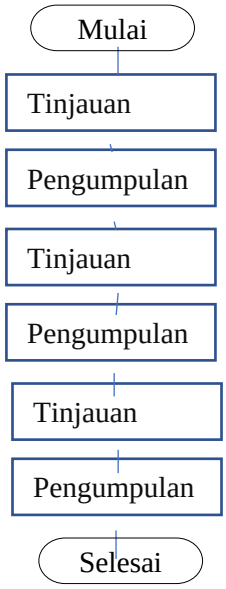
BAB III
METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif. Tempat Penelitian ini adalah SMK SMK Negeri 1 Penawartama Tulang Bawang.

Prosedur penelitian menggunakan tahapan-tahapan KDD (Knowledge Data Discovery). Tahapan-tahapannya adalah: Pre-processing/Cleaning, sebelum proses data mining.

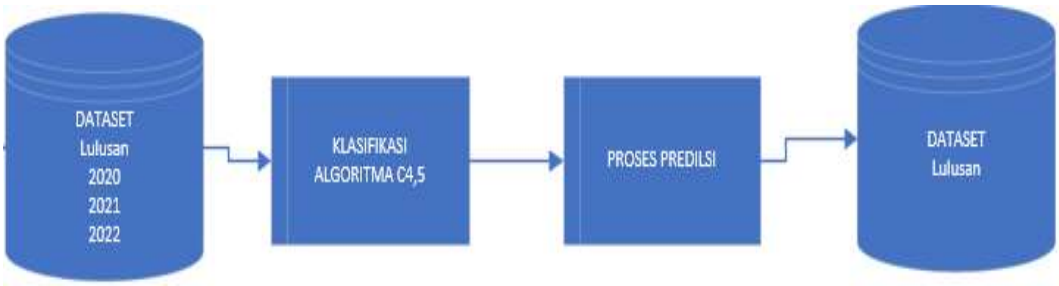
Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi); (3) Transformation, coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining; (4) Analisis data; (5) Pattern Evaluation, merupakan tahapan evaluasi untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan sumber data yang ada. Kerangka pikir yang digunakan dalam penelitian ini dapat dilihat pada gambar 3.1

Tahapan Penelitian dapat dijelaskan pada Gambar 3.1 sebagai berikut :



Gambar 3.1 Tahapan Penelitian

3.1 Tahapan Pengolahan Data



Gambar 3.2. Tahapan Pengolahan Data

3.2 Teknik Analisis Data

Penelitian ini mengunakan teknik Decision Tree, CHAID dan regresi ganda untuk melakukan prediksi prestasi belajar siswa SMK 1 Penawartama Tulang Bawang.berdasarkan status ekonomi orang tua, motivasi, kedisiplinan dan prestasi masa lalu.

3.3 Decision Tree

Decision Tree akan memperlihatkan faktor-faktor kemungkinan (probabilitas) yang akan mempengaruhi alternatif-alternatif prestasi belajar siswa, disertai dengan prediksi hasil akhir yang akan didapat bila faktor-faktor dalam Decision Tree terpenuhi. Decision Tree akan mengubah data kedalam bentuk visual berupa diagram pohon dan aturan-aturan keputusan. Data dalam Decision Tree dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Salah satu atribut yang merupakan atribut yang menyatakan data solusi per-item data yang disebut dengan target atribut. Atribut memiliki nilai-nilai yang dinamakan dengan instance. Alur proses analisis dalam decision tree adalah mengubah bentuk data (table) menjadi model tree, mengubah model tree menjadi rule dan menyederhanakan rule (pruning). Data yang diambil dalam penelitian ini adalah populasi sejumlah 416 siswa akan digunakan untuk membuat model prediksi Decision Tree. Model yang telah dibuat kemudian akan dihitung tingkat akurasi prediksinya.

3.4 CHAID

Tujuan dari metode ini adalah untuk memisahkan data secara berurutan dengan pembagian biner menjadi beberapa subgrup. Pada tiap tahap, pembagian sebuah grup menjadi dua bagian didefinisikan oleh salah satu variabel prediktor, sebuah himpunan bagian dari kategori-kategorinya mendefinisikan salah satu bagian, dan sisa kategori lainnya mendefinisikan bagian yang lain. Pada AID, prediktornya memiliki dua tipe utama, yaitu monotonik dan bebas[13]. Alur proses analisis data dengan CHAID adalah memeriksa tiap variabel independen menggunakan uji chi-square, menentukan variabel independen mana yang paling signifikan, membagi data menggunakan kategori variabel independen tersebut dengan peringkat yang paling signifikan, mengulangi langkah ke-4 untuk semua subgrup sampai teridentifikasi semua pembagian yang secara statistik telah signifikan.

3.5 Regresi

Linier adalah metode statistika yang digunakan untuk membentuk model hubungan antara variabel terikat (dependen) dengan satu atau lebih variabel bebas (independen). Apabila banyaknya variabel bebas hanya ada satu, disebut sebagai regresi linier sederhana, sedangkan apabila terdapat lebih dari 1 variabel[14]

analisis regresi setidak-tidaknya memiliki 3 kegunaan, yaitu untuk tujuan deskripsi dari fenomena data atau kasus yang sedang diteliti, untuk tujuan kontrol, serta untuk tujuan prediksi. Regresi mampu mendeskripsikan fenomena data melalui terbentuknya suatu model hubungan yang bersifatnya numerik. Regresi juga dapat digunakan untuk melakukan pengendalian (kontrol) terhadap suatu kasus atau hal-hal yang sedang diamati melalui penggunaan model regresi yang diperoleh. Selain itu, model regresi juga dapat dimanfaatkan untuk melakukan prediksi untuk variabel terikat. Namun yang perlu diingat, prediksi di dalam konsep regresi hanya boleh dilakukan di dalam rentang data dari variabel-variabel bebas yang digunakan untuk membentuk model regresi tersebut.[15]

3.6 Algoritma C4.5

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Keterangan: Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti berikut:

$$Gain\ S, A = Entropy\ S - Si\ S\ n\ i{=}1 * Entropy\ A$$

S: himpunan kasus

A: atribut

N: jumlah partisi atribut

A |Si |: jumlah kasus pada partisi ke-i

|S|: jumlah kasus dalam S

Sebelum mendapatkan nilai Gain adalah dengan mencari nilai Entropy. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan sebuah atribut. Rumus dasar dari Entropy adalah sebagai berikut:

$Entropy\ S = -pi * log2p$

Keterangan:

S: himpunan kasus

A: fitur

n: jumlah partisi

S Pi: proporsi dari Si terhadap S

Contoh perhitungan

Dari kutipan di atas Perhitungan C4.5 dimulai dengan mencari nilai entropy dari semua data. Nilai entropy digunakan sebagai dasar perhitungan gain tiap atribut, nilai gain tertinggi menjadi akar dari pohon keputusan perhitungan dilakukan hingga semua atribut terdefinisi.

3.7 Split Info

Split Info merupakan rumus yang menyatakan informasi potensial atau entropy. dapat dilihat dalam rumus (3). Dan keterangannya:

Split Info (S, A) = -$\sum Si\ S\ log2\ Si\ S\ n\ i$=1 .....(3)

S: Himpunan kasus

A: Atribut

Si: Jumlah kasus pada partisi ke- i

3.8 Gain Ratio

Gain Ratio adalah modifikasi dari information gain yang digunakan untuk mengurangi bias atribut yang memiliki banyak cabang. Gain ratio memiliki sifat:

Bernilai besar jika data menyebar rata

➢ Bernilai kecil jika semua data masuk ke dalam satu cabang GainRatio

(S, A) = $Gai\ (S, A)\ SpiltInfo\ (S, A)$ .....(4) dan Keterangannya sebagai berikut:

S: Himpunan kasus

A: Atribut Gain

(S, A): Information gain pada atribut A

SpiltInfo (S, A): SplitInfo pada atribut A

3.9 Pengujian Algoritma Decision Tree C4.5

Pengujian data yang digunakan dalam algoritma C4.5 adalah dataset prediksi bantuanbeasiswa yang didapatkan dari data 3 tahun sebelumnya. Data yang digunakan dalam pengujian ini dapat dilihat pada gambar berikut.

Row No.	Prestasi	No	Nis	Status Ekon...	Kedisiplinan	Presensi
1	tidak ada	1	181444023	mampu	kurang baik	baik sekali
2	ada	2	181444028	mampu	baik	baik sekali
3	tidak ada	3	181444035	mampu	kurang baik	baik sekali
4	tidak ada	4	181444037	mampu	kurang baik	baik sekali
5	tidak ada	5	181444040	mampu	kurang baik	baik sekali
6	tidak ada	6	181444043	mampu	kurang baik	baik sekali
7	tidak ada	7	181444048	mampu	kurang baik	baik sekali
8	tidak ada	8	181444050	mampu	kurang baik	baik sekali
9	tidak ada	9	181444066	mampu	kurang baik	baik sekali
10	tidak ada	10	181444069	mampu	kurang baik	baik sekali
11	tidak ada	11	181444084	mampu	kurang baik	baik sekali
12	tidak ada	12	18144088	mampu	kurang baik	baik sekali
13	tidak ada	13	181444089	mampu	kurang baik	baik sekali
14	tidak ada	14	181444093	mampu	kurang baik	baik sekali
15	tidak ada	15	181444100	mampu	kurang baik	baik sekali

Gambar 3.3. Dataset Prediksi beasiswa

Table 3.1 perhitungan Entropi

		Jumlah (s)	A (Si)	B (Si)	Entrophy	Gain	Split Info	Gain Ratio
Total		606	44	562	0,375578962			
Status Ekonomi						0,00682399	0.23645260385	-0,02885986404
	Mampu	65	10	55	0,619382195			
	Tidak Mampu	541	34	507	0,338642642			
Kedisiplinan						0,15032371		
	Baik Sekali	14	11	3	0,749595257			
	Baik	24	17	7	0,870864469			
	Kurang Baik	567	16	551	0,185378491			
Presensi						0,00276641		
	Baik Sekali	388	27	361	0,364382875			
	Baik	205	17	188	0,412409015			

Perhitungan algoritma C4.5

Entropy Total $\left(\frac{606}{44} x \log 2 \frac{606}{44}\right) + \frac{562}{562} x \log 2 \frac{562}{44}$=0,375578962

Status Ekonomi Mampu

$= -\left(\frac{65}{10} x \log 2 \frac{65}{10}\right) + \frac{55}{65} x \log 2 \frac{65}{10}$=0,619382195

Status Ekonomi Tidak Mampu

$= -\left(\frac{541}{34} x \log 2 \frac{541}{34}\right) + \frac{507}{34} x \log 2 \frac{65}{10}$=0,338642642

Perhitungan Gain

$= -\left(\frac{65}{606}\right) x$ 0,619382195 + ($\left(\frac{541}{606}\right) x$ 0,338642642= 0,00682399

Perhitungan Sinfo

$= -\left(\frac{65}{606} x \log 2 \frac{65}{606}\right) - \frac{541}{606} x \log 2 \frac{541}{606}$ = -0.23645260385

Perhitungan Gratio

$= \frac{0,00682399}{-0.23645260385}$= -0,02885986404

Semua atribut/ fitur didalam data set dihitung nilai entropy, gain, split info dan gain ratio menggunakan tools rapidminer yang ada dipenjelasan