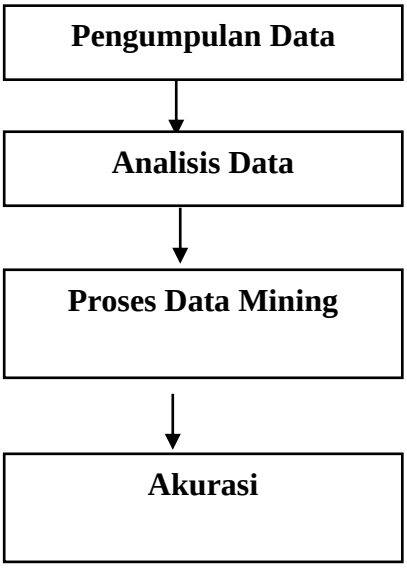
BAB III

METODOLOGI PENELITIAN

Penelitian ini akan dilaksanakan berdasarkan alur penelitian seperti yang ditunjukkan pada Gambar 3.1.



Gambar 3.1 Alur Penelitian

3.1 Tempat dan Waktu Penelitian

1. Tempat

   Penelitian akan dilakukan di Kampus Sekolah Tinggi Manajemen Informatika Dan Komputer (STMIK Pringsewu)

2. Waktu

   Dalam melaksanakan tahapan penelitian, peneliti merencakan waktu penelitian dari bulan september 2019sampai januari2020.

3.2 Alat dan Bahan

Adapun alat dan bahan yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Hardware

Kebutuhan perangkat keras (hardware) yang digunakan : Laptop Samsung RV511 Processor : Intel® CORE ™ i3 CPU M380 @2.53 GHz (4 CPUs) ~ 2,5 GHz.RAM : 3072 MB. harddisk dengan kapasitas 500 GB.

2. Software

Kebutuhan perangkat lunak (software) yang digunakan :

a. Sistem Operasi Windows 07
b. Aplikasi Matlab R2014 .
c. Google Earth .

3. Data

Data yang akan digunakan dalam penelitian ini adalah data citra satelit yang didapatkan dari aplikasi Google Earth pada citra kabupaten Pringsewu Lampung Indonesia.

3.3 Teknik Pengambilan Data

Didalam penelitian ini peneliti menggunakan beberapa metode yang akan digunakan untuk melakukan penelitian yang berkaitan dengan pengumpulan data Berikut adalah beberapa metode yang digunakan;

1. Studi Lapangan (*Field Research*)

   Studi lapangan merupakan metode pengumpulan data untuk memperoleh data dan informasi dengan mengadakan pengamatan secara langsung. Adapun teknik Pengumpulan data dan informasi yang dilakukan pada saat studi lapangan pada titik daratan yang akan dianalisa adalahPengamatan Langsung (*Observation*) yaitu Pengumpulan data yang dilakukan penulis pada saat pengamatan langsung pada data mahasiswa STMIK Pringsewu.

2. Tinjauan Pustaka (*Research Library*)

   Tinjauan pustaka dilakukan dengan cara membaca, mengutip dan membuat catatan yang bersumber pada bahan-bahan pustaka yang mendukung dan berkaitan dengan penelitian dalam hal ini mengenai data mining C4.5, Naïve Bayes, K-NN.

3.4. Tahapan Penelitian

Penelitian ini menggunakan model standarisasi data mining yaitu Cross Industry Standart Process for Data Mining (CRISP-DM) yang digambarkan pada skema tahap penelitian di atas dengan langkah-langkah sebagai berikut:

1. Fase Pemahaman Bisnis (*Bussiness Understanding Phase*)

Pada tahap ini berfokus pada tujuan penelitian yaitu untuk mengetahui algoritma terbaik dalam Waktu Mahasiswa Mendapatkan Pekerjaan, dengan menerjemahkan data mahasiswa , sehingga didapatkan model yang terbaik untuk memenuhi dari tujuan penelitian.

2. Fase Pemahaman Data (*Data Understanding Phase*)

Data diperoleh dari Kampus STMIK Pringsewu, atribut yang disajikan pada Table 3.1.

Tabel 3.1 Keterangan Kriteria

No	Kriteria	Keterangan
1	Jurusan	Jurusan Mahasiswa
2	Keanggotaan Organisasi Kampus	Organisasi yang diikuti oleh mahasiswa
3	IPK	IPK Mahasiswa
4	Prestasi Selama Perkuliahan	Prestasi Yang didapatkan oleh mahasiswa
5	Lama Study	Lama study yang di tempuh oleh mahasiswa

3. Data intergration dan data transformasi

Untuk meningkatkan dan memudahkan dalam proses analisis maka dari 5 Kriteria yang diperoleh akan dipilih beberapa kriteria inti yang digunakan.

Tabel 3.2 Keterangan Kriteria

No	Kriteria	Skala	Penjelasan
1	Jurusan	Binomial	- 1=Sistem Informasi (SI) - 2=Manajemen Informatika (MI)
2	Keanggotaan Organisasi Kampus	Binomial	- 1= Mengikuti UKM - 2= Tidak Mengikuti UKM
3	IPK	Polinomial	- 1=<=3.00(Cukup) - 2= >3.00&<3.50(Memuaskan) - 3= >=3.50 (Dengan Pujian)
4	Prestasi Selama Perkuliahan	Binomial	- 1=Berprestasi - 2=Tidak Berprestasi
5	Lama Study	Binomial	- 1=Tepat Waktu - 2=Lambat

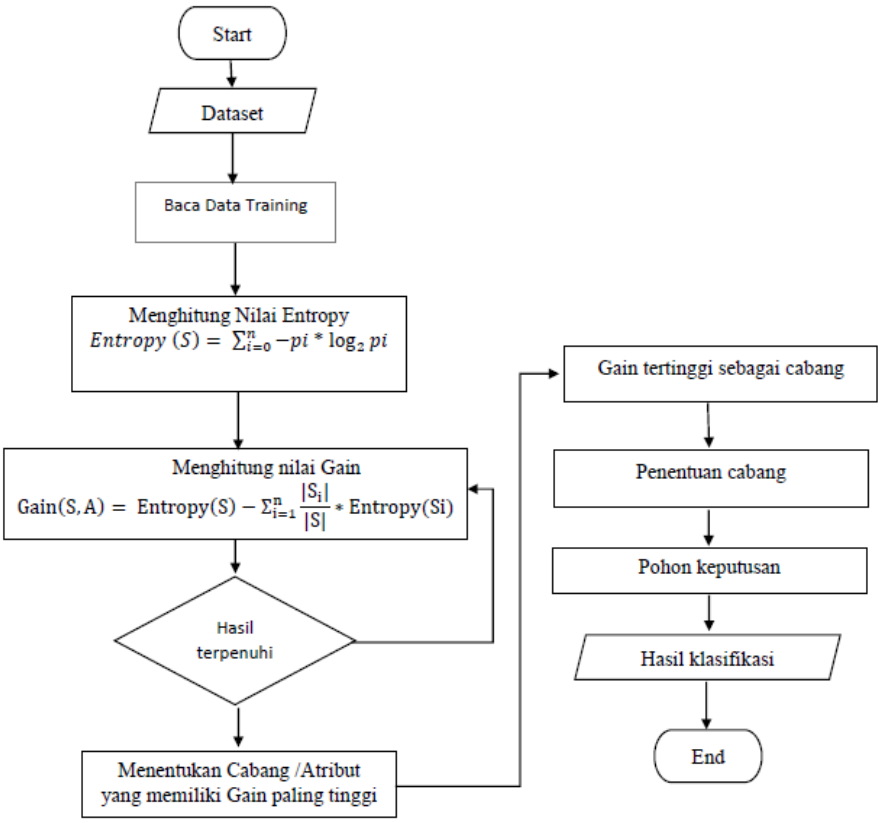
4. Pemodelan (Modeling Phase)

Algoritma yang digunakan dalam penelitian ini yaitu algoritma C4.5, Naïve Bayes dan K-Nearest Neighbor (KNN), untuk mengklasifikasikan dalam memperikrakanmasa tunggu mahasiswa dalam mendapatkan pekerjaan dan untuk mengukur akurasi ketiga algoritma menggunakan metode confusion matrix yang hasil akhir menampilkan hitungan dari rata-rata presentasi akurasi, recall dan error

rate. Setelah diketahui hasil dari akurasi tiap algoritma kemudian dibandingkan untuk mencari algoitma terbaik

a. Flowchart Algoritma C4.5

Dibawah ini akan dijelaskan flowchart atau urutan algorithma C4.5 .



Gambar 3.1 Flowchart Algoritma C4.5

Algoritma C4.5 memiliki proses alur sebagai berikut :

1. Mempersiapkan data *training*
2. Menentukan nilai entropy dengan menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing variabel. Sebelumnya dihitung terlebih dahulu nilai *entropy*, dengan persamaan 3.1

$$Entropy(S) = \sum_{i=0}^{n} -p\,i_{*} \qquad \log_2 pi$$

(3.1)

Dimana :

S = Himpunan kasus

n = Jumlah kasus pada partisi S

pi = Proporsi Si terhadap S

3. Menghitung nilai gain dengan persamaan 3.2

$$\text{Gain(S,A)} = \text{Entropy(S)} - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy(Si)}$$

(3.2)

Dimana :

S : Himpunan Kasus

A : Atribut

n : jumlah partisi atribut A

|*Si*| *:* jumlah kasus pada partisi ke –i

|*S*| *:* jumlah kasus dalam S

4. Mengulangi langkah ke-2 hingga semua record terpartisi. Proses partisi pohon akan berhenti disaat semua kasus pada cabang memiliki kelas yang sama dan tidak ada variabel dalam record yang dipartisi lagi.
5. Gain yang tertinggi akan dijadikan sebagai cabang untuk menentukan cabang pohon keputusan.
6. Jika sudah mendapatkan cabang pohon keputusan maka proses selanjutnya membuat pohon keputusan.

Tabel 3.1 Keputusan Lama Waktu Tunggu Mendapat pekerjaan

Mahasiswa	Jurusan	UKM	IPK	Prestasi	Lama Study	LAMA MENDAPAT PEKERJAAN
1	SI	Mengikuti UKM	CUKUP	Berprestasi	Tepat Waktu	Cepat
2	SI	Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Lambat	Lambat
3	MI	Mengikuti UKM	DENGAN PUJIAN	Tidak Berprestasi	Tepat Waktu	Lambat
4	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
5	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Cepat
6	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
7	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Lambat	Lambat
8	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
9	MI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
10	SI	Mengikuti UKM	MEMUASKAN	Berprestasi	Tepat Waktu	Cepat
11	MI	Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Cepat
12	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
13	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
14	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Cepat
15	SI	Tidak Mengikuti UKM	DENGAN PUJIAN	Tidak Berprestasi	Tepat Waktu	Lambat
16	SI	Mengikuti UKM	DENGAN PUJIAN	Berprestasi	Tepat Waktu	Lambat
17	MI	Mengikuti UKM	CUKUP	Berprestasi	Lambat	Cepat
18	MI	Mengikuti UKM	CUKUP	Berprestasi	Tepat Waktu	Cepat
19	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
20	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Cepat

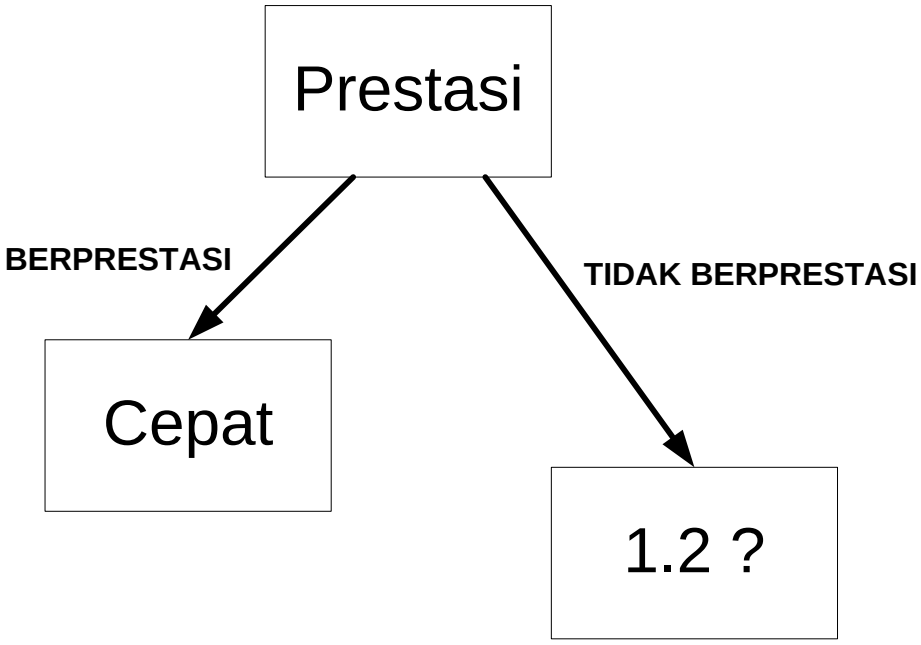
Untuk memilih atribut sebagai node, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Maka dilakukan perhitungan gain dan entropi sesuai dengan rumus yang berlaku bagi algoritma decision tree C4.5

Tabel 3.2 Perhitungan node 1

Node			jml(S)	Lambat(s1)	Cepat(s2)	Entropy	Gain
1	**Total**		20	11	9	0.9928	
	Jurusan						**0,0418**
		SI	12	8	4	0,9183	
		MI	8	4	4	1	
	IPK						**0,0102**

		Cukup	10	5	5	1	
		Memuaskan	7	4	3	0,9852	
		Dengan Pujian	3	2	1	.9183	
	UKM						**0,1815**
		Ikut	8	2	6	0,8113	
		Tidak Ikut	12	9	3	0,8113	
	Prestasi					0	**0,3438**
		Berprestasi	5	0	5	0	
		Tidak Berprestasi	15	11	4	0,8113	
	Lama Study					0	**0,0072**
		Tepat Waktu	17	9	8	0,9975	
		Lambat	3	2	1	0,9183	

Selanjutnya dilakukan perhitungan untuk entropi dan gain pada masing-masing atribut. Setelah perhitungan selesai dilakukan akan diketahui nilai gain yang terbesar yaitu pada atribut **Prestasi**dengan nilai **0,3438**maka atribut **0,3438**dapat dijadikan sebagai *root*. Pohon keputusan yang terbentuk dapat digambarkan sebagai berikut.



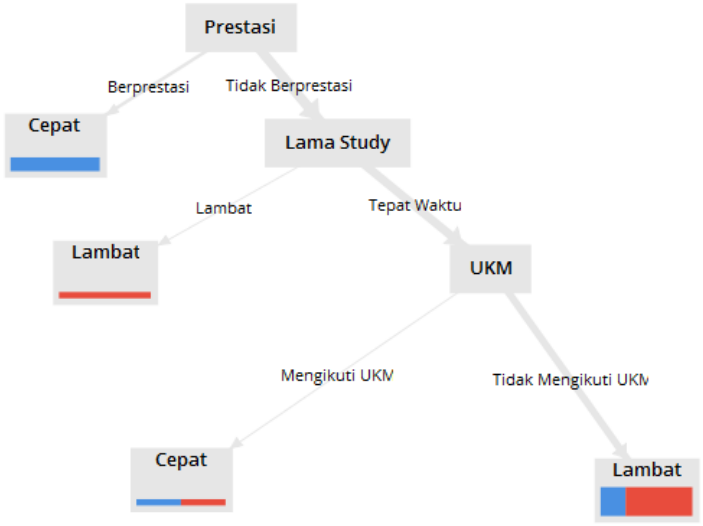
Gambar 3.2 Node 1.1

Pada node 1.2 masih belum diketahui , sehingga untuk nilai atribut tidakberprestasi harus dilakukan perhitungan lagi

Tabel 3.3Perhitungan node 2

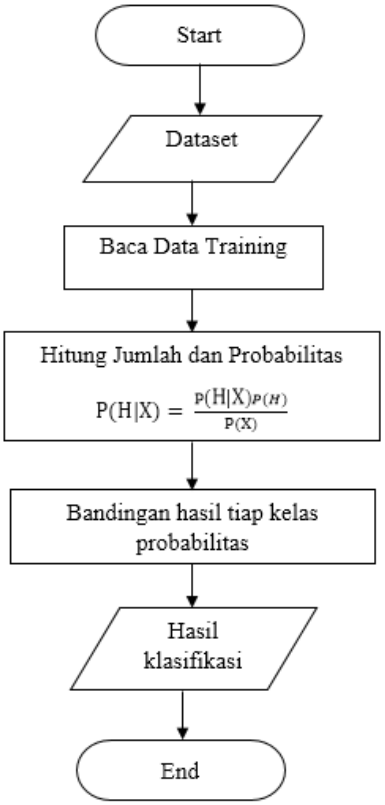
Node			jml(S)	Lambat(s1)	Cepat(s2)	Entropy	Gain
1.2	**Berprestasi**		15	11	4	0,8366	
	Jurusan						**0,0108**
		SI	9	7	2	0,7642	
		MI	6	4	2	0,9183	
	IPK						**0,0665**
		Cukup	7	5	4	0,8631	
		Memuaskan	6	4	2	0,9183	
		Dengan Pujian	2	2	0	0	
	UKM						**0,0040**
		Ikut	3	2	1	0,9183	
		Tidak Ikut	12	9	3	0,8113	
	Lama Study						**0,0649**
		Tepat Waktu	13	9	4	0,8905	
		Lambat	2	2	0	0	

Selanjutnya dilakukan perhitungan untuk entropi dan gain pada masing-masing atribut. Setelah perhitungan selesai dilakukan akan diketahui nilai gain yang terbesar yaitu pada atribut **Lama Study**dengan nilai **0,3438**maka atribut **0,0649**dapat dijadikan sebagai *root*. Proses ini terus berlanjut sehingga menghasilkan Pohon keputusan yang terbentuk dapat digambarkan sebagai berikut.



Gambar 3.2Pohon Keputusan

b. Flowchart Naïve Bayes (NB)



Gambar 3.2 Flowchart Algoritma Naïve Bayes

Algoritma Naïve Bayes memiliki proses alur sebagai berikut :

1. Mempersiapkan data *training*
2. Lalu melakukan perhitungan jumlah dan probabilitas algoritma Naïve Bayes dengan rumus 3.3 $P(H|X) = \frac{P(H|X)P(H)}{P(X)}$
3. Membandingkan hasil probabilitas dari tiap kelas probabilitas
4. Setelah selesai di bandingkan maka tahap selanjutnya hitung hasil klasifikasi
5. Tahap akhir hitung nilai akurasi algoritma naïve bayes dan tahap selesai.

Tabel 3.4 Data Training

Mahasiswa	Jurusan	UKM	IPK	Prestasi	Lama Study	LAMA MENDAPAT PEKERJAAN
1	SI	Mengikuti UKM	CUKUP	Berprestasi	Tepat Waktu	Cepat
2	SI	Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Lambat	Lambat
3	MI	Mengikuti UKM	DENGAN PUJIAN	Tidak Berprestasi	Tepat Waktu	Lambat
4	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
5	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Cepat
6	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
7	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Lambat	Lambat
8	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
9	MI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
10	SI	Mengikuti UKM	MEMUASKAN	Berprestasi	Tepat Waktu	Cepat
11	MI	Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Cepat
12	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Lambat
13	SI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
14	SI	Tidak Mengikuti UKM	MEMUASKAN	Tidak Berprestasi	Tepat Waktu	Cepat
15	SI	Tidak Mengikuti UKM	DENGAN PUJIAN	Tidak Berprestasi	Tepat Waktu	Lambat
16	SI	Mengikuti UKM	DENGAN PUJIAN	Berprestasi	Tepat Waktu	Lambat
17	MI	Mengikuti UKM	CUKUP	Berprestasi	Lambat	Cepat
18	MI	Mengikuti UKM	CUKUP	Berprestasi	Tepat Waktu	Cepat
19	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Lambat
20	MI	Tidak Mengikuti UKM	CUKUP	Tidak Berprestasi	Tepat Waktu	Cepat

Setelah didapatkan data training kemudian dilanjutkan dengan menghitung probabilitas

- Probabilitas Lambat : 12/20 = 0,6
- Probabilitas Cepat : 8/20 = 0,4

Membandingkan setiap Probabilitas dengan tiap kelas Probabilitas

Tabel 3.5 Probabilitas

Case 1	Jurusan SI dan Lama mendapat Pekerjaan Cepat	5/12	0,4167
Case 2	Jurusan SI dan Lama mendapat Pekerjaan Lambat	7/12	0,563
Case 3	Jurusan MI dan Lama mendapat Pekerjaan Cepat	4/8	0,5

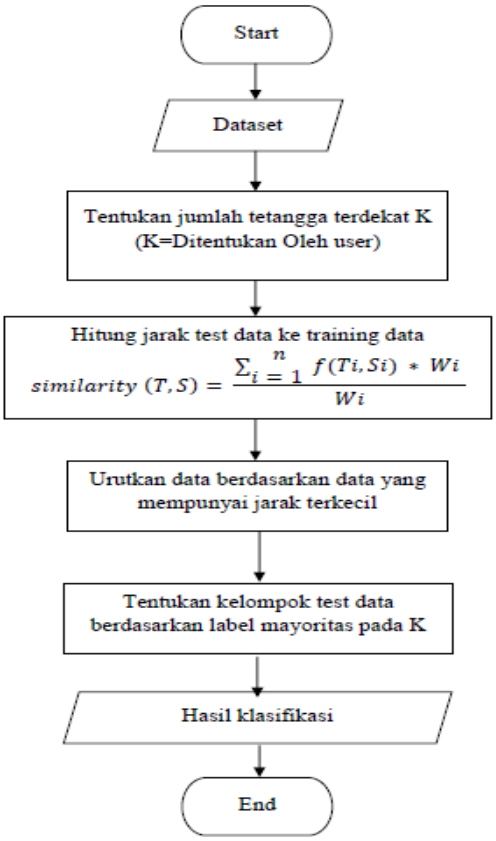
Case 4	Jurusan MI dan Lama mendapat Pekerjaan Lambat	4/8	0,5
Case 5	Mengikuti UKM dan Lama mendapat Pekerjaan Cepat	6/8	0,75
Case 6	Mengikuti UKM dan Lama mendapat Pekerjaan Lambat	2/8	0,25
Case 7	Tidak Mengikuti UKM dan Lama mendapat Pekerjaan Cepat	3/12	0,25
Case 8	Tidak Mengikuti UKM dan Lama mendapat Pekerjaan Lambat	8/12	0,75
Case 9	IPK Cukup dan Lama mendapat Pekerjaan Cepat	5/10	0,5
Case 10	IPK Cukup dan Lama mendapat Pekerjaan Lambat	5/10	0,5
Case 11	IPK Memuaskan dan Lama mendapat	3/7	0,43

	Pekerjaan Cepat		
Case 12	IPK Memuaskan dan Lama mendapat Pekerjaan Lambat	4/7	0,57
Case 13	IPK Dengan Pujian dan Lama mendapat Pekerjaan Cepat	1/3	0,33
Case 14	IPK Dengan Pujian dan Lama mendapat Pekerjaan Lambat	2/3	0,67
Case 15	Berprestasi dan Lama mendapat Pekerjaan Cepat	5/5	1
Case 16	Berprestasi dan Lama mendapat Pekerjaan Lambat	5/0	0
Case 17	Tidak Berprestasi dan Lama mendapat Pekerjaan Cepat	4/15	0,27
Case 18	Tidak Berprestasi dan Lama mendapat Pekerjaan Lambat	11/15	0,73

Case 19	Tepat Waktu dan Lama mendapat Pekerjaan Cepat	8/17	0,47
Case 20	Tepat Waktu dan Lama mendapat Pekerjaan Lambat	9/17	0,53
Case 21	Tidak Tepat Waktu dan Lama mendapat Pekerjaan Cepat	1/3	0,33
Case 22	Tidak Tepat Waktu dan Lama mendapat Pekerjaan Cepat	2/3	0,67

- Selanjutnya mengalikan semua probabilitas cepat dan lambat
- Membandingakan Probabilitas Cepat dan Lambat Jika Cepat > Lambat maka = Cepat

C. Flowchart K-Nearest Neighbors (K-NN)



Gambar 3.3 Flow Chart Algoritma K-Nearest Neighbors

Penjelasan Flowchart diatas sebagai berikut :

1. Menentukan jumlah tetangga terdekat K ditentukan oleh user
2. Selanjuntnya hitung dengan menggunakan rumus 3.4

$$similarity\ (T,S) = \frac{\sum_{i=1}^{n} f(Ti,Si) * Wi}{Wi} \quad (3.4)$$

3. Setelah proses perhitungan telah selesai lalu urutkan data berdasarkan data yang mempunyai jarak terkecil.
4. Menentukan kelompok test data berdasarkan label mayoritas pada k

Jumlah Tetangga ditentukan sebanyak 5

-data akan ditransofrmasi menjadi numerik

Tabel 3.6Transformasi Data Training

Jurusan	UKM	IPK	Prestasi	Lama Study	Lama Mendapat Pekerjaan
1	1	1	1	1	Cepat
1	1	2	2	2	Lambat
2	1	3	2	1	Lambat
2	2	1	2	1	Lambat
1	2	1	2	1	Cepat
1	2	1	2	1	Lambat
1	2	1	2	2	Lambat
1	2	2	2	1	Lambat
2	2	2	2	1	Lambat
1	1	2	1	1	Cepat
2	1	2	2	1	Cepat
1	2	2	2	1	Lambat
1	2	1	2	1	Lambat
1	2	2	2	1	Cepat
1	2	3	2	1	Lambat
1	1	3	1	1	Cepat
2	1	1	1	2	Cepat

2	1	1	1	1	Cepat
2	2	1	2	1	Lambat
2	2	1	2	1	Cepat

Dengan data testing

Tabel 3.7Transformasi Data Testing

1	2	2	1	2	?

Yang berasal dari transformasi data

Tabel 3.8 Data Testing

SI	Tidak Mengikuti UKM	MEMUASKAN	Berprestasi	Lambat	?

- Kemudian ditentunkan Euclidean Distance nya..
- Lalu label dari jumlah tetangga sesuai dengan dominan nya label dari tetangga tersebut

5. Fase Evaluasi (*Evaluation Phase*)

Pada tahap ini dilakukan evaluasi kinerja dari algoritma C4.5, Naïve Bayes dan K-Nearest Neighbor dengan membandingkan hasil nilai rata-rata akurasi, recall dan error rate yang terdapat pada tabel confusion matrix. Beberapa persyaratan standar yang telah ditetapkan untuk matriks klasifikasi dua kelas:

a. Akurasi (AC), adalah proporsi jumlah prediksi yang benar. Hal ini ditentukan dengan menggunakan persamaan 3.5.

$$AC = \frac{a+d}{a+b+c+d}$$

(3.5)

b. *Recall* atau *true positive rate* (TP), adalah proporsi kasus positif yang diidentifikasi dengan benar, yang dihitung dengan menggunakan persamaan 3.6.

$$TP = \frac{d}{c+d} \quad (3.6)$$

c. *False positive rate* (FP), adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dapat dihitung dengan menggunakan persamaan 3.7.

$$FP = \frac{b}{a+b} \quad (3.7)$$

d. *True negative rate* (TN), didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan 3.8.

$$TN = \frac{a}{a+b} \quad (3.8)$$

e. False negative rate (FN) adalah proporsi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitung dengan menggunakan persamaan 3.9.

$$FN = \frac{c}{c+d} \quad (3.9)$$

f. *Precision* (P) adalah proporsi kasus positif diprediksi dengan benar, yang dihitung dengan menggunakan persamaan 3.10.

$$FN = \frac{d}{b+d}$$

(3.10)