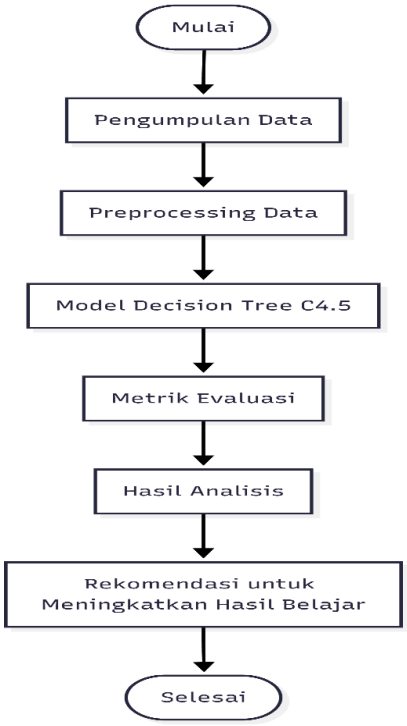
BAB III

METODELOGI PENELITIAN

3.1 Kerangka Penelitian

Kerangka kerja penelitian ini memberikan alur sistematis untuk menganalisis pola belajar siswa menggunakan data mining dengan algoritma C4.5. Kerangka ini mencakup tahapan mulai dari pengumpulan dan pengolahan data hingga pembentukan serta interpretasi model klasifikasi. Tujuannya adalah mengidentifikasi pola belajar yang memengaruhi hasil belajar Bahasa Inggris, sehingga dapat memberikan informasi bagi pengambilan keputusan dalam penyusunan strategi pembelajaran. Setiap tahap dirancang agar analisis lebih efisien, akurat, dan sesuai kebutuhan pendidik.



Gambar 3.1. Kerangka Penelitian

3.2 Populasi dan Sampel

3.2.1 Populasi

Populasi dalam penelitian ini adalah siswa kelas X (Sepuluh) SMK Swasta Al-Washliyah 2 Marbau yang dipilih berdasarkan ketersediaan dan kelengkapan data pembelajaran Bahasa Inggris. Populasi penelitian tidak mencakup seluruh siswa di sekolah tersebut, melainkan hanya siswa kelas X (Sepuluh) yang memiliki data terkait absensi, keikutsertaan dalam kelas tambahan, penggunaan media digital, serta nilai hasil belajar Bahasa Inggris. Pemilihan populasi ini dilakukan karena penelitian menggunakan pendekatan data mining yang membutuhkan data lengkap dan relevan untuk dianalisis menggunakan algoritma Decision Tree C4.5.

3.2.2 Sampel

Sampel dalam penelitian ini merupakan sebagian siswa yang diambil dari populasi penelitian. Teknik pengambilan sampel yang digunakan adalah purposive sampling, yaitu pemilihan sampel berdasarkan kriteria tertentu yang sesuai dengan tujuan penelitian. Sampel terdiri dari siswa kelas X yang memiliki data lengkap mengenai absensi, kelas tambahan, penggunaan media digital, serta nilai Bahasa Inggris. Sampel yang diperoleh dianggap telah mewakili karakteristik populasi dan digunakan sebagai dasar dalam pembentukan model pohon keputusan dengan algoritma Decision Tree C4.5 untuk menganalisis pengaruh variabel absensi, kelas tambahan, dan penggunaan media digital terhadap nilai hasil belajar Bahasa Inggris siswa.

3.3 Lokasi dan Waktu Penelitian

3.3.1 Lokasi Penelitian

Penelitian ini dilaksanakan di SMK Swasta Al-Washliyah 2 Marbau, yang berada di bawah naungan Yayasan Al-Washliyah di Kecamatan Marbau, Kabupaten Labuhanbatu Utara, Provinsi Sumatera Utara. Pemilihan sekolah ini didasarkan pada karakteristik institusi yang meliputi pendidikan vokasional, sehingga dapat memberikan gambaran yang lebih fokus tentang pola belajar siswa dalam konteks pendidikan kejuruan. Penelitian ini bertujuan untuk mendapatkan data yang lebih mendalam dalam menganalisis pola belajar Bahasa Inggris, khususnya yang berkaitan dengan kehadiran, waktu belajar, dan pemanfaatan teknologi digital dalam pembelajaran.

3.3.2 Waktu Penelitian

Proses penelitian ini direncanakan berlangsung dari bulan Januari – Maret 2026. Rentang waktu ini dipilih untuk memberikan cukup waktu dalam pengumpulan data dan analisis mendalam terhadap pola belajar siswa, khususnya pada mata pelajaran Bahasa Inggris. Durasi yang cukup panjang juga memberikan kesempatan untuk mengkaji perubahan atau perkembangan dalam pola belajar siswa selama periode tersebut, serta memastikan data yang dikumpulkan memiliki validitas yang tinggi.

3.4 Variabel Penelitian

Variabel yang diteliti dalam penelitian ini terdiri atas dua kategori utama, yaitu variabel bebas dan variabel terikat. Variabel penelitian merupakan objek yang dianalisis untuk menjawab rumusan masalah dan mencapai tujuan penelitian.

Variabel bebas meliputi absensi, kelas tambahan, dan penggunaan media digital, sedangkan variabel terikat adalah nilai siswa. Penelitian ini menggunakan pendekatan data mining dengan algoritma Decision Tree C4.5 untuk menganalisis pola belajar siswa dan hubungannya dengan hasil belajar.

Tabel 3.1 Variabel penelitian

Kode	**Variabel**	**Kategori**
X1	Absensi	≥7, <7
X2	Kelas Tambahan	Ya, Tidak
X3	Penggunaan Media Digital	Ya, Tidak
Y	Nilai	KKM, Diatas KKM

3.5 Langkah – Langkah Pengelolaan Algoritma C.45

3.5.1 Pengumpulan Data

Pengumpulan data dalam penelitian ini bertujuan untuk memperoleh data yang akurat dan relevan guna menganalisis pola belajar siswa menggunakan algoritma Decision Tree C4.5. Pada tahap pertama pengumpulan data, informasi yang dikumpulkan meliputi absensi, kelas tambahan, penggunaan media digital, dan nilai yang diperoleh. Data tersebut digunakan sebagai dasar dalam proses pengolahan data mining untuk menentukan pola belajar siswa terhadap hasil belajar Bahasa Inggris.

Tabel 3.2 Sampel Data Mentah

Nama	Absensi (X1)	Kelas Tambahan (X2)	Media Digital (X3)	Nilai (Y)
S01	3	Ya	Tidak	80
S02	8	Tidak	Ya	77
S03	5	Ya	Tidak	77
S04	7	Ya	Tidak	77
S05	2	Ya	Ya	79
S06	4	Tidak	Tidak	77
S07	8	Tidak	Ya	77
S08	4	Ya	Tidak	77
S09	7	Ya	Ya	77
S10	2	Tidak	Tidak	80
S11	3	Ya	Tidak	77
S12	7	Tidak	Tidak	77
S13	5	Ya	Tidak	77
S14	7	Ya	Ya	77
S15	3	Ya	Tidak	81
S16	3	Ya	Ya	79
S17	8	Ya	Tidak	77
S18	4	Ya	Ya	78
S19	9	Tidak	Ya	77
S20	3	Tidak	Tidak	77
S21	4	Ya	Ya	80
S22	8	Ya	Tidak	77
S23	2	Tidak	Ya	77
S24	7	Ya	Ya	77
S25	5	Ya	Tidak	77
S26	8	Tidak	Tidak	77
S27	2	Ya	Ya	79
S28	4	Tidak	Tidak	77
S29	8	Ya	Ya	77
S30	3	Tidak	Tidak	77

3.5.2 Preprocessing Data

Tahap 1 Preprocessing Data dan Transformasi Data

Tabel 3.3 Data Preprocessing dan Transformasi

Nama	Absensi (X1)	KelasTambahan (X2)	Media Digital (X3)	Nilai (Y)
S01	<7	Ya	Tidak	Diatas KKM
S02	≥7	Tidak	Ya	KKM
S03	<7	Ya	Tidak	KKM
S04	≥7	Ya	Tidak	KKM
S05	<7	Ya	Ya	Diatas KKM
S06	<7	Tidak	Tidak	KKM
S07	≥7	Tidak	Ya	KKM
S08	<7	Ya	Tidak	KKM
S09	≥7	Ya	Ya	KKM
S10	<7	Tidak	Tidak	Diatas KKM
S11	<7	Ya	Tidak	KKM
S12	≥7	Tidak	Tidak	KKM
S13	<7	Ya	Tidak	KKM
S14	≥7	Ya	Ya	KKM
S15	<7	Ya	Tidak	Diatas KKM
S16	<7	Ya	Ya	Diatas KKM
S17	≥7	Ya	Tidak	KKM
S18	<7	Ya	Ya	Diatas KKM
S19	≥7	Tidak	Ya	KKM
S20	<7	Tidak	Tidak	KKM
S21	<7	Ya	Ya	Diatas KKM
S22	≥7	Ya	Tidak	KKM
S23	<7	Tidak	Ya	KKM
S24	≥7	Ya	Ya	KKM
S25	<7	Ya	Tidak	KKM
S26	≥7	Tidak	Tidak	KKM
S27	<7	Ya	Ya	Diatas KKM
S28	<7	Tidak	Tidak	KKM
S29	≥7	Ya	Ya	KKM
S30	<7	Tidak	Tidak	KKM

Tabel preprocessing data ini menyajikan hasil pengolahan awal dataset penelitian yang masih mempertahankan bentuk data kategorikal tanpa melakukan engkodean numerik. Pada tahap preprocessing, data terlebih dahulu diseleksi

dengan memilih atribut yang relevan dengan tujuan penelitian, yaitu absensi (X1), kelas tambahan (X2), penggunaan media digital (X3), dan nilai hasil belajar (Y). Selanjutnya dilakukan pembersihan data untuk memastikan tidak terdapat data ganda, data kosong, maupun kesalahan penulisan pada setiap atribut. Seluruh data dinyatakan lengkap dan konsisten sehingga tidak memerlukan penghapusan atau perbaikan. Tahap transformasi dilakukan dengan menyeragamkan penulisan kategori, seperti penggunaan nilai ≥7 dan <7 pada variabel absensi, kategori Ya dan Tidak pada variabel kelas tambahan dan media digital, serta kategori Di atas KKM, dan KKM pada variabel nilai. Dataset yang ditampilkan dalam tabel ini terdiri dari 30 data siswa dan telah siap digunakan dalam proses analisis menggunakan algoritma Decision Tree C4.5, karena algoritma tersebut mampu mengolah data kategorikal secara langsung.

Tahap 2 Split Data

Tabel 3.4 Data Training

Nama	**Absensi (X1)**	**Kelas Tambahan (X2)**	**Media Digital (X3)**	**Nilai (Y)**
S01	<7	Ya	Tidak	Diatas KKM
S02	≥7	Tidak	Ya	KKM
S03	<7	Ya	Tidak	KKM
S04	≥7	Ya	Tidak	KKM
S05	<7	Ya	Ya	Diatas KKM
S06	<7	Tidak	Tidak	KKM
S07	≥7	Tidak	Ya	KKM
S08	<7	Ya	Tidak	KKM
S09	≥7	Ya	Ya	KKM
S10	<7	Tidak	Tidak	Diatas

				KKM
S11	<7	Ya	Tidak	KKM
S12	≥7	Tidak	Tidak	KKM
S13	<7	Ya	Tidak	KKM
S14	≥7	Ya	Ya	KKM
S15	<7	Ya	Tidak	Diatas KKM
S16	<7	Ya	Ya	Diatas KKM
S17	≥7	Ya	Tidak	KKM
S18	<7	Ya	Ya	Diatas KKM
S19	≥7	Tidak	Ya	KKM
S20	<7	Tidak	Tidak	KKM
S21	<7	Ya	Ya	Diatas KKM
S22	≥7	Ya	Tidak	KKM
S23	<7	Tidak	Ya	KKM
S24	≥7	Ya	Ya	KKM

Tabel data training ini terdiri dari 24 data siswa (S01–S24) yang digunakan untuk menganalisis hubungan antara absensi (X1), kelas tambahan (X2), dan penggunaan media digital (X3) terhadap hasil nilai siswa (Y). Variabel absensi dibedakan menjadi kurang dari 7 (<7) dan lebih dari atau sama dengan 7 (≥7), kelas tambahan dan media digital masing-masing memiliki kategori Ya dan Tidak, sedangkan variabel nilai diklasifikasikan menjadi KKM dan Di atas KKM. Data ini menunjukkan bahwa sebagian besar siswa berada pada kategori nilai KKM, sementara siswa dengan nilai Di atas KKM cenderung mengikuti kelas tambahan dan memanfaatkan media digital dalam proses pembelajaran. Data training ini digunakan sebagai dasar pembentukan model klasifikasi untuk menemukan pola dan aturan yang memengaruhi pencapaian nilai siswa berdasarkan variabel yang diamati.

Tabel 3.5 Data Testing

Nama	Absensi (X1)	Kelas Tambahan (X2)	Media Digital (X3)	Nilai (Y)
S25	<7	Ya	Tidak	KKM
S26	≥7	Tidak	Tidak	KKM
S27	<7	Ya	Ya	Diatas KKM
S28	<7	Tidak	Tidak	KKM
S29	≥7	Ya	Ya	KKM
S30	<7	Tidak	Tidak	KKM

Tabel ini merupakan data testing yang terdiri dari 6 siswa (S25–S30) dan digunakan untuk menguji model klasifikasi yang telah dibangun sebelumnya. Variabel yang digunakan meliputi absensi (X1) dengan kategori <7 dan ≥7, kelas tambahan (X2) dengan kategori Ya dan Tidak, serta media digital (X3) dengan kategori Ya dan Tidak, sedangkan variabel hasil (Y) adalah nilai siswa yang diklasifikasikan menjadi KKM dan Di atas KKM. Data testing ini berfungsi untuk mengevaluasi kemampuan model dalam memprediksi hasil nilai siswa berdasarkan pola yang diperoleh dari data training.

Model Decision Tree C4.5

Tahap 1 Perhitungan Entropy, Information Gain, Split Information, dan Gain Ratio Secara Manual

Tahap ini merupakan tahap awal dalam algoritma Decision Tree C4.5 yang bertujuan untuk menentukan atribut terbaik dalam membentuk pohon keputusan. Entropy digunakan untuk mengukur tingkat ketidakpastian data, information gain untuk mengetahui pengaruh suatu atribut terhadap variabel target, split information untuk melihat pola pembagian data, dan gain ratio untuk memilih atribut yang paling optimal. Atribut dengan nilai gain ratio tertinggi akan digunakan pada tahap selanjutnya.

Menghitung Entropy Total (S)

rumus:

$$Entropy(S) = -\sum p_i \log_2(p_i)$$

Langkah:

Total data = 24

KKM = 17

Diatas KKM = 7

Langkah 1: Probabilitas

$$P(KKM) = \frac{17}{24} = 0{,}708$$

$$P(Diatas) = \frac{7}{24} = 0{,}292$$

Hitung:

$$Entropy(S) = -(p_1 log_2\, p_1 + p_2 log_2\, p_2)$$

$$Entropy(S) = -(0{,}708 log_2\, 0{,}708 + 0{,}292 log_2\, 0{,}292)$$

$$= -(0{,}708 \times -0{,}50 + 0{,}292 \times -1{,}78)$$

$$= -(-0{,}354 - 0{,}520)$$

$$= 0{,}874$$

Menghitung Entropy Absensi

Diketahui:

< 7 = 14 data

≥ 7 = 10 data

Entropy Absensi <7

KKM = 10

Diatas KKM = 4

Langkah 1 : Probabilitas

$$P(KKM) = \frac{10}{14} = 0{,}714$$

$$P(Diatas) = \frac{4}{14} = 0{,}286$$

Hitung

$$Entropy = -(0{,}714\log_2 0{,}714 + 0{,}286\log_2 0{,}286)$$

$$= -(0{,}714 \times -0{,}48 + 0{,}286 \times -1{,}81)$$

$$= -(-0{,}343 - 0{,}518) = 0{,}861$$

Entropy absensi ≥ 7

KKM = 7

Diatas KKM = 3

Langkah 1 : Probabilitas

$$P(KKM) = \frac{7}{10} = 0{,}7$$

$$P(Diatas) = \frac{3}{10} = 0{,}3$$

Hitung

Entropy =$-(0{,}7\log_2 0{,}7+0{,}3\log_2 0{,}3)=0{,}881$

Information Gain

$$Gain = 0{,}874 - \left(\frac{14}{24} \times 0{,}861 + \frac{10}{24} \times 0{,}881\right)$$

$$= 0{,}874 - (0{,}583 \times 0{,}861 + 0{,}417 \times 0{,}881)$$

$$= 0{,}874 - (0{,}502 + 0{,}367)$$

$$= 0{,}874 - 0{,}869 = 0{,}005$$

Split Info

$$SplitInfo = -(0{,}583\log_2 0{,}583 + 0{,}417\log_2 0{,}417)$$

$$= -(0{,}583 \times -0{,}78 + 0{,}417 \times -1{,}26)$$

$$= -(-0{,}455 - 0{,}526) = 0{,}981$$

Gain Ratio

$$GainRatio = \frac{0{,}005}{0{,}981} = 0{,}005$$

Menghitung Entropy Kelas Tambahan

Diketahui:

Ya = 16

Tidak = 8

Entropy Ya

KKM = 11

Diatas = 5

Langkah 1 : Probabilitas

$$P(KKM) = \frac{11}{16} = 0{,}688$$

$$P(Diatas) = \frac{5}{16} = 0{,}312$$

Hitung

Entropy =−(0,688$\log_2$0,688+0,312$\log_2$0,312) =0,896

Entropy Tidak

KKM = 6

Diatas = 2

Langkah 1 : Probabilitas

$$P(KKM) = \frac{6}{8} = 0{,}75$$

$$P(Diatas) = \frac{2}{8} = 0{,}25$$

Hitung

Entropy =−(0,75$\log_2$0,75+0,25$\log_2$0,25) =0,811

Information Gain

$$Gain = 0{,}874 - \left(\frac{16}{24} \times 0{,}896 + \frac{8}{24} \times 0{,}811\right)$$

$$= 0{,}874 - (0{,}667 \times 0{,}896 + 0{,}333 \times 0{,}811)$$

$$= 0{,}874 - (0{,}597 + 0{,}270)$$

$$= 0{,}874 - 0{,}867 = 0{,}007$$

Split Info

$$SplitInfo = -(0{,}667log_2\,0{,}667 + 0{,}333log_2\,0{,}333)$$

$$= 0{,}918$$

Gain Ratio

$$GainRatio = \frac{0{,}007}{0{,}918} = 0{,}008$$

Menghitung Entropy Media Digital

Diketahui:

Ya = 11

Tidak = 13

Entropy Ya

KKM = 8

Diatas = 3

Langkah 1 : Probabilitas

$$P(KKM) = \frac{8}{11} = 0{,}727$$

$$P(Diatas) = \frac{3}{11} = 0{,}273$$

Hitung

Entropy =−(0,727$\log_2$0,727+0,273$\log_2$0,273) =0,845

Entropy Tidak

KKM = 9

Diatas = 4

Langkah 1 : Probabilitas

$$P(KKM) = \frac{9}{13} = 0{,}692$$

$$P(Diatas) = \frac{4}{13} = 0{,}308$$

Hitung

$$Entropy = -(0{,}692\log_2 0{,}692 + 0{,}308\log_2 0{,}308) = 0{,}881$$

Information Gain

$$Gain = 0{,}874 - \left(\frac{11}{24} \times 0{,}845 + \frac{13}{24} \times 0{,}881\right)$$

$$= 0{,}874 - (0{,}458 \times 0{,}845 + 0{,}542 \times 0{,}881)$$

$$= 0{,}874 - (0{,}387 + 0{,}477)$$

$$= 0{,}874 - 0{,}864 = 0{,}010$$

Split Info

$$SplitInfo = -(0{,}458\log_2 0{,}458 + 0{,}542\log_2 0{,}542) = 0{,}995$$

Gain Ratio

$$GainRatio = \frac{0{,}010}{0{,}995} = 0{,}010$$

Tulisan ini merupakan hasil perhitungan yang dilakukan dengan menggunakan metode *Decision Tree*. Proses perhitungan manual ini dilakukan secara sistematis melalui tahapan pengukuran *Entropy*, *Information Gain*, *Split Information*, dan *Gain Ratio* pada setiap variabel yang digunakan. Hasil dari proses tersebut kemudian diolah untuk menemukan pola yang digunakan, sehingga mampu menghasilkan model yang terstruktur dan konsisten

Tahap 2 Perhitungan Entropy, Information Gain, Split Information, dan Gain Ratio Menggunakan Aplikasi Excel.

Pada tahap perhitungan selanjutnya, proses dilakukan dengan bantuan aplikasi Microsoft Excel untuk mempermudah pengolahan data. Penggunaan Excel membantu dalam melakukan perhitungan secara lebih cepat, sistematis, dan akurat melalui pemanfaatan rumus-rumus yang sesuai, sehingga dapat mengurangi kesalahan dalam perhitungan manual serta memudahkan proses

analisis data.

Tabel 3.6 Perhitungan Entropy, Information Gain, Split Information, dan Gain Ratio

		Jumlah	KKM	Diatas KKM	Entropy	Gain	Split Information	Gain Ratio
Total		24	17	7	0,870864			
Absensi						0,287531	0,979868757	0,293438416
	<7	14	7	7	1			
	≥7	10	10	0	0			
Kelas Tambahan						0,053387	0,918295834	0,058137023
	Ya	16	10	6	0,954434			
	Tidak	8	7	1	0,543564			
Media Digital						0,882144	0,036751814	0,041661908
	Ya	11	7	4	0,94566			
	Tidak	13	10	3	0,77935			

Tabel ini menunjukkan hasil perhitungan nilai entropy, gain, split information, dan gain ratio pada data training yang terdiri dari 24 data, dengan 17 data bernilai KKM dan 7 data Di atas KKM, sehingga diperoleh nilai entropy total sebesar 0,870864. Analisis dilakukan terhadap tiga atribut, yaitu Absensi, Kelas Tambahan, dan Media Digital, untuk menentukan atribut terbaik dalam pembentukan pohon keputusan menggunakan algoritma C4.5. Hasil perhitungan menunjukkan bahwa atribut Absensi memiliki nilai gain ratio tertinggi sebesar 0,2934 dibandingkan dengan Kelas Tambahan sebesar 0,0581 dan Media Digital sebesar 0,0417, sehingga atribut Absensi dipilih sebagai root node.

Hal ini mengindikasikan bahwa tingkat kehadiran siswa merupakan faktor yang paling berpengaruh dalam membedakan kategori nilai siswa pada data yang dianalisis.

Tahap 3 Pemilihan Atribut Terbaik (Root Node)

Pemilihan atribut terbaik sebagai root node dalam algoritma Decision Tree C4.5 dilakukan dengan membandingkan nilai gain ratio dari setiap atribut

Tabel 3.7 Root Node 1 Absensi

Node 1		Jumlah	KKM	Diatas KKM	Entropy	Gain	Split Information	Gain Ratio
Total		24	17	7	0,870864			
Absensi						0,2875531	0,979868757	0,293438416
	<7	14	7	7	1			
	≥7	10	10	0	0			
Kelas Tambahan						0,053387	0,918295834	0,058137023
	Ya	16	10	6	0,954434			
	Tidak	8	7	1	0,543564			
Media Digital						0,882144	0,036751814	0,041661908
	Ya	11	7	4	0,94566			
	Tidak	13	10	3	0,77935			

Pada Node 1 (Absensi), total terdapat 24 data, dengan 17 siswa memenuhi KKM dan 7 siswa berada di atas KKM. Atribut absensi memberikan nilai gain sebesar 0,2875 dan gain ratio 0,2934, sehingga cukup berpengaruh dalam pemisahan data.

Absensi kurang dari 7 memiliki entropy tertinggi, sedangkan absensi ≥7 menunjukkan kondisi paling homogen karena seluruh data berada pada satu kelas hasil.

Tabel 3.8 Node 1.1 Media Digital

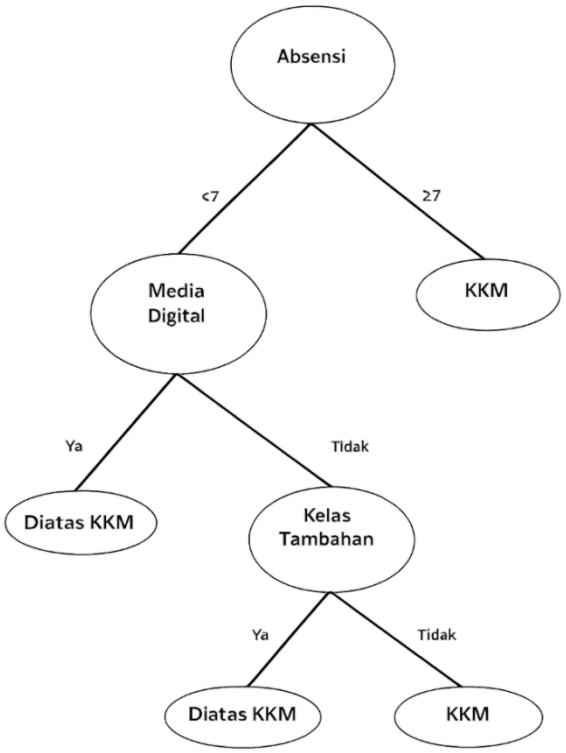
Node 1.1		Jumlah	KKM	Diatas KKM	Entropy	Gain	Split Information	Gain Ratio
Total		14	7	7	1			
Kelas Tambahan						0,07467	0,863120569	0,086511797
	Ya	10	4	6	0,970951			
	Tidak	4	3	1	0,811278			
Media Digital						0,667498	0,120733204	0,18087414
	Ya	5	1	4	0,721928			
	Tidak	9	6	3	0,918296			

Tabel Node 1.1 ini menunjukkan hasil perhitungan entropy, gain, split information, dan gain ratio pada subset data berjumlah 14 data, yang terdiri dari 7 data bernilai KKM dan 7 data Di atas KKM, sehingga diperoleh nilai entropy total sebesar 1. Perhitungan dilakukan terhadap atribut Kelas Tambahan dan Media Digital untuk menentukan atribut terbaik pada node lanjutan dalam algoritma C4.5. Hasil perhitungan menunjukkan bahwa atribut Media Digital memiliki nilai gain ratio tertinggi sebesar 0,1809 dibandingkan dengan Kelas Tambahan sebesar 0,0865, sehingga Media Digital dipilih sebagai node berikutnya pada Node 1.1. Hal ini menunjukkan bahwa penggunaan media digital memiliki pengaruh yang lebih besar dalam membedakan hasil nilai siswa pada subset data tersebut.

Tahap 4 Pembentukan pohon keputusan

Pembentukan pohon keputusan dilakukan secara bertahap dengan membagi data berdasarkan nilai dari atribut root node yang telah dipilih. Setiap pembagian akan menghasilkan cabang-cabang baru yang mewakili nilai atribut tersebut.

Pada setiap node selanjutnya, dilakukan kembali perhitungan entropy, information gain, split information, dan gain ratio untuk menentukan atribut terbaik berikutnya. Proses ini dilakukan secara berulang hingga memenuhi kondisi penghentian, yaitu ketika seluruh data dalam satu node berada pada kelas yang sama atau tidak terdapat atribut lain yang dapat digunakan untuk pemisahan.



Gambar 3.2 Pohon Keputusan

Gambar tersebut menunjukkan pohon keputusan (decision tree) hasil penerapan algoritma C4.5 untuk mengklasifikasikan hasil nilai siswa, yaitu KKM dan Di atas KKM, berdasarkan beberapa atribut. Atribut Absensi menjadi root node karena memiliki nilai gain ratio tertinggi. Jika absensi siswa ≥ 7, maka hasil nilai langsung diklasifikasikan sebagai KKM. Sebaliknya, jika absensi < 7, proses klasifikasi dilanjutkan ke atribut Media Digital. Pada kondisi ini, siswa yang menggunakan media digital (Ya) akan diklasifikasikan Di atas KKM. Namun, jika tidak menggunakan media digital, maka klasifikasi dilanjutkan ke atribut Kelas Tambahan, di mana siswa yang mengikuti kelas tambahan (Ya) dikategorikan Di atas KKM, sedangkan yang tidak mengikuti kelas tambahan (Tidak) diklasifikasikan sebagai KKM. Pohon keputusan ini menggambarkan aturan klasifikasi yang jelas dan sistematis berdasarkan pola data training yang telah dianalisis.

3.5.3 Metrik Evaluasi

Metrik evaluasi adalah ukuran untuk menilai kinerja model dalam memprediksi data. Metrik ini menunjukkan keakuratan, ketepatan, dan kemampuan model menangkap kelas yang sebenarnya. Pada model klasifikasi, metrik umum meliputi accuracy, precision, recall, F1-score, serta confusion matrix untuk melihat prediksi benar dan salah. Metrik evaluasi membantu menilai efektivitas model dan menjadi dasar perbaikan untuk meningkatkan kinerjanya.

1. Confusion Matrix

Confusion matrix adalah tabel yang menunjukkan perbandingan antara nilai

aktual (sebenarnya) dan prediksi model. Tabel ini membantu menilai kinerja model, termasuk melihat berapa prediksi yang benar (TP) dan berapa prediksi yang salah (FP atau FN) untuk setiap kelas.

Tabel 3.9 Hasil prediksi data testing

Nama	**Nilai Aktual**	**Nilai Prediksi**	**Keterangan**
S25	KKM	Di atas KKM	Salah
S26	KKM	KKM	Benar
S27	Di atas KKM	Di atas KKM	Benar
S28	KKM	KKM	Benar
S29	KKM	KKM	Benar
S30	KKM	KKM	Benar

Tabel ini menunjukkan hasil evaluasi prediksi pohon keputusan C4.5 menggunakan data testing. Dari enam data siswa, lima data berhasil diprediksi dengan benar dan satu data mengalami kesalahan prediksi. Kesalahan terjadi pada S25, sedangkan data lainnya sesuai antara nilai aktual dan nilai prediksi. Hasil ini menunjukkan kinerja model cukup baik.

Tabel 3.10 Confusion Matrix

	Aktual Di atas KKM	**Aktual KKM**
Prediksi Di atas KKM	1 (TP)	1 (FP)
Prediksi KKM	0 (FN)	4 (TN)

Tabel confusion matrix ini menunjukkan hasil evaluasi model klasifikasi. Terdapat 1 data True Positive (TP)**,** yaitu prediksi Di atas KKM yang sesuai dengan nilai aktual, serta 4 data True Negative (TN)**,** yaitu prediksi KKM yang benar. Selain itu, terdapat 1 False Positive (FP) dan tidak ditemukan False Negative (FN)**,** yang menunjukkan model cukup akurat.

2. Accuracy

Accuracy adalah ukuran yang menunjukkan seberapa tepat atau benar prediksi model secara keseluruhan.

Rumusnya :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total data}}$$

Contoh :

$$\text{Accuracy} = \frac{1+4}{6} = \frac{5}{6} = 0{,}83\ (83\%)$$

Hasil perhitungan menunjukkan bahwa nilai akurasi model adalah 0,83 atau 83%. Nilai ini diperoleh dari perbandingan antara jumlah data yang berhasil diklasifikasikan dengan benar (sebanyak 5 data) terhadap total seluruh data yang diuji (sebanyak 6 data), sehingga dapat disimpulkan bahwa model memiliki tingkat ketepatan yang cukup baik dalam melakukan klasifikasi.

3. Precision

Precision adalah ukuran yang menunjukkan seberapa tepat prediksi model untuk suatu kelas tertentu. Dengan kata lain, precision menghitung proporsi prediksi yang benar dari semua data yang diprediksi sebagai kelas tersebut.

Rumusnya :

$$\text{Precisioan} = \frac{\text{True Positive (TP)}}{\textit{True Positive (TP)} + \textit{False Positive (FP)}}$$

Contoh :

$$\text{Precisioan} = \frac{1}{1+1} = 0{,}5\ (50\%)$$

Hasil perhitungan menunjukkan bahwa nilai precision adalah 0,5 atau 50%. Nilai ini diperoleh dari perbandingan antara jumlah data yang diprediksi positif dengan benar (sebanyak 1 data) terhadap seluruh data yang diprediksi positif (sebanyak 2 data). Artinya, dari 2 data yang diprediksi sebagai positif, hanya 1 data yang benar-benar sesuai, hal ini menunjukkan bahwa ketepatan prediksi positif model masih tergolong rendah dan perlu dilakukan perbaikan agar hasil prediksi menjadi lebih akurat

4. Recall

Recall adalah ukuran yang menunjukkan seberapa baik model dapat menangkap semua data yang sebenarnya termasuk dalam suatu kelas tertentu. Dengan kata lain, recall mengukur proporsi data aktual dari suatu kelas yang berhasil diprediksi dengan benar oleh model.

Rumus :

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\textit{True Positive (TP)} + \textit{False Negative (FN)}}$$

Contoh :

$$\text{Recall} = \frac{1}{1 + 0} = 1\ (100\%)$$

Hasil perhitungan menunjukkan bahwa nilai recall adalah 1 atau 100%. Nilai ini diperoleh dari perbandingan antara jumlah data positif yang berhasil diprediksi dengan benar (sebanyak 1 data) terhadap seluruh data positif yang sebenarnya ada (sebanyak 1 data). Artinya, model berhasil mengenali seluruh data positif tanpa ada yang terlewat, sehingga kemampuan model dalam menangkap data positif tergolong sangat baik.

5. F1-score

F1-Score adalah ukuran kinerja model yang menggabungkan precision dan recall menjadi satu angka. F1-Score digunakan untuk menilai keseimbangan antara ketepatan prediksi (precision) dan kemampuan menangkap semua anggota kelas (recall).

Rumus :

$$\text{F1} = 2\,x\,\frac{\text{Precision x Recall}}{Precision\ +\ Recall}$$

Contoh :

$$\text{F1} = 2\,x\,\frac{0{,}5 \text{ x } 1}{0{,}5 + 1} = 0{,}67\ (67\%)$$

Hasil perhitungan menunjukkan bahwa nilai F1-Score adalah 0,67 atau 67%. Nilai ini diperoleh dari penggabungan antara precision sebesar 50% dan recall sebesar 100%, sehingga F1-Score menggambarkan keseimbangan antara ketepatan dan kelengkapan model dalam melakukan klasifikasi. Nilai ini menunjukkan bahwa meskipun model sangat baik dalam menangkap seluruh data positif (recall tinggi), ketepatan prediksi positifnya masih perlu ditingkatkan (precision rendah), sehingga secara keseluruhan performa model berada pada kategori cukup baik