		Kacamata
		Tipe wajah	Bentuk *Jawline*/rahang	
1	Oval	*Leptoprosop/ Hyperleptoprosop*		Semua tipe
2	Bundar (Round)	*Euryprosop/ Hypereuriprosop*		Rectangle Square
3	Kotak (Square)	*Mesoprosopic*		Round Oval

Hidung, tulang pipi, dan rahang merupakan tiga tatanan wajah yang mendasar untuk keharmonisan wajah (Terino dan Edwards, 2008). Bentuk wajah adalah salah satu fitur menonjol dalam sistem pengambilan wajah. Berdasarkan struktur, wajah oval mempunyai garis rahang tidak terlalu bersudut. Bentuk wajah bundar mempunyai bentuk rahang tidak bersudut yang lebar, sedangkan bentuk wajah kotak mempunyai rahang lebih bersudut (Khan dan Jalal, 2020).

2.2.5 Metode Deteksi Wajah

Deteksi wajah dapat dilakukan dengan mudah oleh manusia, tetapi dalam hal *Computer Vision*, itu bukan tugas yang mudah (Singh dkk., 2017). Deteksi wajah memutuskan apakah ada wajah atau tidak di dalam *input* gambar. Ada dua pendekatan deteksi wajah yaitu pendekatan berbasis fitur dan berbasis Model dapat dilihat seperti gambar 2.3 berikut:



Gambar 2.3. Pendekatan Deteksi Wajah (Singh dkk., 2017)

Pendekatan berbasis fitur mencoba mengekstrak fitur yang ada pada gambar untuk mencocokkannya dengan pengetahuan fitur wajah, sedangkan pendekatan berbasis model mencoba untuk mendapatkan kecocokan terbaik antara pelatihan dan pengujian gambar. Metode Linear subspace, Neural network, dan pendekatan statistic seperti Support Vector Machine (SVM) merupakan pendekatan berbasis model. Linear subspace didasarkan pada perhitungan atau wajah sendiri dengan menurunkan vektor sendiri. Gambar wajah manusia terletak dalam ruang bagian atau ruang gambar keseluruhan. Ada banyak teknik terkenal seperti Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), dan Factor Analysis (FA). Support Vector Machines (SVM) adalah teknik pembelajaran *supervised* berbasis kernel yang banyak digunakan untuk klasifikasi dan regresi. Ide dasar SVM adalah untuk menemukan hyperplane yang optimal untuk pola yang dapat dipisahkan secara linier. Memaksimalkan margin geometris pada set pelatihan dan meminimalkan kesalahan pelatihan. Kemudian, fungsi kernel memetakan data asli ke dalam ruang baru untuk kasus-kasus yang terpisah secara non-linear. SVM juga digunakan untuk mendeteksi wajah. Sedangkan pendekatan berbasis fitur dibagi menjadi tiga:

1. Analisis tingkat rendah: mengolah inputan dengan segmentasi komponen visual seperti deteksi tepi, analisis skala abu-abu, *motion*, dan informasi warna.

a) Deteksi wajah berdasarkan tepi menganalisis tepi wajah dari gambar untuk mencari berbagai fitur wajah.
b) Deteksi wajah berdasarkan informasi abu-abu juga dapat digunakan sebagai fitur wajah yang penting. Fitur wajah seperti alis, bibir, dan pupil umumnya tampak lebih gelap daripada daerah wajah di sekitarnya. Dalam algoritma berbasis analisis tingkat abu-abu, gambar input ditingkatkan dengan kontras dan rutinitas morfologi skala-abu untuk meningkatkan kualitas *patch* gelap dan mempermudah pendeteksian.
c) Dalam pendekatan berbasis gerakan, informasi gerakan digunakan untuk mencari objek bergerak di hadapan urutan video. Memindahkan siluet (garis luar gelap) seperti wajah dan bagian tubuh lainnya dapat diekstraksi pada dasarnya dengan mengkalkulasi perbedaan bingkai yang terakumulasi. Fitur wajah dapat ditemukan oleh perbedaan bingkai di samping daerah wajah.
d) Dalam metode berbasis warna, warna kulit dapat digunakan dalam informasi fitur wajah seperti geometri, bentuk, dll. Ada berbagai ruang warna seperti HSV (Hue, Saturation, Value). Ruang warna HSV adalah salah satu dari beberapa sistem warna yang digunakan orang untuk memilih warna.

2. Analisis fitur: Tujuan dari algoritma analisis fitur untuk mencari fitur struktural dengan berbagai pose, sudut pandang, atau kondisi pencahayaan. Fitur struktural digunakan untuk mencari wajah. Beberapa algoritma fitur analisis adalah metode Viola Jones, metode fitur Gabor dan metode Constellation.

3. *Active shape models*: *Active shape models* (ASM) berkonsentrasi pada fitur non-kaku yang kompleks seperti tampilan fitur fisik dan tingkat yang lebih tinggi. ASM diharapkan secara otomatis menemukan *landmark* yang mencirikan bentuk objek dan fitur wajah yang dimodelkan secara statistik seperti mata, hidung, bibir, mulut, dan alis dalam sebuah gambar.

2.2.6 *Machine Learning*

Machine learning merupakan cabang dari kecerdasan buatan yang menghubungkan masalah belajar dari sampel data dengan konsep umum inferensi (Kourou dkk., 2015). Algoritma pembelajaran mesin telah menjadi alat yang berguna untuk analisis data dalam banyak aplikasi. Algoritma pembelajaran mesin dapat mengekstrak informasi, dengan demikian, sistem yang dibantu komputer berdasarkan pembelajaran mesin membantu manusia untuk membuat keputusan berdasarkan informasi.

Ada dua jenis umum metode *machine learning* yang dikenal sebagai *supervised learning* dan *unsupervised learning*. Dalam pembelajaran *supervised*, data pelatihan yang digunakan untuk memperkirakan atau memetakan adalah input data ke output yang diinginkan. Sebaliknya, metode pembelajaran *unsupervised* tidak ada contoh berlabel yang disediakan dan tidak ada gagasan tentang output selama proses pembelajaran.

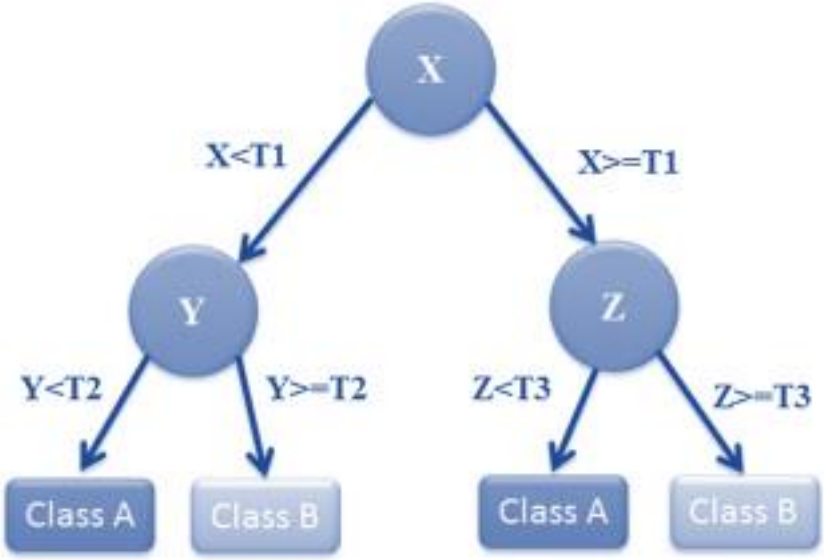
Machine learning tanpa data maka tidak akan bisa bekerja. Data dibagi menjadi 2 kelompok, yaitu data *training* dan data *testing*. *Data training* digunakan untuk melatih algoritma untuk mencari model yang cocok, sementara data *testing* dipakai untuk menguji dan mengetahui performa model yang didapatkan pada tahapan *testing.*

Dari model yang didapatkan, dapat dilakukan prediksi yang dibedakan menjadi dua macam, tergantung tipe keluarannya. Jika hasil prediksi bersifat diskrit, maka dinamakan proses klasifikasi. Sedangkan jika kelurannya bersifat kontinyu, maka dinamakan proses regresi.

2.2.7 *Decision Tree*

Pohon keputusan adalah salah satu pendekatan klasifikasi yang paling populer dalam pembelajaran mesin. *Decision Tree* merupakan metode pembelajaran *supervised* non-parametrik yang digunakan untuk klasifikasi dan regresi. Tujuannya adalah untuk membuat model yang memprediksi nilai variabel target dengan mempelajari aturan keputusan sederhana yang disimpulkan dari fitur data. Pohon keputusan terdiri dari "*root*", "daun", dan *node* internal. *Node* internal

menggunakan fitur tertentu untuk membagi ruang *instance* menjadi dua atau lebih subruang. Setiap daun mewakili satu kelas. Daun dapat mewakili nilai target yang paling tepat atau menunjukkan kemungkinan target memiliki nilai spesifik. Pohon keputusan belajar dari data untuk memperkirakan keputusan dengan seperangkat aturan keputusan if-then-else. Semakin dalam pohon, semakin kompleks aturan keputusan. Contoh model pohon keputusan dapat dilihat pada Gambar 2.4



Gambar 2.4 Ilustrasi *Decision Tree* yang menunjukkan struktur pohon

Pada gambar 2.4 Setiap variabel (X, Y, Z) diwakili oleh lingkaran dan hasil keputusan dengan kuadrat (Kelas A, Kelas B). T (1-3) mewakili ambang (aturan klasifikasi) agar berhasil mengklasifikasikan setiap variabel ke label kelas. *Decision tree* ini digambarkan sebagai sebuah diagram alir yang berbentuk seperti struktur pohon yang mana setiap *internal node* menyatakan pengujian terhadap suatu atribut, setiap cabang menyatakan output dari pegujian tersebut dan *leaf node* menyatakan kelas–kelas atau distribusi kelas. *Node* yang paling atas disebut sebagai *root node* atau *node* akar. Sebuah *root node* akan memiliki beberapa *edge* keluar tetapi tidak memiliki *edge* masuk, *internal node* akan memiliki satu *edge* masuk dan beberapa *edge* keluar, sedangkan *leaf node* hanya akan memiliki satu *edge* masuk tanpa memiliki *edge* keluar. *Decision tree* digunakan untuk

mengklasifikasikan suatu sampel data yang belum diketahui kelasnya ke dalam kelas–kelas yang sudah ada.

Pemilihan atribut untuk menjadi rootnode atau internal node sebagai atribut test berdasarkan atas ukuran impurity dari masing–masing atribut. Ukuran–ukuran impurity yang umumnya digunakan adalah *information gain, gain ratio dan gini index*. Information Gain merupakan suatu ukuran korelasi pada model parametrik yang menggambarkan ketergantungan antara dua peubah acak X dan Y. *Information Gain* memiliki rumus:

$$Gain(A) = I(S_1, S_2, .., S_n) - E(A) \quad (2.2)$$

Gain Ratio merupakan modifikasi dari information gain untuk mengurangi bias atribut yang memiliki banyak cabang. Gain ratio memiliki rumus:

$$Gain\ Ratio = \frac{Gain}{Split\ Info} \quad (2.3)$$

Gini index merupakan suatu ukuran ketidaksamaan pada distribusi pendapatan dan memiliki nilai antara 0 sampai 1. Semakin rendah nilai Gini index maka semakin besar pula ukuran kesamaannya. Gini index atribut t untuk data dengan m kelas didefinisikan sebagai berikut:

$$Gini(t) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2.4)$$

Ukuran–ukuran tersebut biasanya hanya digunakan pada algoritma tertentu, jadi penentuan ukuran untuk digunakan dalam memilih atribut test sangat dipengaruhi algoritma yang dipilih. Berdasarkan banyaknya edge keluar dari suatu atribut, maka terdapat dua jenis pemisahan yaitu binary split yang menghasilkan dua buah edge keluar dan multyway split yang menghasilkan lebih dari dua edge keluar.

2.2.8 Algoritma CART

Decision Tree klasifikasi CART (*Classification and Regression Trees*) merupakan metode nonparametrik yang berguna untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. Metode klasifikasi CART terdiri dari dua metode yaitu metode pohon regresi dan pohon klasifikasi. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (classification trees). Sedangkan jika variabel

dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (regression trees). CART sangat mirip dengan C4.5, tetapi berbeda karena mendukung variabel target numerik (regresi) dan tidak menghitung set aturan. CART membangun pohon biner menggunakan fitur dan ambang batas yang menghasilkan perolehan informasi terbesar di setiap node. Langkah-langkah penerapan metode CART:

1. Pembentukan pohon

   Proses pembentukan pohon kalsifikasi terdiri atas 3 tahapan yaitu:

a. Pemilahan Pemilah (Classifier)

   Untuk membentuk pohon klasifikasi digunakan sampel data *learning* (L) yang masih bersifat heterogen. Sampel tersebut akan dipilah berdasarkan aturan pemilahan. Pemilahan pemilah tergantung pada jenis tree atau lebih tepatnya tergantung pada jenis varietas responnya. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu simpul tertentu dalam pohon klasifikasi dikenal dengan istilah *impurity measure* $\mathrm{i}(t)$. ukuran ini membantu menemukan fungsi pemilah yang optimal. Kualitas ukuran dari seberapa baik pemilah dalam menyaring data menurut kelas merupakan ukuran penurunan keheterogenan dari suatu kelas

b. Penentuan Simpul Terminal

   Suatu simpul t akan menjadi simpul terminal atau tidak akan dipilah kembali apabila pada simpul t tidak terdapat penurunan keheterogenan secara berat atau adanya batasan minimum n.

c. Penandaan Label Kelas

   Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak.

2. Pemangkasan Pohon Klasifikasi

   Pemangkasan dilakukan dengan jalan memangkas bagian tree yang kurang penting sehingga didapatkan pohon optimal. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran tree yang layak adalah *cost complexity minimum*.

3. Penentuan Pohon Kasifikasi Optimal

   Pohon klasifikasi yang berukuran besar akan memberikan nilai penduga pengganti paling kecil, sehingga tree ini cenderung dipilih untuk mendunga nilai respon.

Algoritma CART dimulai dengan satu simpul L_0 - root. Selama proses pembelajaran, di setiap simpul L_q yang dibuat subset tertentu S_q dari dataset pelatihan S diproses (untuk root S_0=S). Jika semua elemen atau set S_q memiliki kelas yang sama, node ditandai sebagai daun dan pemisahan tidak dibuat. Jika tidak, menurut fungsi ukuran split, atribut terbaik untuk dibagi dipilih di antara atribut yang tersedia di simpul yang dipertimbangkan. Untuk setiap atribut yang tersedia a^i, set nilai atribut A^i dipartisi menjadi dua himpunan bagian yang terpisah A_L^idan $A_R^i(A^i = A_L^i \cup A_R^i)$. Pilihan A_L^i secara otomatis adalah subset pelengkap A_R^i, oleh karena itu partisi hanya diwakili oleh A_L^i. Himpunan semua partisi atau himpunan yang mungkin A^i dilambangkan dengan V_i. Himpunan bagian A_L^i dan A_R^i membagi dataset S_q menjadi dua himpunan bagian yang terpisah: kiri $L_q(A_L^i)$ dan kanan $R_q(A_L^i)$

$$L_q(A_L^i) = \{S_j \in S_q | v_j^i \in A_L^i\}, \quad (2.5)$$

$$R_q(A_L^i) = \{S_j \in S_q | v_j^i \in A_R^i\}, \quad (2.6)$$

Himpunan $L_q(A_L^i)$ dan $R_q(A_L^i)$ bergantung pada atribut dan partisi yang dipilih dari nilainya. $p_{L,q}(A_L^i)$ $\left(p_{R,q}(A_L^i)\right)$ menunjukkan bagian elemen data dari S_q, yang termasuk dalam subset $L_q(A_L^i)\left(R_q(A_L^i)\right)$. Karena pecahan $p_{L,q}(A_L^i)$ dan $p_{R,q}(A_L^i)$ tergantung pada

$$p_{R,q}(A_L^i) = 1 - p_{L,q}(A_L^i), \quad (2.7)$$

hanya satu dari parameter ini yang perlu dipertimbangkan, misal $p_{L,q}(A_L^i)$. Fraksi elemen dari $L_q(A_L^i)\left(R_q(A_L^i)\right)$, dari kelas k, dilambangkan dengan $p_{kL,q}(A_L^i)\left(p_{kR,q}(A_L^i)\right)$. Fraksi semua elemen data S_q dalam simpul yang dianggap L_q, dari kelas k, dilambangkan dengan $p_{k,q}$. Perhatikan bahwa $p_{k,q}$, k=1,…., K tidak tergantung pada atribut a^i dan partisi A_L^i yang dipilih. Seperti yang

disebutkan sebelumnya, ukuran pengotor yang digunakan dalam algoritma CART adalah indeks Gini. Untuk setiap subset S_q dari set data pelatihan diberikan oleh :

$$Gini(S_q) = 1 - \sum_{k=1}^{K}(p_{k,q})^2. \quad (2.8)$$

Indeks Gini mencapai minimum (nol) ketika semua kasus masuk ke dalam kategori target tunggal, dan maksimum diperoleh ketika catatan didistribusikan secara merata di antara semua kelas. Selanjutnya, indeks dari subset S_q Gini tertimbang, yang dihasilkan dari pilihan partisi A_L^i, didefinisikan sebagai berikut

$$wGini(S_q, A_L^i) = p_{L,q}(A_L^i)\, Gini\left(L_q(A_L^i)\right) + \left(1 - p_{L,q}(A_L^i)\right) Gini\left(R_q(A_L^i)\right) \quad (2.9)$$

Gini mengindeks dari set $L_q(A_L^i)$ dan $R_q(A_L^i)$ diberikan secara analog seperti pada rumus (2.8)

$$Gini\left(L_q(A_L^i)\right) = 1 - \sum_{k=1}^{K}\left(p_{kL,q}(A_L^i)\right)^2 \quad (2.10)$$

$$Gini\left(R_q(A_L^i)\right) = 1 - \sum_{k=1}^{K}\left(p_{kR,q}(A_L^i)\right)^2 \quad (2.11)$$

Fungsi ukuran terpisah dalam algoritma CART didefinisikan sebagai perbedaan antara indeks Gini (2.8) dan indeks Gini tertimbang (2.9).

2.2.9 Pengukuran Pengujian Kinerja dan Validasi Prediksi

Suatu sistem yang menggunakan algoritma klasifikasi perlu dilakukan pengujian kinerja, hal ini dilakukan untuk mengetahui seberapa baik sistem dalam menjalankan klasifikasi data. Confusion Matrix digunakan untuk mengevaluasi kualitas output atau classifier pada set data iris. Elemen-elemen diagonal mewakili jumlah titik label yang diprediksi sama dengan label yang sebenarnya, sedangkan elemen off-diagonal adalah mereka yang salah diberi label oleh classifier. Semakin tinggi nilai diagonal dari matriks kebingungan semakin baik, menunjukkan banyak prediksi yang benar. Jumlah prediksi yang benar dan salah dirangkum dengan nilai-nilai hitung dan dipecah oleh masing-masing kelas. Penggambaran Confusion Matrik dapat di lihat pada Gambar 2.5.

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

Gambar 2.5 Confusion Matrix

Ada dua kemungkinan kelas yang diprediksi: "ya" dan "tidak". Hasil dari pengamatan dan prediksi klasifikasi diistilahkan sebagai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). True Positive (TP) adalah jumlah data untuk pengamatan positif dan diprediksi positif, False Negative (FN) jumlah data untuk pengamatan positif, tetapi diprediksi negative, True Negative (TN) jumlah data untuk pengamatan negatif dan diprediksi negative, sedangkan False Positive (FP) adalah jumlah data untuk pengamatan negatif, tetapi diprediksi positif. Berdasarkan jumlah nilai tersebut dapat diperoleh nilai akurasi, recall dan presisi. Tingkat Klasifikasi atau Akurasi diberikan persamaan:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.12)$$

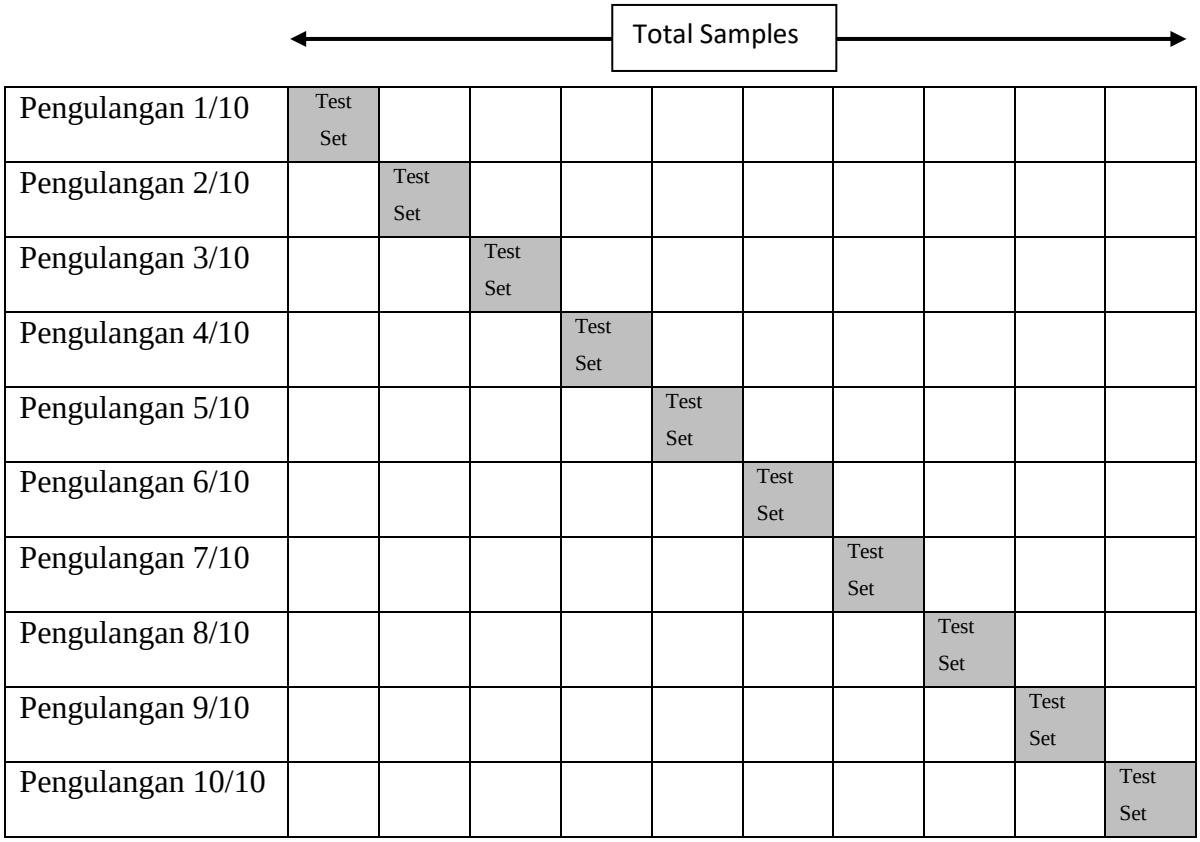
Recall dapat didefinisikan sebagai rasio dari jumlah total positif yang diklasifikasikan dengan benar dibagi dengan jumlah total positif. High Recall menunjukkan kelas dikenali dengan benar. Persamaan untuk menghitung recall adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (2.13)$$

Sedangkan untuk mendapatkan nilai presisi, dapat dilakukan dengan membagi jumlah total positif yang diklasifikasikan dengan benar dengan jumlah total positif yang diprediksi. Presisi Tinggi menunjukkan contoh berlabel positif memang positif. Persamaan untuk menghitung presisi adalah sebagai berikut;

$$Presisi = \frac{TP}{TP+FP} \quad (2.14)$$

Evaluasi validasi dari suatu model dalam proses klasifikasi perlu dilakukan. Penelitian ini menggunakan metode K-fold Cross-Validation untuk memvalidasi model algoritma CART Decision Three. Cross-Validation merupakan salah satu teknik untuk menilai/memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Dalam teknik Cross-Validation ini dataset dibagi menjadi sejumlah K-buah partisi secara acak. Kemudian dilakukan sejumlah K-kali eksperimen, masing-masing eksperimen menggunakan data partisi ke-K sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training. Pada penelitian ini K-fold cross-validation menggunakan 10-fold yang artinya data dibagi menjadi 10-fold berukuran kira-kira sama, sehingga dimiliki 10 subset data untuk mengevaluasi kinerja model atau algoritma. Dalam setiap subset, cross-validation memisahkan data set menjadi 2 bagian kemudian menggunakan satu bagian fold sebagai data uji dan 9-fold lainnya sebagai data latihan. Untuk mendapatkan nilai akurasi ataupun ukuran penilaian lainnya dari hasil eksperimen yang dilakukan, dapat diambil nilai rataan dari seluruh eksperimen tersebut. Sebagai gambaran jika dilakukan 10-Fold Cross-Validation maka skema desain data eksperimennya adalah seperti Gambar 2.6.



Gambar 2.6 Skema 10-Fold Cross-Validation