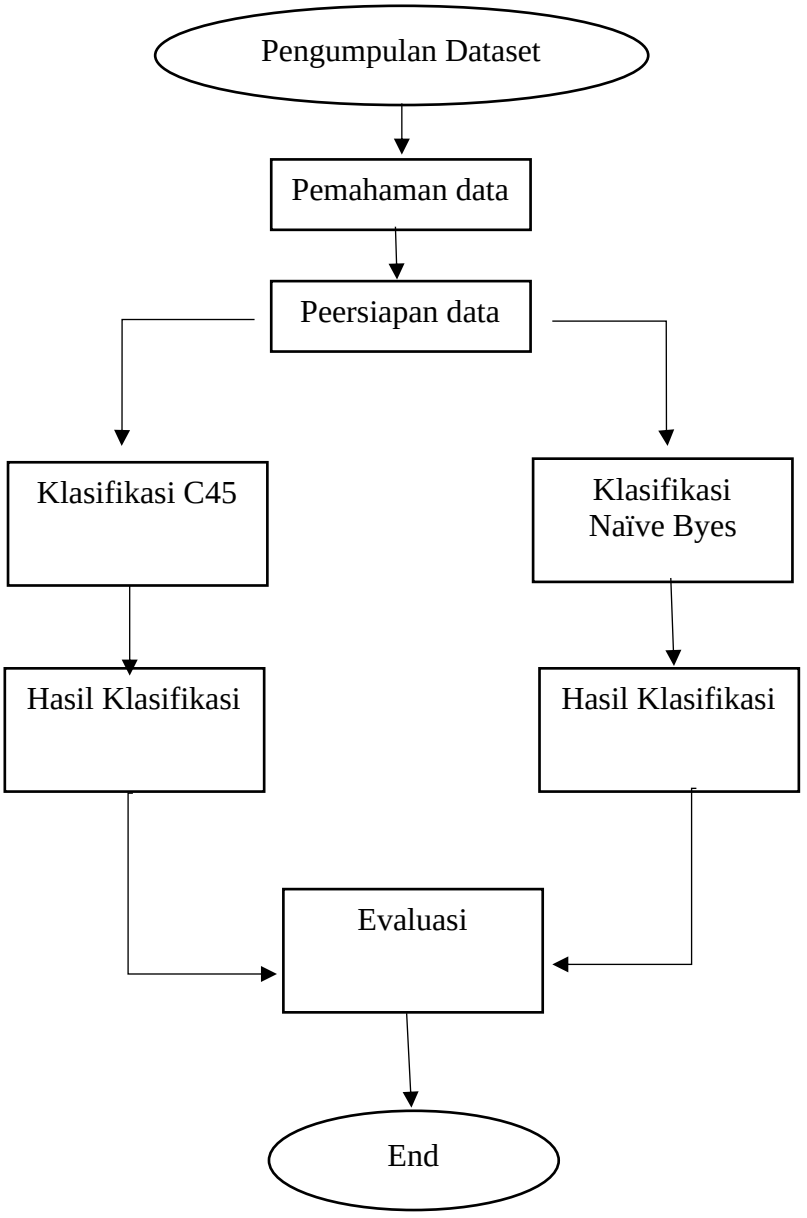
BAB III

METODOLOGI PENELITIAN

3.1 Metode Penelitian

Dalam melakukan analisa dan mencari pola data untuk dijadikan sebuah dataset dalam memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan maka dibuat alur dalam tahapan penelitian yang akan dilakukan sebagai berikut :



3.1.1 Tahap Pengumpulan Dataset

Pada tahap ini dilakukan proses pengumpilam dataset yang relevan serta atribut yang saling berkaitan dengan Kesehatan ibu hamil untuk nantinya dapat di proses dalam Data Mining melalui Website yang menyediakan Dataset secara publik. Dalam hal ini pengumpulan data dilakukan melalui Website, *https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set* dengan jumlah data sampling sebesar 1014 data record

3.1.2 Tahap Pemahaman data

Pada tahapan dilakukan dengan cara membuat rincian data sehingga dapat memudahkan dalam memahami data.

Tabel 3. 1 Attribut Dataset Kesehatan Ibu Hamil

Nama Attribut	Keterangan
Usia	20-45 tahun
SystolicBP (Nilai atas Tekanan Darah dalam mmHg)	70-160 mmHg
DiastolicBP (Nilai Tekanan Darah yang lebih rendah dalam mmHg)	49-100 mmHg
BS (Kadar glukosa darah dalam hal konsentrasi molar, mmol/L.)	5-20 mmol

BodyTemp (Suhu Badan)	95-105 drajat
HeartRate (Denyut jantung istirahat normal dalam denyut per menit.)	60-90
RiskLevel (Prediksi Tingkat Intensitas Risiko selama kehamilan dengan mempertimbangkan atribut sebelumnya.)	Class 1 (Low Risk) Class 2 (Mid Risk) Class 3 (High Risk)

3.1.3 Tahap Persiapan data

Pada tahapan dilakukan Persiapan Data seperti menghapus Data Duplikat, mencari Data yang kosong atau tidak sesuai sebelum masuk ke dalam tahapan klasifikasi Algoritma.

3.1.4 Klasifikasi menggunakan Algoritma C.45

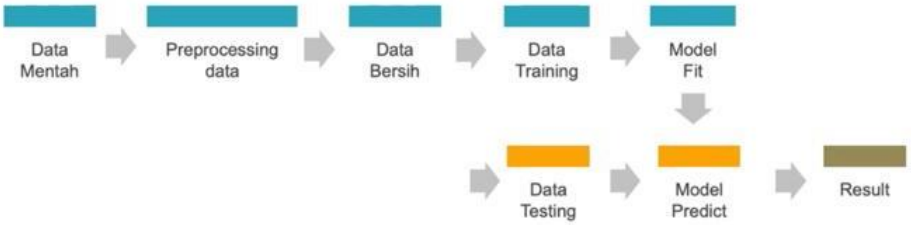
Pada tahapan ini dilakukan Pengklasifikasian Menggunakan Algoritma C4.5

3.1.4.1 Tahap Pemodelan Data

Pada tahap ini terdapat beberapa proses dalam pemodelan data diantaranya sebagai berikut.

3.1.4.2 Merancang Design Model Analisis

Pada tahap ini dilakukan proses merancang model analisis yang akan digunakan pada tools Rapid Miner.

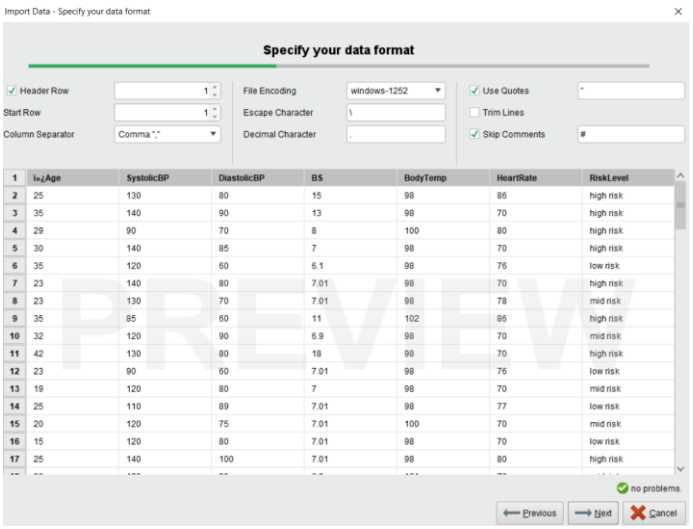


Gambar 3. 1 Rancang Model Analisis

3.1.4.2.1 Data Mentah

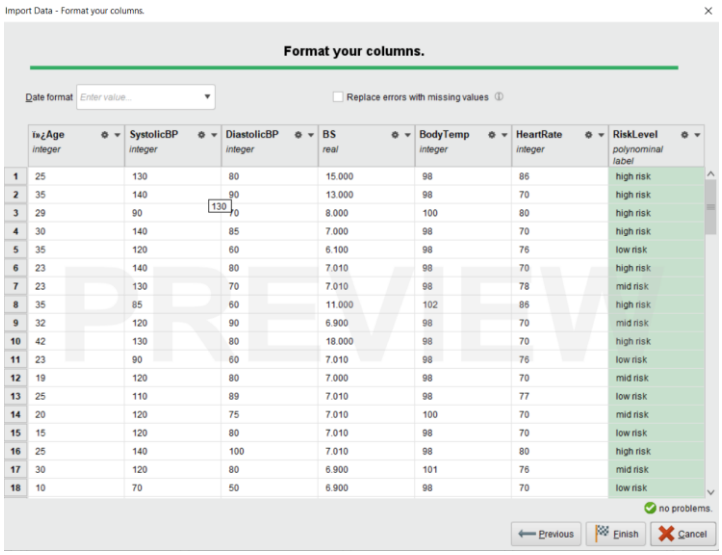
Data mentah yang telah dipersiapkan kemudian di import menggunakan tools Rapid miner dengan menggunakan operator Read CSV kemudian dilakukan pemberian label pada data atrribut yang akan dijadikan label kelas.

Gambar 3. 2 Operator Read CSV



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
2	25	130	80	15	98	86	high risk
3	35	140	90	13	98	70	high risk
4	29	90	70	8	100	80	high risk
5	30	140	85	7	98	70	high risk
6	35	120	60	6.1	98	76	low risk
7	23	140	80	7.01	98	70	high risk
8	23	130	70	7.01	98	78	mid risk
9	35	85	60	11	102	86	high risk
10	32	120	90	6.9	98	70	mid risk
11	42	130	80	18	98	70	high risk
12	23	90	60	7.01	98	76	low risk
13	19	120	80	7	98	70	mid risk
14	25	110	89	7.01	98	77	low risk
15	20	120	75	7.01	100	70	mid risk
16	15	120	80	7.01	98	70	low risk
17	25	140	100	7.01	98	80	high risk

Gambar 3. 3 Data mentah yang telah di import kedalam Tools Rapid Miner



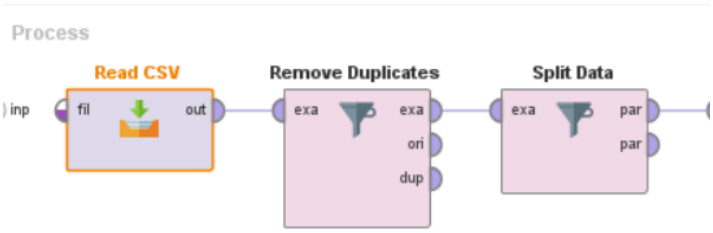
Gambar 3. 4 Data mentah yang telah diberikan label pada kolom hijau

Table 3. 1 Type Data Attribut

N0	Attribute	Tipe
1	Usia	Integer
2	SystolicBP (Nilai atas Tekanan Darah dalam mmHg)	Integer
3	DiastolicBP (Nilai Tekanan Darah yang lebih rendah dalam mmHg)	Integer
4	BS (Kadar glukosa darah dalam hal konsentrasi molar, mmol/L.)	Real
5	BodyTemp (Suhu Badan)	Integer
6	HeartRate (Denyut jantung istirahat normal dalam denyut per menit.)	Integer
7	RiskLevel (Prediksi Tingkat Intensitas Risiko selama kehamilan dengan mempertimbangkan atribut sebelumnya.)	Polynominal

3.1.4.2.2 Data Training

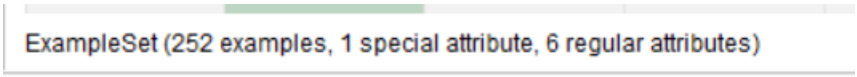
Tahapan selanjutnya adalah melakukan *split data*, operator ini digunakan untuk membagi jumlah data training dan data testing sebesar 70:30.



Gambar 3. 5 Operator Split Data

Maka akan didapatkan hasil sebagai berikut :

Gambar 3. 6 Data Training



Gambar 3. 7 Data Testing

Data Training : 70% x 840 (*record* dataset)
Data Testing : 30% x 840 (*record* dataset)

3.1.4.2.3 Model Fit

Pada tahap ini adalah menentukan model Algoritma yang akan digunakan, Untuk penelitian ini menggunakan Algoritma Decision Tree sehingga menghasilkan pohon keputusan. Adapun tahapanya sebagai berikut :

3.1.4.2.3.1 Menentukan Nilai Entrophy Dan Gain (Rumus)

1. Nilai Entrophy Total

Perhitungan entropy untuk semua data terhadap komposisi kelas, dimana diketahui sebagai berikut :

Tabel 3. 2 Entrophy Total

Attribut	Kriteria	Jumlah Data	Obesity			Entrophy
			Level 1	Level II	Level III	Nilai
RiskLevel	Label	840	359	287	194	0,973

$$E\ (semua(Total))$$

$$= -\left(\left(\frac{359}{840}\right) x \log 3 \left(\frac{359}{840}\right)\right) + \left(\left(\frac{287}{840}\right) x \log 3 \left(\frac{287}{840}\right)\right)$$

$$+ \left(\left(\frac{194}{840}\right) x \log 3 \left(\frac{194}{840}\right)\right) = 0{,}973$$

2. Nilai Entrophy Dan Gain Atrribut

2.1 Atribut Age (kontinyu).

dimana diketahui data sebagai berikut :

Tabel 3. 3 Attribut Age

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
Age	<=25	506	266	170	70
	>25	334	93	117	124

Total		840	359	287	194

$$E\ (semua(<= 25))$$

$$= -\left(\left(\frac{266}{506}\right) x \log 3 \left(\frac{266}{506}\right)\right) + \left(\left(\frac{170}{506}\right) x \log 3 \left(\frac{170}{506}\right)\right)$$

$$+ \left(\left(\frac{70}{506}\right) x \log 3 \left(\frac{70}{506}\right)\right) = 0.890$$

$$E\ (semua(> 25))$$

$$= -\left(\left(\frac{93}{334}\right) x \log 3 \left(\frac{93}{334}\right)\right) + \left(\left(\frac{117}{334}\right) x \log 3 \left(\frac{117}{334}\right)\right)$$

$$+ \left(\left(\frac{124}{334}\right) x \log 3 \left(\frac{124}{334}\right)\right) = 0.993$$

$$Gain\ (semua(Age)) = -\left(\left(\frac{454}{840}\right) x 0.890\right) + \left(\left(\frac{356}{840}\right) x 0.993\right) = 0.041$$

$$Split\ Info\ (Age) = -\left(\left(\frac{506}{840}\right) x \log 2 \left(\frac{506}{840}\right)\right) + \left(\left(\frac{334}{840}\right) x \log 2 \left(\frac{334}{840}\right)\right)$$

$$= 0.970$$

$$Gain\ Rasio\ (Age) = \frac{0.041}{0.970} = 0.043$$

2.2 Atribut SystolicBP (kontinyu).

dimana diketahui data sebagai berikut :

Tabel 3. 4 Attribut SystolicBP

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
SystolicBP	<=111	336	178	97	61
	>111	504	181	190	133

$$E\ (semua(<= 111))$$

$$= -\left(\left(\frac{178}{336}\right) x \log 3 \left(\frac{178}{336}\right)\right) + \left(\left(\frac{97}{336}\right) x \log 3 \left(\frac{97}{336}\right)\right)$$

$$+ \left(\left(\frac{61}{336}\right) x \log 3 \left(\frac{61}{336}\right)\right) = 0.915$$

$$E\ (semua(> 111))$$

$$= -\left(\left(\frac{181}{504}\right) x \log 3 \left(\frac{181}{504}\right)\right) + \left(\left(\frac{190}{504}\right) x \log 3 \left(\frac{190}{504}\right)\right)$$

$$+ \left(\left(\frac{133}{504}\right) x \log 3 \left(\frac{133}{504}\right)\right) = 0.990$$

$$Gain\ (semua(SystolicBP)) = -\left(\left(\frac{336}{840}\right) x 0.915\right) + \left(\left(\frac{504}{840}\right) x 0.990\right) = 0.013$$

$$Split\ Info\ (SystolicBP) = -\left(\left(\frac{336}{840}\right) x \log 2 \left(\frac{336}{840}\right)\right) + \left(\left(\frac{504}{840}\right) x \log 2 \left(\frac{504}{840}\right)\right)$$

$$= 0.971$$

$$Gain\ Rasio\ (SystolicBP) = \frac{0.013}{0.971} = \ 0.014$$

2.3 Atribut DiastolicBP (kontinyu).

dimana diketahui data sebagai berikut :

Tabel 3. 5 Attribut DiastolicBP

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
DiastolicBP	<=75	431	196	162	73
	>75	409	163	125	121
Total					

$$E\ (semua(<= 75))$$

$$= -\left(\left(\frac{196}{431}\right) x \log 3 \left(\frac{196}{431}\right)\right) + \left(\left(\frac{162}{431}\right) x \log 3 \left(\frac{162}{431}\right)\right)$$

$$+ \left(\left(\frac{73}{431}\right) x \log 3 \left(\frac{73}{431}\right)\right) = 0.935$$

$$E\ (semua(> 75))$$

$$= -\left(\left(\frac{163}{409}\right) x \log 3 \left(\frac{163}{409}\right)\right) + \left(\left(\frac{125}{409}\right) x \log 3 \left(\frac{125}{409}\right)\right)$$

$$+ \left(\left(\frac{121}{409}\right) x \log 3 \left(\frac{121}{409}\right)\right) = 0.991$$

$$Gain\ (semua(\text{DiastolicBP})) = -\left(\left(\frac{431}{840}\right) x 0.935\right) + \left(\left(\frac{409}{840}\right) x 0.991\right) = 0.010$$

$$Split\ Info\ (\text{DiastolicBP}) = -\left(\left(\frac{431}{840}\right) x \log 2 \left(\frac{431}{840}\right)\right) + \left(\left(\frac{409}{840}\right) x \log 2 \left(\frac{409}{840}\right)\right)$$

$$= 1.000$$

$$Gain\ Rasio\ (\text{DiastolicBP}) = \frac{0.010}{1.000} = 0.010$$

2.4 Atribut BS (kontinyu).

dimana diketahui data sebagai berikut :

Table 3. 6 Attribut BS

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
BS	<=8	700	359	259	82
	>8	140	0	28	112
Total					

$$E\ (semua(<= 8))$$

$$= -\left(\left(\frac{359}{700}\right) x \log 3 \left(\frac{359}{700}\right)\right) + \left(\left(\frac{259}{700}\right) x \log 3 \left(\frac{259}{700}\right)\right)$$

$$+ \left(\left(\frac{82}{700}\right) x \log 3 \left(\frac{82}{700}\right)\right) = 0.875$$

$$E\ (semua(> 8))$$

$$= -\left(\left(\frac{0}{140}\right) x \log 3 \left(\frac{0}{140}\right)\right) + \left(\left(\frac{28}{140}\right) x \log 3 \left(\frac{28}{140}\right)\right)$$

$$+ \left(\left(\frac{112}{140}\right) x \log 3 \left(\frac{112}{140}\right)\right) = 0.455$$

$$Gain\ (semua(\text{BS})) = -\left(\left(\frac{700}{840}\right) x 0.875\right) + \left(\left(\frac{140}{840}\right) x 0.455\right) = 0.168$$

$$Split\ Info\ (\text{BS}) = -\left(\left(\frac{700}{840}\right) x \log 2 \left(\frac{700}{840}\right)\right) + \left(\left(\frac{140}{840}\right) x \log 2 \left(\frac{140}{840}\right)\right) = 0.650$$

$$Gain\ Rasio\ (\text{BS}) = \frac{0.168}{0.650} = \ 0.258$$

2.5 Atribut BodyTemp (kontinyu).

dimana diketahui data sebagai berikut :

Table 3. 7 Attribut BodyTemp

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
BodyTemp	<=99	652	323	203	126
	>99	188	36	84	68
Total					

$$E\ (semua(<= 99))$$

$$= -\left(\left(\frac{323}{652}\right) x \log 3 \left(\frac{323}{652}\right)\right) + \left(\left(\frac{203}{652}\right) x \log 3 \left(\frac{203}{652}\right)\right)$$

$$+ \left(\left(\frac{126}{652}\right) x \log 3 \left(\frac{126}{652}\right)\right) = 0.937$$

$$E\ (semua(> 99))$$

$$= -\left(\left(\frac{36}{188}\right) x \log 3 \left(\frac{36}{188}\right)\right) + \left(\left(\frac{84}{188}\right) x \log 3 \left(\frac{84}{188}\right)\right)$$

$$+ \left(\left(\frac{68}{188}\right) x \log 3 \left(\frac{68}{188}\right)\right) = 0.951$$

$$Gain\ (semua(\text{BodyTemp})) = -\left(\left(\frac{652}{840}\right) x 0.937\right) + \left(\left(\frac{188}{840}\right) x 0.951\right) = 0.033$$

$$Split\ Info\ (\text{BodyTemp}) = -\left(\left(\frac{652}{840}\right) x \log 2 \left(\frac{652}{840}\right)\right) + \left(\left(\frac{188}{840}\right) x \log 2 \left(\frac{188}{840}\right)\right)$$

$$= 0.767$$

$$Gain\ Rasio\ (\text{BodyTemp}) = \frac{0.033}{0.767} = 0.043$$

2.6 Atribut HeartRate (kontinyu).

dimana diketahui data sebagai berikut :

Table 3. 8 Attribut HeartRate

Attribut	Kriteria	Jumlah Data	Risk Level		
			Low Risk	Mid Risk	High RIsk
HeartRate	<=74	386	191	127	68
	>74	454	168	160	126
Total					

$$E\ (semua(<= 74))$$

$$= -\left(\left(\frac{191}{386}\right) x \log 3 \left(\frac{191}{386}\right)\right) + \left(\left(\frac{127}{386}\right) x \log 3 \left(\frac{127}{386}\right)\right)$$

$$+ \left(\left(\frac{68}{386}\right) x \log 3 \left(\frac{68}{386}\right)\right) = 0.928$$

$$E\ (semua(> 74))$$

$$= -\left(\left(\frac{168}{454}\right) x \log 3 \left(\frac{168}{454}\right)\right) + \left(\left(\frac{160}{454}\right) x \log 3 \left(\frac{160}{454}\right)\right)$$

$$+ \left(\left(\frac{126}{454}\right) x \log 3 \left(\frac{126}{454}\right)\right) = 0.993$$

$$Gain\ (semua(\text{HeartRate})) = -\left(\left(\frac{386}{840}\right) x 0.928\right) + \left(\left(\frac{454}{840}\right) x 0.993\right) = 0.009$$

$$Split\ Info\ (\text{HeartRate}) = -\left(\left(\frac{386}{840}\right) x \log 2 \left(\frac{386}{840}\right)\right) + \left(\left(\frac{454}{840}\right) x \log 2 \left(\frac{454}{840}\right)\right)$$

$$= 0.995$$

$$Gain\ Rasio\ (\text{HeaertRate}) = \frac{0.009}{0.995} = 0.009$$

3. Hasil Entrophy Attribut Dan Gain dari Entrophy Total

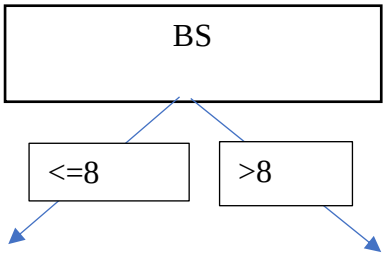
Table 3. 9 Hasil Entrophy Attribut Dan Gain dari Entrophy Total

No	Atribut	Value	Entrophy	Gain	Split Info	Gain rasio
1	Age	<=25	0.890	0.041	0.970	0.043
		>25	0.993			
2	SystolicBP	<=111	0.915	0.013	0.971	0.014
		>111	0.990			
3	DiastolicBP	<=75	0.935	0.010	1.000	0.010

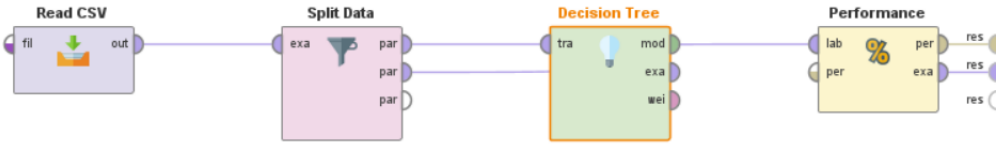
		>75	0.991			
4	BS	<=8	0.875	0.168	0.650	0.258
		>8	0.455			
5	BodyTemp	<=99	0.937	0.033	0.767	0.043
		>99	0.951			
6	Heartrate	<=74	0.928	0.009	0.995	0.009
		>74	0.993			

3.1.4.2.3.2 Menentukan Node Root Keputusan

Berdasarkan hasil dari perhitungan Manual untuk mencari gain tertinggi dari setiap attribut maka dihasilkan Node root Sebagai berikut :

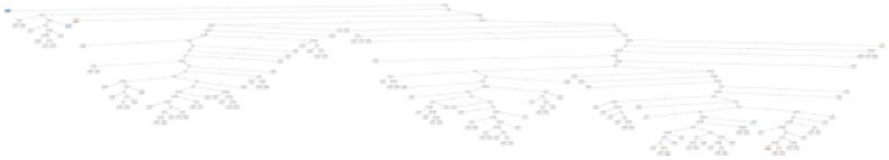


3.1.4.2.3.3 Menentukan Model Algoritma



Gambar 3. 8 Model Algoritma Decision Tree

Berikut ini hasil Pohon Keputusan dari *Operator Decision Tree* dengan *Node Root* sebagai Blood Glucose.



Gambar 3. 9 Pohon Keputusan dari Operator Decision Tree

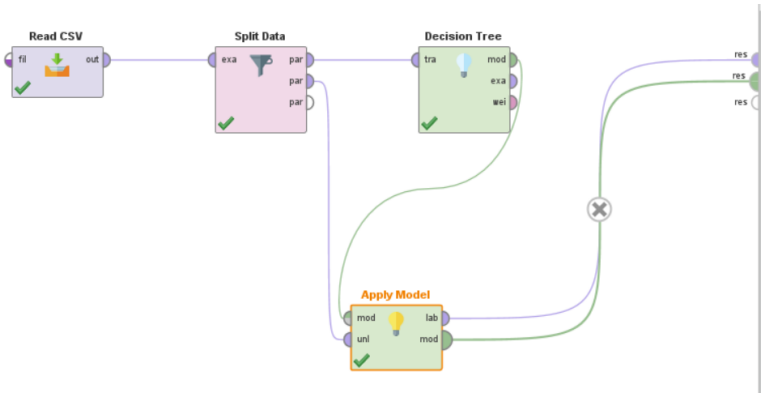
Tree

```
SystolicBP > 132.500: high risk (high risk=64, low risk=0, mid risk=0)
SystolicBP ≤ 132.500
|   BS > 9.500
|   |   i»¿Age > 43.500
|   |   |   i»¿Age > 44.500: high risk (high risk=1, low risk=0, mid risk=0)
|   |   |   i»¿Age ≤ 44.500: mid risk (high risk=0, low risk=0, mid risk=3)
|   |   i»¿Age ≤ 43.500
|   |   |   i»¿Age > 38.500
|   |   |   |   i»¿Age > 42.500: high risk (high risk=1, low risk=0, mid risk=0)
|   |   |   |   i»¿Age ≤ 42.500
|   |   |   |   |   HeartRate > 65
|   |   |   |   |   |   DiastolicBP > 92.500: high risk (high risk=4, low risk=0, mid risk=2)
|   |   |   |   |   |   DiastolicBP ≤ 92.500
|   |   |   |   |   |   |   i»¿Age > 41: high risk (high risk=2, low risk=0, mid risk=1)
|   |   |   |   |   |   |   i»¿Age ≤ 41: high risk (high risk=2, low risk=0, mid risk=0)
|   |   |   |   |   HeartRate ≤ 65: high risk (high risk=1, low risk=0, mid risk=0)
|   |   |   i»¿Age ≤ 38.500: high risk (high risk=25, low risk=0, mid risk=0)
|   BS ≤ 9.500
|   |   SystolicBP > 125: mid risk (high risk=0, low risk=0, mid risk=22)
|   |   SystolicBP ≤ 125
|   |   |   BodyTemp > 99.500
|   |   |   |   SystolicBP > 76.500
|   |   |   |   |   SystolicBP > 84
|   |   |   |   |   |   HeartRate > 66.500
|   |   |   |   |   |   |   i»¿Age > 30.500: mid risk (high risk=0, low risk=0, mid risk=8)
|   |   |   |   |   |   |   i»¿Age ≤ 30.500
|   |   |   |   |   |   |   |   BS > 6.050
|   |   |   |   |   |   |   |   |   BS > 6.500
|   |   |   |   |   |   |   |   |   |   BS > 6.750
|   |   |   |   |   |   |   |   |   |   |   DiastolicBP > 85
|   |   |   |   |   |   |   |   |   |   |   |   i»¿Age > 26: mid risk (high risk=0, low risk=0, mid risk=1)
|   |   |   |   |   |   |   |   |   |   |   |   i»¿Age ≤ 26: low risk (high risk=0, low risk=2, mid risk=0)
```

Gambar 3. 10 Deskripsi dari pohon keputusan

3.1.4.2.3.4 Menentukan Model *Predict*

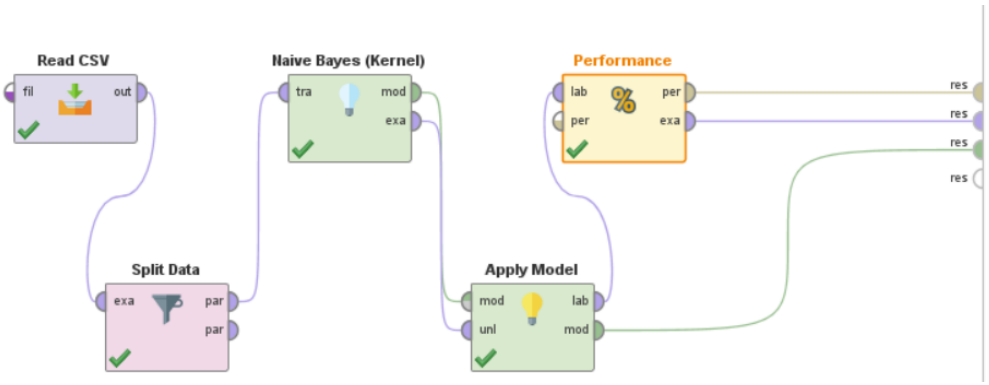
Tahapan selanjutnya adalah menggunakan *Model Predict* untuk melihat prediksi Data Testing berdasarkan Algoritma yang digunakan..



Gambar 3. 11 Model Prediksi untuk uji coba Algoritma

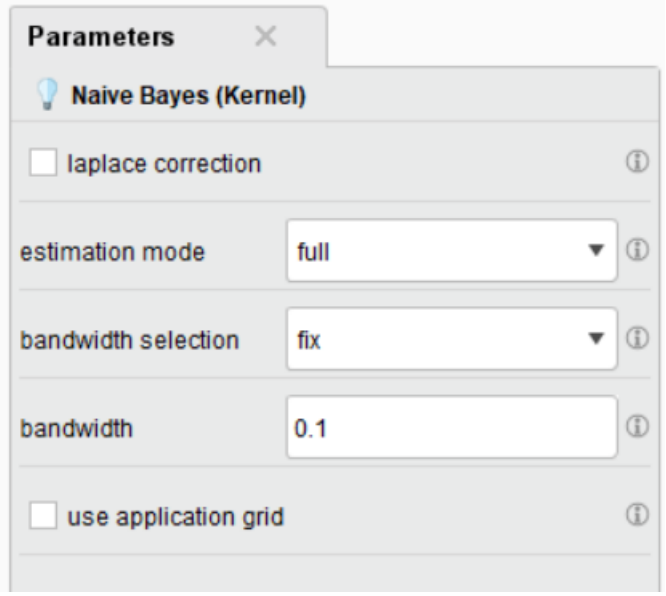
3.1.5 Klasifikasi menggunakan Algoritma Naïve Byes

Pada tahapan ini dilakukan Pengklasifikasian Menggunakan Algoritma Naive Byes dengan model design seperti dibawah ini,

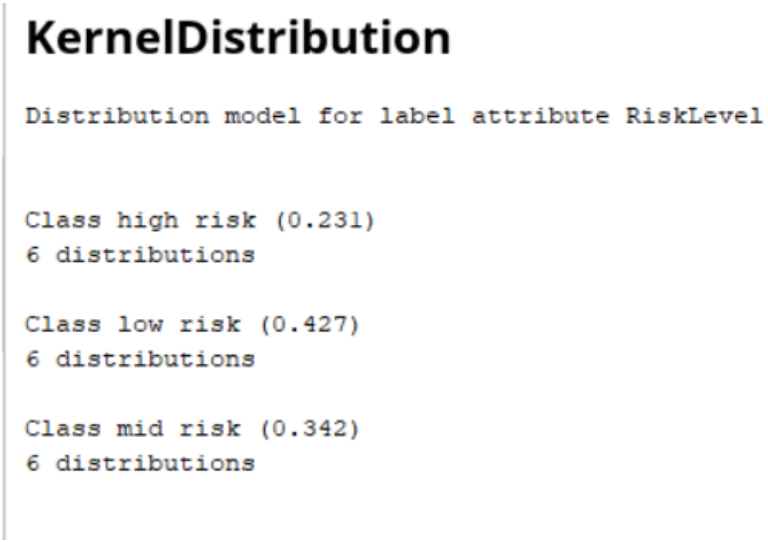


Gambar 3. 12 Model Algoritma Naïve Byes

kemudian pada Parameter disesuaikan seperti gambar dibawah ini untuk menghasilkan akurasi yang maksimal.



Gambar 3. 13 Parameter Naïve byes



KernelDistribution

```
Distribution model for label attribute RiskLevel


Class high risk (0.231)
6 distributions

Class low risk (0.427)
6 distributions

Class mid risk (0.342)
6 distributions
```

Gambar 3. 14 Model Naïve Byes