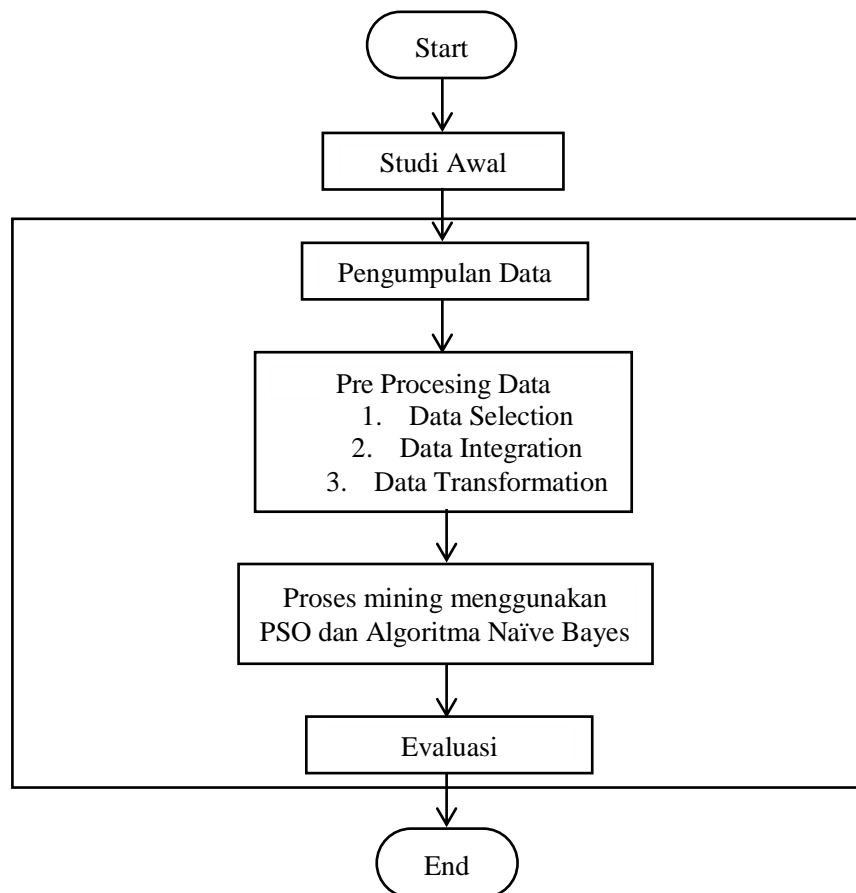


BAB III

METODOLOGI PENELITIAN

3.1. Metode Penelitian

Pada bab ini akan membahas langkah-langkah dari proses penelitian yang akan dilaksanakan, 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 berikut :



Gambar 3. 1 Alur Tahapan Penelitian

Tahapan pada gambar 3.1 . Adalah proses data mining pada penelitian ini :

1. Tahapan studi awal merupakan tahap awal dalam melakukan penelitian. Pada tahap ini, peneliti melakukan observasi atau pemahaman penelitian yang meliputi tujuan dan persyaratan proyek dengan jelas dalam hal bisnis atau unit penelitian secara keseluruhan, menterjemahkan tujuan dan batasan ke dalam

perumusan definisi masalah data mining, menyiapkan strategi awal untuk mencapai tujuan tersebut.

2. Tahap pengumpulan data merupakan suatu tahapan pengumpulan data menggunakan analisis data eksplorasi untuk membiasakan diri dengan data dan menemukan wawasan awal dan mengevaluasi kualitas data. Data penelitian yang digunakan diperoleh dari situs *UCI Machine Learning Repository* dengan nama *heart-disease*. Berikut adalah *link url* dari dataset tersebut. <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>.

3. *Preprocessing data* merupakan proses pengolahan data mentah kedalam bentuk yang lebih mudah dipahami, proses ini penting dilakukan karena data mentah sering kali tidak memiliki format yang teratur.

a) *Data Selection*

Pada proses seleksi ini, dipilih dari keseluruhan data dari dataset *UCI Machine Learning Repository* yaitu data prediksi gejala penyakit jantung. Dari hasil seleksi akan digunakan sebanyak 303 data pasien untuk perhitungan data mining.

b) *Data Integration*

Proses ini dilakukan karena perlu adanya perubahan skema struktur data dari data mentah ke data yang siap akan dilakukan proses *preprocessing* lebih lanjut dengan tahap *transformation*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	68.1,145.233,1.0,150.0,2.3,0.0,1.1																	
2	37,1,2,130,250,0,1,187,0,3,5,0,0,2,1																	
3	41,0,1,130,204,0,0,1,172,0,1,4,2,0,2,1																	
4	56,1,1,120,236,0,1,178,0,0,8,2,0,2,1																	
5	57,0,0,120,354,0,1,163,1,0,6,2,0,2,1																	
6	57,1,0,140,192,0,1,148,0,0,4,1,0,1,1																	
7	56,0,1,140,294,0,0,153,0,1,3,1,0,2,1																	
8	44,1,1,120,263,0,1,173,0,0,2,0,3,1																	
9	52,1,2,172,199,1,1,162,0,0,5,2,0,3,1																	
10	57,1,2,150,168,0,1,174,0,1,6,2,0,2,1																	
11	64,1,0,140,239,0,1,160,0,1,2,2,0,2,1																	
12	48,0,2,130,275,0,1,139,0,0,2,2,0,2,1																	
13	49,1,1,130,266,0,1,171,0,0,6,2,0,2,1																	
14	64,1,3,110,211,0,0,144,1,1,8,1,0,2,1																	
15	58,0,3,150,283,1,0,162,0,1,2,0,2,1																	
16	59,0,3,130,219,0,1,158,0,1,6,1,0,2,1																	
17	58,0,2,120,340,0,1,172,0,0,2,0,2,1																	
18	66,0,3,150,226,0,1,114,0,2,6,0,0,2,1																	
19	43,1,0,150,247,0,1,171,0,1,5,2,0,2,1																	
20	69,0,3,140,239,0,1,151,0,1,4,2,2,2,1																	
21	59,1,0,135,234,0,1,161,0,0,5,1,0,3,1																	
22	44,1,2,130,233,0,1,179,1,0,4,2,0,2,1																	
23	42,1,0,140,226,0,1,178,0,0,2,0,2,1																	
24	61,1,2,150,243,1,1,137,1,1,1,0,2,1																	
25	60,1,3,140,199,0,1,178,1,1,4,2,0,3,1																	
26	71,0,1,160,302,0,1,162,0,0,4,2,2,2,1																	
27	59,1,2,150,212,1,1,157,0,1,6,2,0,2,1																	
28	51,1,2,110,175,0,1,123,0,0,6,2,0,2,1																	
29	65,0,2,140,417,1,0,157,0,0,8,2,1,2,1																	
30	53,1,2,130,197,1,0,152,0,1,2,0,0,2,1																	
31	41,0,1,105,198,0,1,168,0,0,2,1,2,1																	
32																		

Gambar 3. 2 Data Sebelum *Integration*

Data akan berubah skema atau struktur menjadi seperti pada Gambar 3.3 yang mudah untuk diproses. Data tabel didefinisikan memanjang ke samping kanan dengan penggunaan kolom yang lebih banyak dan sesuai secara rapi serta *type file* yang sebelumnya *.csv* di *export* menjadi *type data excel* sehingga memudahkan untuk merubah data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalachh	exang	oldpeak	sip	caa	thal	output
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1

Gambar 3. 3 Data Sesudah *Integration*

c) *Data Transformation*

Proses *Transformation* merupakan perubahan data tahap akhir yang siap diolah dan dimasukkan kedalam *tools* atau aplikasi *rapidminer*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalachh	exang	oldpeak	sip	caa	thal	output
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1

Gambar 3. 4 *Dataset Transformation*

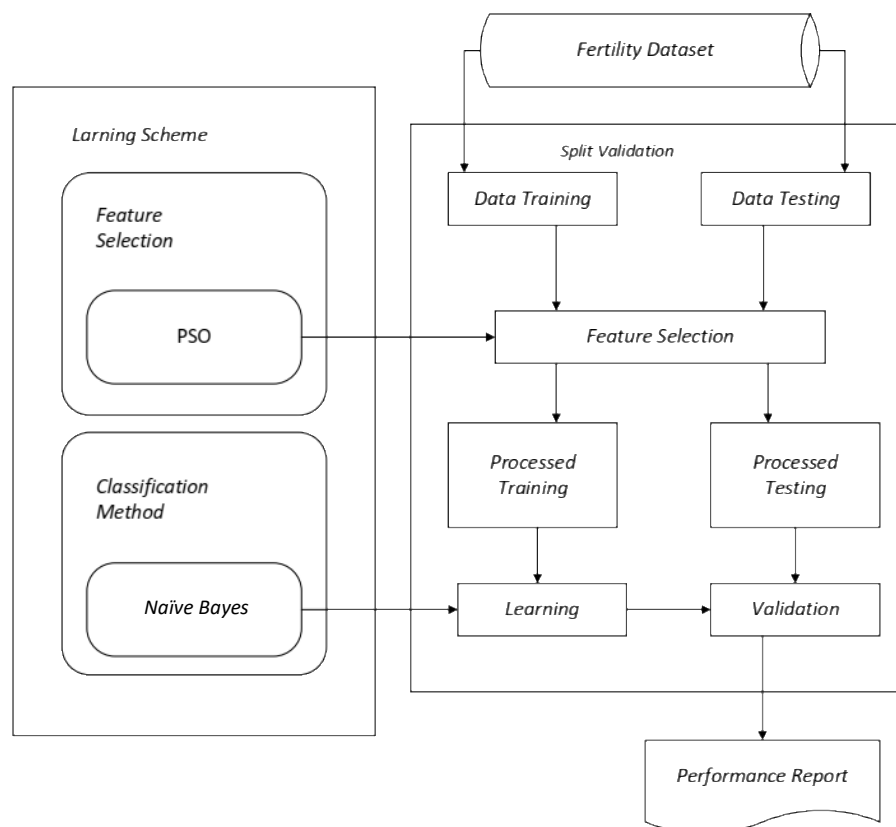
Semua atribut berjumlah 14 dimana terdapat 1 atribut label (*num/output*) dan 13 atribut fitur. Terdapat 303 data pasien penderita penyakit jantung dan 14 atribut meliputi *age*, *sex*, *cp*, *trestbps*, *chol*, *fbs*, *restecg*, *thalach*, *exang*, *oldpeak*, *slope*, *ca*, *thal*, *num*. Penjelasan data dapat dilihat dibawah ini.

Tabel 3. 1 Atribut Dataset Penyakit Jantung

Atribut	Keterangan
(Age)	Umur, [29-77]
(Sex)	jenis kelamin (1 = laki-laki; 0 = perempuan)
(cp)	tipe nyeri dada a. Nilai 1: khas angina b. Nilai 2: angina atipikal c. Nilai 3: nyeri non-angina d. Nilai 4: tanpa gejala
(trestbps)	tekanan darah istirahat (saat otot jantung istirahat) (dalam mm Hg saat masuk ke RSUD), [94-200]
(chol)	serum kolestoral (seluruh jumlah kolesterol dalam darah) dalam m/dl [126--564]
(fbs)	(gula darah puasa / sebelum makan > 120m/dl) (1 = benar; 0 = salah)
(restecg)	beristirahat hasil elektrokardiografi (alat pemeriksa otot jantung) a. Nilai 0: normal b. Nilai 1: memiliki kelainan gelombang ST-T (inversi gelombang T dan / atau ST c. Nilai 2: hipertrofi ventrikel kiri oleh Estes' kriteria
(thalach)	denyut jantung maksimum tercapai, [71-202]
(exang)	olahraga yang diinduksi angina (1 = ya; 0 = tidak)
(oldpeak)	ST depresi yang disebabkan oleh olahraga relatif terhadap istirahat, [0--6.2]
(slope)	kemiringan segmen ST latihan puncak a. Nilai 1: menanjak b. Nilai 2: datar c. Nilai 3: downsloping
(ca)	jumlah pembuluh darah (0-4) dijelaskan dengan fluoroskopi
(thal)	0 = tidak normal; 1 = normal; 2 = cacat ; 3 = cacat tetap

(num)/output	(atribut yang diprediksi) diagnosis penyakit jantung (Label) a. Nilai 0: <50% penyempitan diameter (No) b. Nilai 1:> penyempitan diameter 50% (Yes)
--------------	---

4. Proses mining menggunakan metode *Particle Swam Optimization (PSO)* dan *Algoritma Naïve Bayes* untuk meningkatkan akurasi.



Gambar 3. 5 Alur Proses PSO

- Tahapan pertama yaitu pengumpulan *dataset fertility* yang diambil dari *UCI Machine Learning Repository*.
- Berikutnya pembagian data menggunakan *split validation*, yaitu membagi data menjadi data *training* dan data *testing*.
- Pada tahap learning schame menggunakan *Particle Swarm Optimization* untuk optimasi, sehingga akan mengoptimasi nilai bobot pada setiap atribut.
- Selanjutnya proses klasifikasi menggunakan *algoritma Naïve Bayes* dan masuk ketahap validasi atau melihat kebenaran dan hasil dari proses sebelumnya.

- Pengujian *Algoritma Naïve Bayes*

Pengujian data yang digunakan dalam *algoritma Naïve Bayes* adalah dataset prediksi gejala penyakit jantung yang didapatkan dari *UCI Machine Learning Repository*. Data yang digunakan dalam pengujian ini dapat dilihat pada gambar berikut.

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
67	1	0	160	286	0	0	108	1	1.5	1	3	2	0
67	1	0	120	229	0	0	129	1	2.6	1	2	3	0

Gambar 3. 6 Dataset Prediksi Gejala Penyakit Jantung

3.2. Perhitungan Algoritma *Naïve Bayes*

Untuk menghitung nilai probabilitas dari masing-masing atribut menggunakan rumus algoritma *naïve bayes* di bawah ini.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

$$P(X) = \sum_{n=1}^n P(X|H)P(X)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(x_k - \mu_{ik})^2}{2\sigma^2}}$$

3.2.1. Menghitung Probabilitas Kelas

Pada atribut kelas bertuliskan *output* yang didefinisikan sebagai hasil prediksi diagnosis jantung yang terdiri dari angka 0 dan 1. Dimana dijelaskan bahwa

- Nilai 0 : <50% penyempitan diameter (No)
- Nilai 1 : >50% penyempitan diameter (Yes)

Pada atribut *output* pada nilai 0 terdiri dari 138 dan nilai 1 adalah 165 dari total data sebesar 303.

- Menghitung probabilitas *output* 0

$$\frac{\text{Jumlah } output(0)}{\text{Jumlah data}} = \frac{138}{303} = 0.4554$$

- Menghitung probabilitas *output* 1

$$\frac{\text{Jumlah } output(1)}{\text{Jumlah data}} = \frac{165}{303} = 0.5446$$

3.2.2. Menghitung Probabilitas Atribut Kategorikal/Diskrit

Pada atribut yang berjenis kategorikal / diskrit terdiri dari 8 atribut yaitu *sex*, *cp*, *fbs*, *restecg*, *exng*, *slp*, *ca* dan *thall*. Berikut contoh perhitungan probabilitas dari masing-masing atribut.

1. Menghitung probabilitas atribut *sex*

Jumlah *output* (Perempuan) “0” = 138

Jumlah *output* (Laki-laki) “1” = 165

- a. Menghitung Probabilitas *Sex* “0”

Jumlah *Sex* “0” yang berada pada *output* “0” = 24

$$\text{Probabilitas} = \frac{24}{138} = 0.1739$$

Jumlah *Sex* “0” yang berada pada *output* “1” = 72

$$\text{Probabilitas} = \frac{72}{165} = 0.4364$$

- b. Menghitung Probabilitas *Sex* “1”

Jumlah *Sex* “1” yang berada pada *output* “0” = 114

$$\text{Probabilitas} = \frac{114}{138} = 0.8261$$

Jumlah *Sex* “1” yang berada pada *output* “1” = 93

$$\text{Probabilitas} = \frac{93}{165} = 0.5636$$

2. Menghitung probabilitas atribut *cp*

Jumlah *output* “0” = 138

Jumlah *output* “1” = 165

- a. Menghitung Probabilitas *cp* “0”

Jumlah *cp* “0” yang berada pada *output* “0” = 104 sedangkan pada *output* “1” = 39

$$\text{Probabilitas } cp \text{ “0”} | \text{output “0”} = \frac{104}{138} = 0.7536$$

$$\text{Probabilitas } cp \text{ “0”} | \text{output “1”} = \frac{39}{165} = 0.2364$$

- b. Menghitung Probabilitas *cp* “1”

Jumlah *cp* “1” yang berada pada *output* “0” = 9 sedangkan pada *output* “1” = 41

$$\text{Probabilitas } cp \text{ “1”} | \text{output “0”} = \frac{9}{138} = 0.0652$$

$$\text{Probabilitas } cp \text{ “1”} | \text{output “1”} = \frac{41}{165} = 0.2485$$

c. Menghitung Probabilitas cp “2”

Jumlah cp “2” yang berada pada $output$ “0” = 18 sedangkan pada $output$ “1” = 69

$$\text{Probabilitas } cp \text{ “2”} \mid \text{output “0”} = \frac{18}{138} = 0.1304$$

$$\text{Probabilitas } cp \text{ “2”} \mid \text{output “1”} = \frac{69}{165} = 0.4182$$

d. Menghitung Probabilitas cp “3”

Jumlah cp “3” yang berada pada $output$ “0” = 7 sedangkan pada $output$ “1” = 16

$$\text{Probabilitas } cp \text{ “3”} \mid \text{output “0”} = \frac{7}{138} = 0.0507$$

$$\text{Probabilitas } cp \text{ “3”} \mid \text{output “1”} = \frac{16}{165} = 0.0970$$

3. Menghitung probabilitas atribut fs

Jumlah $output$ “0” = 138

Jumlah $output$ “1” = 165

a. Menghitung Probabilitas fs “0”

Jumlah fs “0” yang berada pada $output$ “0” = 116 sedangkan pada $output$ “1” = 142

$$\text{Probabilitas } fs \text{ “0”} \mid \text{output “0”} = \frac{116}{138} = 0.8406$$

$$\text{Probabilitas } fs \text{ “0”} \mid \text{output “1”} = \frac{142}{165} = 0.8606$$

b. Menghitung Probabilitas fs “1”

Jumlah fs “1” yang berada pada $output$ “0” = 22 sedangkan pada $output$ “1” = 23

$$\text{Probabilitas } fs \text{ “1”} \mid \text{output “0”} = \frac{22}{138} = 0.1594$$

$$\text{Probabilitas } fs \text{ “1”} \mid \text{output “1”} = \frac{23}{165} = 0.1394$$

4. Menghitung probabilitas atribut $restecg$

Jumlah $output$ “0” = 138

Jumlah $output$ “1” = 165

a. Menghitung Probabilitas $restecg$ “0”

Jumlah $restecg$ “0” yang berada pada $output$ “0” = 79 sedangkan pada $output$ “1” = 68

$$\text{Probabilitas } restecg \text{ "0"} | \text{output "0"} = \frac{79}{138} = 0.5725$$

$$\text{Probabilitas } restecg \text{ "0"} | \text{output "1"} = \frac{68}{165} = 0.4121$$

b. Menghitung Probabilitas *restecg* "1"

Jumlah *restecg* "1" yang berada pada *output* "0" = 56 sedangkan pada *output* "1" = 96

$$\text{Probabilitas } restecg \text{ "1"} | \text{output "0"} = \frac{56}{138} = 0.4058$$

$$\text{Probabilitas } restecg \text{ "1"} | \text{output "1"} = \frac{96}{165} = 0.5818$$

c. Menghitung Probabilitas *restecg* "2"

Jumlah *restecg* "2" yang berada pada *output* "0" = 3 sedangkan pada *output* "1" = 1

$$\text{Probabilitas } restecg \text{ "2"} | \text{output "0"} = \frac{3}{138} = 0.0217$$

$$\text{Probabilitas } restecg \text{ "2"} | \text{output "1"} = \frac{1}{165} = 0.0061$$

5. Menghitung probabilitas atribut *exng*

Jumlah *output* "0" = 138

Jumlah *output* "1" = 165

a. Menghitung Probabilitas *exng* "0"

Jumlah *exng* "0" yang berada pada *output* "0" = 62 sedangkan pada *output* "1" = 142

$$\text{Probabilitas } exng \text{ "0"} | \text{output "0"} = \frac{62}{138} = 0.4493$$

$$\text{Probabilitas } exng \text{ "0"} | \text{output "1"} = \frac{142}{165} = 0.8606$$

b. Menghitung Probabilitas *exng* "1"

Jumlah *exng* "1" yang berada pada *output* "0" = 76 sedangkan pada *output* "1" = 142

$$\text{Probabilitas } exng \text{ "1"} | \text{output "0"} = \frac{76}{138} = 0.5507$$

$$\text{Probabilitas } exng \text{ "1"} | \text{output "1"} = \frac{142}{165} = 0.8606$$

6. Menghitung probabilitas atribut *slp*

Jumlah *output* "0" = 138

Jumlah *output* "1" = 165

a. Menghitung Probabilitas *slp* “0”

Jumlah *slp* “0” yang berada pada *output* “0” = 12 sedangkan pada *output* “1” = 9

$$\text{Probabilitas } slp \text{ “0”} \mid \textit{output} \text{ “0”} = \frac{12}{138} = 0.0870$$

$$\text{Probabilitas } slp \text{ “0”} \mid \textit{output} \text{ “1”} = \frac{9}{165} = 0.0545$$

b. Menghitung Probabilitas *slp* “1”

Jumlah *slp* “1” yang berada pada *output* “0” = 91 sedangkan pada *output* “1” = 49

$$\text{Probabilitas } slp \text{ “1”} \mid \textit{output} \text{ “0”} = \frac{91}{138} = 0.6594$$

$$\text{Probabilitas } slp \text{ “1”} \mid \textit{output} \text{ “1”} = \frac{49}{165} = 0.2970$$

c. Menghitung Probabilitas *slp* “2”

Jumlah *slp* “2” yang berada pada *output* “0” = 35 sedangkan pada *output* “1” = 107

$$\text{Probabilitas } slp \text{ “2”} \mid \textit{output} \text{ “0”} = \frac{35}{138} = 0.2536$$

$$\text{Probabilitas } slp \text{ “2”} \mid \textit{output} \text{ “1”} = \frac{107}{165} = 0.6485$$

7. Menghitung probabilitas atribut *ca*

Jumlah *output* “0” = 138

Jumlah *output* “1” = 165

a. Menghitung Probabilitas *ca* “0”

Jumlah *ca* “0” yang berada pada *output* “0” = 45 sedangkan pada *output* “1” = 130

$$\text{Probabilitas } ca \text{ “0”} \mid \textit{output} \text{ “0”} = \frac{45}{138} = 0.3261$$

$$\text{Probabilitas } ca \text{ “0”} \mid \textit{output} \text{ “1”} = \frac{130}{165} = 0.7879$$

b. Menghitung Probabilitas *ca* “1”

Jumlah *ca* “1” yang berada pada *output* “0” = 44 sedangkan pada *output* “1” = 21

$$\text{Probabilitas } ca \text{ “1”} \mid \textit{output} \text{ “0”} = \frac{44}{138} = 0.3188$$

$$\text{Probabilitas } ca \text{ “1”} \mid \textit{output} \text{ “1”} = \frac{21}{165} = 0.1273$$

c. Menghitung Probabilitas *ca* “2”

Jumlah *ca* “2” yang berada pada *output* “0” = 31 sedangkan pada *output* “1” = 7

$$\text{Probabilitas } ca \text{ “2”} \mid \textit{output} \text{ “0”} = \frac{31}{138} = 0.2246$$

$$\text{Probabilitas } ca \text{ “2”} \mid \textit{output} \text{ “1”} = \frac{7}{165} = 0.0424$$

d. Menghitung Probabilitas *ca* “3”

Jumlah *ca* “3” yang berada pada *output* “0” = 17 sedangkan pada *output* “1” = 3

$$\text{Probabilitas } ca \text{ “3”} \mid \textit{output} \text{ “0”} = \frac{17}{138} = 0.1232$$

$$\text{Probabilitas } ca \text{ “3”} \mid \textit{output} \text{ “1”} = \frac{3}{165} = 0.0182$$

e. Menghitung Probabilitas *ca* “4”

Jumlah *ca* “4” yang berada pada *output* “0” = 1 sedangkan pada *output* “1” = 4

$$\text{Probabilitas } ca \text{ “4”} \mid \textit{output} \text{ “0”} = \frac{1}{138} = 0.0072$$

$$\text{Probabilitas } ca \text{ “4”} \mid \textit{output} \text{ “1”} = \frac{4}{165} = 0.0242$$

8. Menghitung probabilitas atribut *thall*

Jumlah *output* “0” = 138

Jumlah *output* “1” = 165

a. Menghitung Probabilitas *thall* “0”

Jumlah *thall* “0” yang berada pada *output* “0” = 1 sedangkan pada *output* “1” = 1

$$\text{Probabilitas } thall \text{ “0”} \mid \textit{output} \text{ “0”} = \frac{1}{138} = 0.0072$$

$$\text{Probabilitas } thall \text{ “0”} \mid \textit{output} \text{ “1”} = \frac{1}{165} = 0.0061$$

b. Menghitung Probabilitas *thall* “1”

Jumlah *thall* “1” yang berada pada *output* “0” = 12 sedangkan pada *output* “1” = 6

$$\text{Probabilitas } thall \text{ “1”} \mid \textit{output} \text{ “0”} = \frac{12}{138} = 0.0870$$

$$\text{Probabilitas } thall \text{ “1”} \mid \textit{output} \text{ “1”} = \frac{6}{165} = 0.0364$$

c. Menghitung Probabilitas *thall* “2”

Jumlah *thall* “2” yang berada pada *output* “0” = 36 sedangkan pada *output* “1” = 130

$$\text{Probabilitas } thall \text{ “2”} \mid \text{output “0”} = \frac{36}{138} = 0.2609$$

$$\text{Probabilitas } thall \text{ “2”} \mid \text{output “1”} = \frac{130}{165} = 0.7879$$

d. Menghitung Probabilitas *thall* “3”

Jumlah *thall* “3” yang berada pada *output* “0” = 89 sedangkan pada *output* “1” = 28

$$\text{Probabilitas } thall \text{ “3”} \mid \text{output “0”} = \frac{89}{138} = 0.6449$$

$$\text{Probabilitas } thall \text{ “3”} \mid \text{output “1”} = \frac{28}{165} = 0.1697$$

3.2.3. Menghitung Probabilitas Atribut Numerik

Pada atribut yang bertipe numerik untuk algoritma *naïve bayes* harus dilakukan dengan menghitung nilai rata-rata atau *mean* dan standar deviasi dari nilai-nilai pada setiap kelas dan setiap atribut yang ada. Berikut adalah masing-masing nilai rata-rata dan standar deviasi dari atribut yang bertipe numerik.

Tabel 3. 2 Menghitung Nilai MEAN dan Standar Deviasi

MEAN						
Kelas	Atribut	age	trtbps	chol	thalachh	oldpeak
0		56.60	134.40	251.09	139.10	0.39
1		52.50	129.30	242.23	158.47	0.08
Standar Deviasi						
Kelas	Atribut	age	trtbps	chol	thalachh	oldpeak
0		7.96	18.73	49.45	22.60	0.86
1		9.55	16.17	53.55	19.17	0.35

3.2.4. Uji Validasi Menggunakan *Confusion Matrix*

Untuk menguji model dari algoritma *naïve bayes* adalah mencari nilai akurasi, presisi dan recall. Hal tersebut bisa dilakukan dengan menggunakan tabel *confusion matrix* sebagai acuan validasi. Berikut adalah set data uji yang diambil secara acak sebanyak 5 *record data* dari total data latih.

Tabel 3. 3 Uji Validasi Menggunakan *Confusion Matrix*

Atribut	Data uji 1	Data uji 2	Data uji 3	Data uji 4	Data uji 5
age	63	38	38	67	67
sex	1	1	1	1	1
cp	3	2	2	0	0
trtbps	145	138	138	160	120
chol	233	175	175	286	229
fbs	1	0	0	0	0
restecg	0	1	1	0	0
thalachh	150	173	173	108	129
exng	0	0	0	1	1
oldpeak	2.3	0	0	1.5	2.6
slp	0	2	2	1	1
caa	0	4	4	3	2
thall	1	2	2	2	3
output	1	1	1	0	0

3.2.4.1. Menghitung Probabilitas Dengan Fungsi Gaussian

Pada dasarnya, proses pembelajaran *naïve bayes* untuk data numerik sama saja dengan untuk data kategorikal. Sedikit perbedaannya adalah terletak pada perhitungan probabilitasnya yang menggunakan fungsi *gaussian* yang didefinisikan sebagai

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(x_k - \mu_{ik})^2}{2\sigma^2}}$$

Dimana μ_{ik} dan σ_{ik} adalah nilai *mean* dan standar deviasi dari masing-masing atribut yang bertipe numerik pada setiap kelasnya. Berikut contoh perhitungan pada pada kasus data uji 1.

1. Menghitung Probabilitas Atribut Age

a. *Gaussian* atribut age output "0"

$$\sigma_{ik} \text{ (standar deviasi pada age, dan output "0")} = 7.96$$

$$\mu_{ik} \text{ (mean pada age, dan output "0")} = 56.60$$

$$x_k \text{ (nilai dari data uji 1)} = 63$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 7.96}$$

$$f(x) = 1 / 7.070275808 = \mathbf{0.141437198}$$

$$\text{Exponen} - \frac{(63-56.60)}{2*(7.96^2)} = \mathbf{0.724039821}$$

$$\text{Probabilitas } age | \text{output "0"} = 0.141437198 * 0.724039821$$

$$\text{Probabilitas } age | \text{output "0"} = \mathbf{0.102392777}$$

b. *Gaussian* atribut *age* output "1"

$$\sigma_{ik} \text{ (standar deviasi pada } age, \text{ dan output "1")} = 9.55$$

$$\mu_{ik} \text{ (mean pada } age, \text{ dan output "1")} = 52.50$$

$$x_k \text{ (nilai dari data uji 1)} = 63$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 9.55}$$

$$f(x) = 1 / 7.070275808 = \mathbf{0.129123026}$$

$$\text{Exponen} - \frac{(63-52.60)}{2*(9.55^2)} = \mathbf{0.546243335}$$

$$\text{Probabilitas } age | \text{output "1"} = 0.129123026 * 0.546243335$$

$$\text{Probabilitas } age | \text{output "1"} = \mathbf{0.070532593}$$

2. Menghitung Probabilitas Atribut *trtbps*

a. *Gaussian* atribut *trtbps* output "0"

$$\sigma_{ik} \text{ (standar deviasi pada } trtbps, \text{ dan output "0")} = 18.73$$

$$\mu_{ik} \text{ (mean pada } trtbps, \text{ dan output "0")} = 134.40$$

$$x_k \text{ (nilai dari data uji 1 atribut } trtbps) = 145$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 18.73} = \mathbf{0.092204462}$$

$$\text{Exponen} - \frac{(145-134.40)}{2*(18.73^2)} = \mathbf{0.851984371}$$

$$\text{Probabilitas } trtbps | \text{output "0"} = 0.092204462 * 0.851984371$$

$$\text{Probabilitas } trtbps | \text{output "0"} = \mathbf{0.078556761}$$

b. *Gaussian* atribut *trtbps* output “1”

$$\sigma_{ik} \text{ (standar deviasi pada } trtbps, \text{ dan output “1”)} = 16.17$$

$$\mu_{ik} \text{ (mean pada } trtbps, \text{ dan output “1”)} = 129.30$$

$$x_k \text{ (nilai dari data uji 1 atribut } trtbps) = 145$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 16.17} = \mathbf{0.099236254}$$

$$\text{Exponen} - \frac{(145-129.30)}{2*(16.17^2)} = \mathbf{0.624254682}$$

$$\text{Probabilitas } trtbps | \text{output “1”} = 0.099236254 * 0.624254682$$

$$\text{Probabilitas } trtbps | \text{output “1”} = \mathbf{0.061948696}$$

3. Menghitung Probabilitas Atribut *chol*

a. *Gaussian* atribut *chol* output “0”

$$\sigma_{ik} \text{ (standar deviasi pada } chol, \text{ dan output “0”)} = 49.45$$

$$\mu_{ik} \text{ (mean pada } chol, \text{ dan output “0”)} = 251.09$$

$$x_k \text{ (nilai dari data uji 1 atribut } trtbps) = 233$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 49.45} = \mathbf{0.056743585}$$

$$\text{Exponen} - \frac{(233-251.09)}{2*(49.45^2)} = \mathbf{0.935308713}$$

$$\text{Probabilitas } chol | \text{output “0”} = 0.056743585 * 0.935308713$$

$$\text{Probabilitas } chol | \text{output “0”} = \mathbf{0.05307277}$$

b. *Gaussian* atribut *chol* output “1”

$$\sigma_{ik} \text{ (standar deviasi pada } chol, \text{ dan output “1”)} = 53.55$$

$$\mu_{ik} \text{ (mean pada } chol, \text{ dan output “1”)} = 242.23$$

$$x_k \text{ (nilai dari data uji 1 atribut } trtbps) = 233$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 53.55}$$

$$f(x) = 1 / 7.070275808 = \mathbf{0.054529159}$$

$$\text{Exponen} - \frac{(233-242.23)}{2*(53.55^2)} = \mathbf{0.985256013}$$

$$\text{Probabilitas } chol | \text{output "1"} = 0.054529159 * 0.985256013$$

$$\text{Probabilitas } chol | \text{output "1"} = \mathbf{0.053725181}$$

4. Menghitung Probabilitas Atribut *thalachh*

a. *Gaussian* atribut *thalachh* output "0"

$$\sigma_{ik} \text{ (standar deviasi pada } thalachh, \text{ dan output "0")} = 22.60$$

$$\mu_{ik} \text{ (mean pada } thalachh, \text{ dan output "0")} = 139.10$$

$$x_k \text{ (nilai dari data uji 1 atribut } thalachh) = 150$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 22.60} = \mathbf{0.08394168}$$

$$\text{Exponen} - \frac{(150-139.10)}{2*(22.60^2)} = \mathbf{0.890218201}$$

$$\text{Probabilitas } thalachh | \text{output "0"} = 0.08394168 * 0.890218201$$

$$\text{Probabilitas } thalachh | \text{output "0"} = \mathbf{0.074726412}$$

b. *Gaussian* atribut *thalachh* output "1"

$$\sigma_{ik} \text{ (standar deviasi pada } thalachh, \text{ dan output "1")} = 19.17$$

$$\mu_{ik} \text{ (mean pada } thalachh, \text{ dan output "1")} = 158.47$$

$$x_k \text{ (nilai dari data uji 1 atribut } thalachh) = 150$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 19.17} = \mathbf{0.091129858}$$

$$\text{Exponen} - \frac{(150-158.47)}{2*(19.17^2)} = \mathbf{0.907112039}$$

$$\text{Probabilitas } thalachh | \text{output "1"} = 0.091129858 * 0.907112039$$

$$\text{Probabilitas } thalachh | \text{output "1"} = \mathbf{0.082664992}$$

5. Menghitung Probabilitas Atribut *thalachh*

a. *Gaussian* atribut *oldpeak* output "0"

$$\sigma_{ik} \text{ (standar deviasi pada } oldpeak, \text{ dan output "0")} = 0.86$$

$$\mu_{ik} \text{ (mean pada } oldpeak, \text{ dan output "0")} = 0.39$$

$$x_k \text{ (nilai dari data uji 1 atribut } oldpeak) = 2.3$$

$$f(x) = 1 / \sqrt{2\pi}\sigma_{ik}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 0.86} = \mathbf{0.42981545}$$

$$\text{Exponen} - \frac{(2.3 - 0.39)}{2 * (0.86^2)} = \mathbf{0.085434118}$$

Probabilitas *oldpeak* | *output* "0" = 0.42981545 * 0.085434118
 Probabilitas *oldpeak* | *output* "0" = **0.036720904**

b. *Gaussian* atribut *oldpeak* *output* "1"

$$\sigma_{ik} \text{ (standar deviasi pada } oldpeak, \text{ dan } output \text{ "1"})} = 0.35$$

$$\mu_{ik} \text{ (mean pada } oldpeak, \text{ dan } output \text{ "1"})} = 0.08$$

$$x_k \text{ (nilai dari data uji 1 atribut } oldpeak) = 2.3$$

$$f(x) = 1 / \sqrt{2\pi\sigma_{ik}}$$

$$f(x) = 1 / \sqrt{2 * 3,14 * 0.35} = \mathbf{0.674593105}$$

$$\text{Exponen} - \frac{(2.3 - 0.08)}{2 * (0.35^2)} = \mathbf{1.82141E-09}$$

Probabilitas *oldpeak* | *output* "1" = 0.674593105 * 1.82141E-09
 Probabilitas *oldpeak* | *output* "1" = **1.22871E-09**

Setelah mencari nilai probabilitas pada atribut data numerik dengan menggunakan fungsi *gaussian* dan untuk atribut data bertipe kategorikal, M masukkan data berdasarkan kelasnya seperti pada table 3.4.

Tabel 3. 4 Hasil Nilai Probabilitas Data Uji 1

Atribut	Data uji 1	Output "0"	Output "1"
age	63	0.102392777	0.070532593
sex	1	0.8261	0.5636
cp	3	0.0507	0.0970
trtbps	145	0.078556761	0.061948696
chol	233	0.05307277	0.053725181
fbs	1	0.1594	0.1394
restecg	0	0.5725	0.4121
thalachh	150	0.074726412	0.082664992
exng	0	0.4493	0.8606
oldpeak	2.3	0.036720904	1.22871E-09
slp	0	0.0870	0.0545
caa	0	0.3261	0.7879
thall	1	0.0870	0.0364
output	1	0.4554	0.5446
Nilai Probabilitas		2.26014E-12	5.48285E-20

Dari tabel diatas nilai probabilitas *output* “0” pada kasus data uji 1 sebesar 2.26014E-12 dan nilai probabilitas *output* “1” sebesar 5.48285E-20. Dari kedua nilai tersebut, nilai probabilitas *output* “0” lebih besar dari pada nilai probabilitas *output* “1”, sehingga prediksi yang dihasilkan tidak sesuai dengan data aktualnya, dimana aktual =1, dari prediksi = 0

3.2.4.2.Menghitung Akurasi, Presisi dan Recall

Berikut adalah 5 contoh *record* data uji yang sudah diketahui nilai probabilitas dan hasil prediksinya.

Tabel 3. 5 Nilai Probabilitas Dan Hasil Prediksi

Atribut	Data uji 1	Data uji 2	Data uji 3	Data uji 4	Data uji 5
age	63	38	38	67	67
sex	1	1	1	1	1
cp	3	2	2	0	0
trtbps	145	138	138	160	120
chol	233	175	175	286	229
fbs	1	0	0	0	0
restecg	0	1	1	0	0
thalachh	150	173	173	108	129
exng	0	0	0	1	1
oldpeak	2.3	0	0	1.5	2.6
slp	0	2	2	1	1
caa	0	4	4	3	2
thall	1	2	2	2	3
output	1	1	1	0	0
Output (prediksi)	0	1	1	0	0

Dari hali kelima *record* data tersebut dapat divisualisasikan dalam bentuk *confusion matrix* untuk mempermudah validasi. Untuk tabel *confusion matrix* bisa dilihat pada tabel 3.6.

Tabel 3. 6 Confusion Matrix

	Prediksi (<i>output</i>)	
Aktual (<i>output</i>)	0	1
0	2	0
1	1	2

Keterangan :

Positive adalah kelas *output* yang bernilai “0”

Negative adalah kelas *output* yang bernilai “1”

TP (True Positive) = Jumlah data uji yang sebenarnya positif diprediksi positif
 TN (True Negative) = Jumlah data uji yang sebenarnya negatif diprediksi negatif
 FP (False Positive) = Jumlah data uji yang sebenarnya negatif diprediksi positif
 FN (False Negative) = Jumlah data uji yang sebenarnya positif diprediksi negatif

- **TP (True Positive) = 2**
- **TN (True Negative) = 2**
- **FP (False Positive) = 0**
- **FN (False Negative) = 1**

a. Menghitung Akurasi

Berikut adalah cara menghitung nilai akurasi dari tabel 3.6 yang diperoleh.

$$\text{Akurasi} = \frac{TP+TN}{\text{Jumlah data}}$$

$$\text{Akurasi} = \frac{(2+2)}{(2+2+1+0)} = \frac{4}{5} = 0,8 \times 100\% = 80\%$$

b. Menghitung Presisi

Menghitung presisi adalah dengan menggunakan rumus sebagai berikut

$$\text{Presisi} = \frac{Tp}{Tp+Fp}$$

$$\text{Presisi} = \frac{2}{2+0} = \frac{2}{2} = 1 \times 100\% = 100\%$$

c. Menghitung Recall

Untuk menghitung recall adalah menggunakan rumus sebagai berikut

$$\text{Recall} = \frac{Tp}{Tp+Fn}$$

$$\text{Recall} = \frac{2}{(2+1)} = \frac{2}{3} = 0.666666667 \times 100\% = 67\%$$