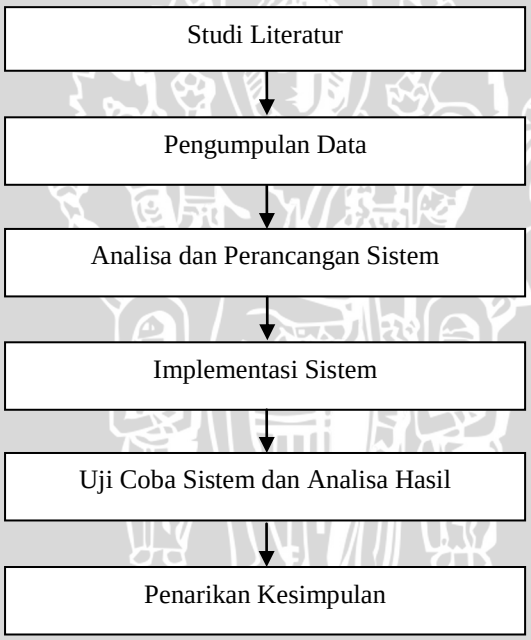
BAB III
METODELOGI PENELITIAN DAN PERANCANGAN

Pada subbab metodologi penelitian ini akan dibahas mengenai tahapan yang merepresentasikan langkah-langkah yang akan dilakukan peneliti dalam proses penelitian yang berjudul "implementasi metode klustering untuk klasifikasi kanker payudara menggunakan algoritma Naïve Bayes dan K-Means".

Tahapan penelitian menunjukkan langkah-langkah yang akan digunakan untuk memecahkan permasalahan *clustering* dan klasifikasi data kanker payudara menggunakan algoritma Naïve Bayes dan K-Means dapat diilustrasikan pada Gambar 3.1.



Gambar 3.1. Tahapan penelitian

Tahap-tahap penelitian yang akan dilakukan sebagai berikut:

1. Studi Literatur.

   Tahap awal dalam penelitian ini adalah mempelajari literatur atau sumber-sumber yang terkait dengan metode yang digunakan (Naïve Bayes dan K-Means) dan obyek penelitian (kanker payudara).

2. Pengumpulan Data.

   Setelah mempelajari literatur terkait dengan metode yang digunakan peneliti akan melakukan pengumpulan data penderita kanker payudara yang didapatkan dari *UCI Machine LearningRepository*(http://archive.ics.uci.edu/ml/machine-learning-databases/breastcancer/breast-cancer-data).

3. Analisa dan Perancangan Sistem.

   Tahap selanjutnya adalah Menganalisa dan melakukan perancangan sistem dengan menggunakan hasil pembelajaran/pelatihan pada tahap sebelumnya.

4. Implementasi Sistem

   Tahap ini mengimplementasikan hasil analisis dan perancangan sistem yang telah dibuat ke dalam komputer dengan menggunakan bahasa pemrograman dengan hasil akhir yang berupa program aplikasi.

5. Uji Coba Sistem dan Analisa Hasil.

   Uji coba ini dilakukan untuk menunjukkan tingkat keberhasilan dari metode klasifikasi yang diuji. Pada tahap ini dilakukan uji coba dengan dataset kanker payudara yang sudah didapat serta nilai-nilai parameter sistem untuk mengetahui apakah sistem yang digunakan telah memberikan hasil yang optimal. Proses pengukuran keberhasilan sistem dilakukan dengan menggunakan akurasi sistem. Setelah dilakukan pengujian, dilakukan analisa terhadap skenario pengujian, sehingga diketahui keefektifan metode dalam mengklasifikasikan data kanker payudara.

6. Penarikan Kesimpulan.

   Tahap penarikan kesimpulan merupakan tahapan akhir pada penelitian. Pada tahap ini akan dilakukan penarikan kesimpulan dari hasil uji coba sistem yang telah dilakukan berdasarkan rumusan masalah.

3.1 Studi Literatur

Dalam pengerjaan penelitian ini dibutuhkan studi literatur untuk mempelajari dasar teori terkait dengan penelitian yang akan dilakukan. Teori-teori tersebut dapat diperoleh dari berbagai macam sumber seperti jurnal, buku, *e-book*, internet, penelitian-penelitian sebelumnya dan sumber pustaka lain yang dapat dipertanggung jawabkan.

Teori yang dipelajari guna melakukan penelitian ini diantaranya adalah data kanker payudara¸ *data maining*, Naïve Bayes, dan K-Means.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah *Breast Cancer Wisconsin (Original)* dataset. Data tersebut diperoleh dari *UCI Machine Learning Repository* (http://archive.ics.uci.edu/ml/machine-learning-databases/breastcancer /breast-cancer-data) dimana data tersebut disumbangkan oleh Dr. WIlliam H. Wolberg dari Amerika. Jumlah data yang digunakan dalam penelitian ini adalah 440 data dengan jumlah data benigna adalah 220 dan 220 data untuk malignan. Dataset tersebut terdiri atas 2 kategori yakni benigna dan malignan. Masing-masing kategori memiliki 11 atribut diantaranya *sample code number* (nomor id pasien), *clump thickness* (ketebalan gumpalan tumor), *uniformity of cell size* (keseragaman ukuran), *uniformity of cell shape* (keseragaman bentuk sel), *marginal adhesion* (kelekatan pinggiran sel), *single epithelial cell size* (ukuran sel tiap jaringan epitel), *bare nuclei* (kekosongan pada inti sel), *bland chromatin* (kromatin lunak), *normal nucleoli* (inti sel normal), *mitoses* (pembelahan mitosis) yang memiliki nilai 1-10 dan *class* yang memiliki nilai 2 atau 4. Atribut yang dimiliki oleh dataset dapat dilihat pada tabel 3.1.

Tabel 3.1 Atribut dataset kanker payudara

	Atribut					
	sample code number	*clump thickness*	*uniformity of cell size*	*uniformity of cell shape*	*marginal adhesion*	*single epithelial cell size*
Range	id number	1-10	1-10	1-10	1-10	1-10
	Atribut					
	bare nuclei	*bland chromatin*	*normal nucleoli*	*mitoses*	*class*	
Range	1-10	1-10	1-10	1-10	*2 for benign, 4 for malignant*	

Berdasarkan 11 atribut tersebut, 9 atribut digunakan untuk proses pengelompokkan K-Means dan 10 atribut untuk klasifikasi Naïve Bayes. Atribut pertama merupakan nomor id dari pasien yang datanya digunakan pada dataset ini, atribut 2 sampai 9 mewakili karakter sitologis dari *Breast Fine-Needle Aspirates* (BFNAs). Setiap atribut memiliki nilai parameter 1-10, nilai 1 merupakan nilai yang paling dekat dengan kanker jinak dan nilai 10 merupakan kanker yang paling ganas.

3.3Analisa Dan Perancangan Sistem

Pada subbab analisa dan perancangan sistem ini akan dibahas mengenai deskripsi umum sistem, perancangan sistem, dan contoh perhitungan manual.

3.3.1 Deskripsi umum sistem

Sistem yang akan dibuat dalam penelitian ini bertujuan untuk mengimplementasikan dan mengetahui kinerja dari metode *clustering* algoritma K-Means dan klasifikasi Naïve Bayes dalam permasalahan kanker payudara.

Dalam mengenali pola pengklasifikasian kanker payudara, sistem membutuhkan masukkan berupa nilai dari 10 indikator yang telah dimasukkan *user*. Sistem akan mengolah inputan yang diberikan *user* dengan cara melakukan pelatihan atau pembelajaran. Sistem ini memiliki 2 proses utama, yaitu proses pelatihan dan proses pengujian. Pada proses pelatihan dibutuhkan masukan berupa data latih. Pada proses pengujian dibutuhkan data uji yang akan diujikan. Proses pada sistem ini adalah sebagai berikut :

1. Proses Pelatihan

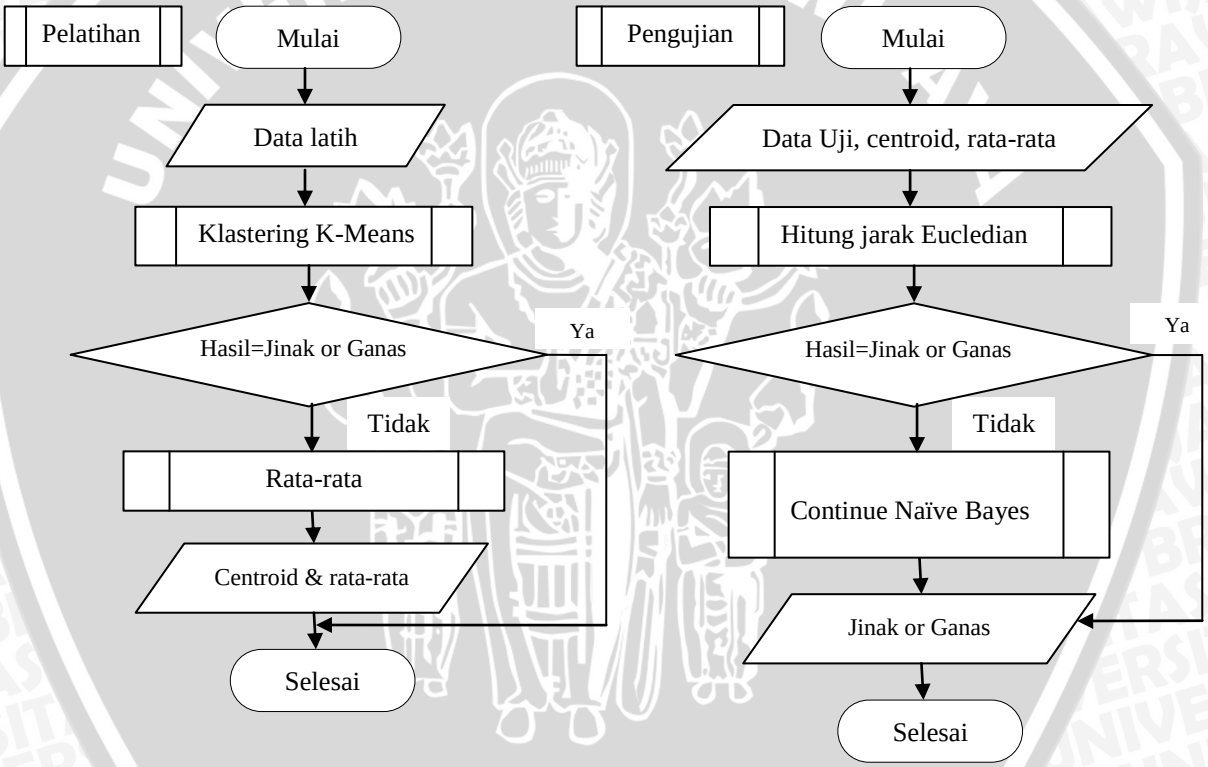
   Tujuan dari proses ini adalah untuk memperoleh nilai *centroid* dari K-Means dan nilai rata-rata dari Naïve Bayes yang akan digunakan selanjutnya untuk data uji. Data latih pada proses K-Means digunakan untuk membentuk *centroid*. Keluaran yang dihasilkan melalui proses ini berupa 3 *centroid* (C=3). Dimana C1 = Jinak, C2 = Ganas, dan C3 =Mungkin. Selanjutnya algoritma Naïve Bayes akan mengolah data dari klaster C3. Tahapan yang dilakukan Naïve Bayes adalah

melakukan perhitungan nilai rata-rata tiap kelas. Naïve bayes akan menghasilkan rata-rata tiap atribut kelas (Jinak dan Ganas).

2. Proses Pengujian

   Pada proses pengujian ini akan mengambil nilai *centroid* dan rata-rata dari proses pelatihan lalu dilakukan perhitungan. Tujuan proses ini adalah untuk penentuan kelas dari data uji yang tidak diketahui kelasnya.

Secara keseluruhan tahapan pelatihan dan pengujian sistem dapat diilustrasikan pada gambar 3.2.



Gambar 3.2 Tahapan pelatihan dan pengujian pada sistem.

3.3.2 Perancangan Sistem

Berdasarkan literatur-literatur yang dipelajari tentang metode pembelajaran algoritma Naïve Bayes dan K-Means pada kasus klasifikasi, maka penelitiakan akan menjabarkan analisa kebutuhan proses yang dibutuhkan sistem.

Pada analisa kebutuhan proses ini peneliti membagi proses pada sistem yang akan dibangun menjadi 2 bagian utama, yaitu proses *clustering* data dengan K-Means dan proses pengklasifikasian data dengan Naïve Bayes.

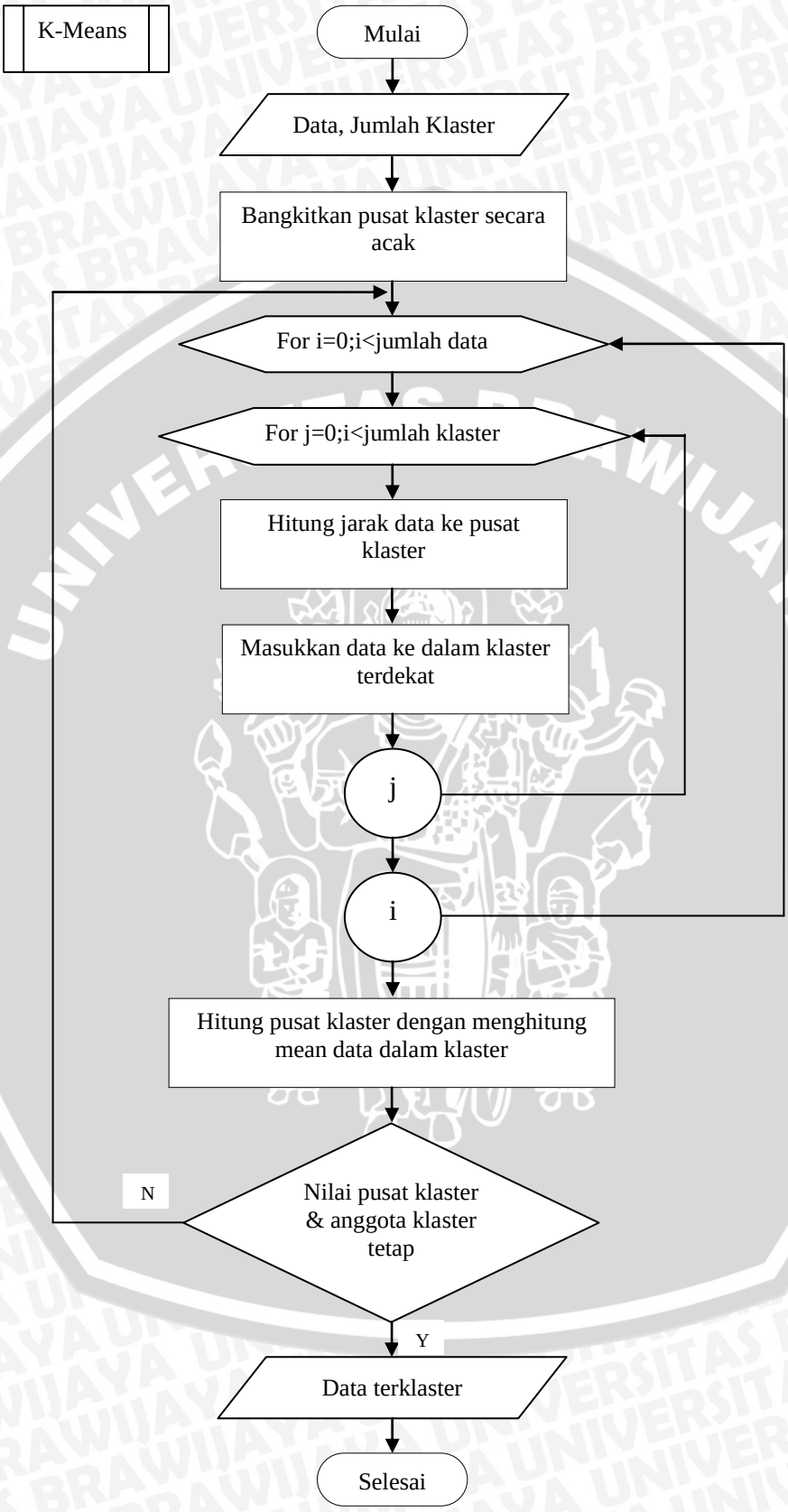
3.3.2.1 Pengklasteran data dengan K-Means

Algoritma K-Means termasuk *portioning clustering* yang memisahkan data ke dalam k daerah bagian yang terpisah. Algoritma K-Means sangat terkenal karena kemudahan dan kemampuannya untuk mengklaster data besar dengan sangat cepat. Prinsip dari proses pengelompokan dengan menggunakan K-Means pada penelitian ini adalah membentuk sebanyak 3 klaster.

Proses ini dijabarkan sebagi berikut :

1. Masukkan (*input*) berupa data kanker payudara.
2. Masukkan (*input*) nilai k.
3. Membangkitkan pusat klaster secara acak.
4. Untuk setiap nilai pada setiap data dilakukan perhitungan d (jarak *euclidean*) menggunakan rumus 2-4.
5. Melakukan *update* klaster yaitu, memasukkan data ke dalam kelompok dengan jarak minimum ke pusat klaster berdasarkan nilai d.
6. Jika sudah satu iterasi, periksa pusat klaster berubah atau tidak, jika berubah maka kembali ke langkah 4, jika tidak maka proses pengklasteran selesai.

Gambar alur sistem pengelompokan K-Means dapat dilihat pada gambar 3.3.



Gambar 3.3 Alur sistem pengelompokkan K-Means

Perhitungan pada program *clustering* data kanker payudara dengan K-Means diawali dengan menentukan jumlah klaster. Penelitian ini menggunakan 3 klaster untuk menentukan jenis kanker payudara, dengan keterangan bahwa klaster 1 adalah jinak, klaster 2 adalah ganas, dan klaster 3 adalah mungkin (tidak teridentifikasi jinak atau ganas oleh K-Means).

Proses K-Means diawali dengan masukan berupa data kanker payudara, dan parameter K-Means yaitu jumlah klaster. Proses selanjutnya yaitu, dengan membangkitkan pusat klaster secara acak untuk dijadikan sebagai pusat klaster. Selanjutnya untuk setiap data dihitung jarak *eucledian* data ke pusat klaster. Data akan dimasukkan ke dalam klaster dengan memilih nilai jarak *eucledian* yang terdekat (terkecil). Setelah seluruh data masuk ke dalam klaster, selanjutnya akan dihitung pusat klaster baru dengan cara menghitung rata-rata (mean) dari seluruh data dalam klaster tersebut dan menghitung jarak *eucledian*-nya. Selanjutnya akan dilihat apakah pusat klaster dan anggota klaster tetap. Jika mengalami perubahan maka dilakukan kembali perhitungan jarak data ke pusat klaster. Jika sudah tidak mengalami perubahan (*konvergen*), maka proses iterasi selesai dan didapatkan keanggotaan masing-masing data dengan klasternya.

3.3.2.2 Pengklasifikasian data dengan Naïve Bayes

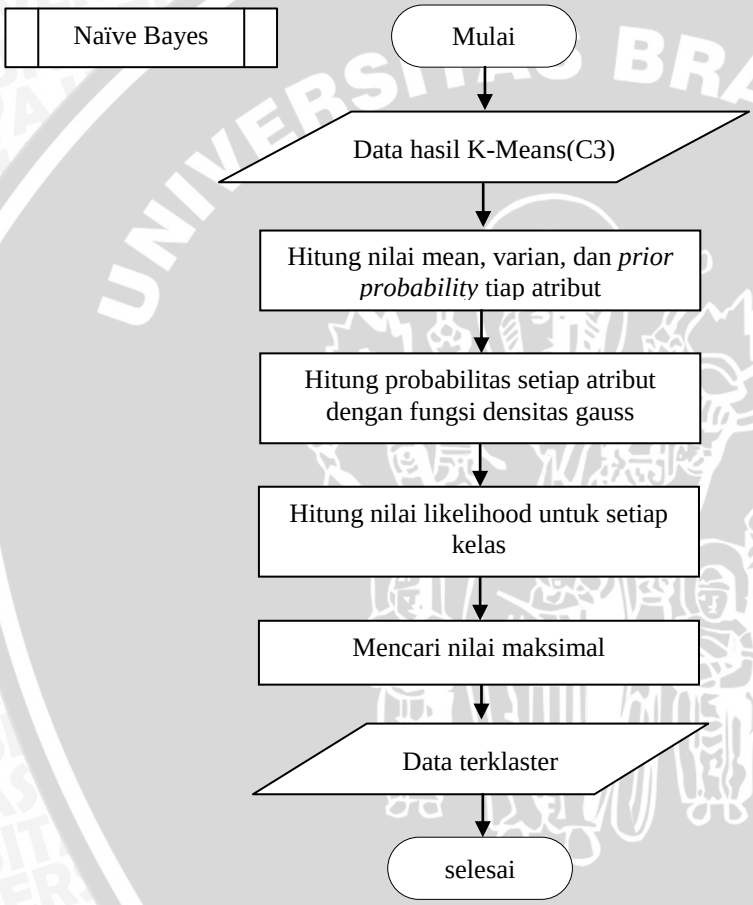
Model statistik merupakan suatu model yang efisien sebagai pendukung pengambilan keputusan. Konsep probabilistik merupakan salah satu bentuk model statistik. Salah satu metode yang menggunakan konsep probabilistik adalah Naïve Bayes. Algoritma Naïve Bayes adalah salah satu algoritma dalam teknik klasifikasi yang mudah diimplementasikan dan cepat prosesnya. Pada metode ini, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain.

Proses ini dijabarkan sebagi berikut :

1. Masukan (*input*) data.
2. Menghitunga rata-rata, varian dan *prior probability* dari setiap kelas untuk satu atribut data kanker payudara.

3. Menghitung probabilitas setiap atribut untuk semua kelas menggunakan perhitungan dengan fungsi densitas gauss.
4. Menghitung nilai *likelihood* untuk setiap kelas.
5. Mencari nilai maksimal untuk menentukan keanggotaan klaster.

Gambar 3.4 berikut adalah ilustrasi perhitungan dengan menggunakan algoritma Naïve Bayes.



Gambar 3.4 Alur sistem klasifikasi Naïve Bayes

Proses awal adalah menginputkan data latih, setelah itu menghitung nilai mean dan varian dari setiap atribut. Setelah mendapatkan nilainya, dilakukan perhitungan probabilitas setiap atribut dengan fungsi densitas gauss (persamaan 2-8) yang ada pada bab 2. Setelah itu hitung nilai *likelihood* untuk setiap kelas. Selanjutnya mencari nilai probabilitas yang paling maksimum untuk setiap kelas yang dihasilkan.

3.3.3 Perhitungan Manual

3.3.3.1 K-Means

Pada proses perhitungan K-Means diperlukan data training yang diambil dari dataset kanker payudara, dimana data diambil 9 atribut yang dibutuhkan untuk proses pengklasifikasian yaitu :

(A) *Clump thickness*

(B) *Uniformity of cell size*

(C) *Uniformity of cell shape*

(D) *Marginal adhesion*

(E) *Single epithelial cell size*

(F) *Bare nucle*

(G) *Bland chromatin*

(H) *Normal nucleoli.*

(I) *Mitoses*

Contoh dari data latih yang digunakan dalam proses K-Means seperti pada tabel 3.2.

Tabel 3.2 Contoh data latih kanker payudara

No.	A	B	C	D	E	F	G	H	I
x1	8	7	4	4	5	3	5	1	1
x2	10	1	7	8	7	1	1	1	3
x3	4	1	6	3	4	10	7	1	1
x4	5	4	6	8	4	1	8	1	1
x5	5	3	2	1	5	1	8	1	2
x6	5	2	1	3	2	1	2	4	1
x7	1	2	3	1	2	1	8	1	5
x8	3	2	7	1	2	1	4	1	1
x9	5	3	3	7	2	1	2	1	5
x10	3	7	1	8	2	2	1	3	1

Pada tabel 3.2 diberikan sejumlah data kanker payudara sejumlah 10 data (n = 10). Sedangkan atribut yang dimiliki adalah 9 atribut (m = 9). Dari sepuluh

data tersebut akan dikelompokkan menjadi 3 klaster (c = 3). Langkah pertama yaitu dilakukan proses pembakitan pusat klaster (*centroid*) awal yang digunakan dalam proses *clustering* di setiap atributnya. Pemilihan pusat klaster dilakukan secara acak, misalnya digunakan data yang ditunjukkan dengan tabel 3.3 :

Tabel 3.3 *Centroid* awal

Centroid awal	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
C1	5	2	1	3	2	1	2	4	1
C2	5	3	3	7	2	1	2	1	5
C3	10	1	7	8	7	1	1	1	3

Selanjutnya dilakukan perhitungan jarak masing-masing data dengan pusat klaster dengan menggunakan rumus jarak *euclidian* (persamaan 2-4).

$$d(x1,C1) = \sqrt{(8-5)^2 + (7-2)^2 + (4-1)^2 + (4-3)^2 + (5-2)^2 + (3-1)^2 + (5-2)^2 + (1-4)^2 + (1-1)^2}$$
$$= 8.6603$$

$$d(x1,C2) = \sqrt{(8-5)^2 + (7-3)^2 + (4-3)^2 + (4-7)^2 + (5-2)^2 + (3-1)^2 + (5-2)^2 + (1-1)^2 + (1-5)^2}$$
$$= 8.5440$$

$$d(x1,C3) = \sqrt{(8-10)^2 + (7-1)^2 + (4-7)^2 + (4-8)^2 + (5-7)^2 + (3-1)^2 + (5-1)^2 + (1-1)^2 + (1-3)^2}$$
$$= 9.6437$$

… …

$$d(x10,C1) = \sqrt{(3-5)^2 + (7-2)^2 + (1-1)^2 + (8-3)^2 + (2-2)^2 + (2-1)^2 + (1-2)^2 + (3-4)^2 + (1-1)^2}$$
$$= 7.5498$$

$$d(x10,C2) = \sqrt{(3-5)^2 + (7-3)^2 + (1-3)^2 + (8-7)^2 + (2-2)^2 + (2-1)^2 + (1-2)^2 + (3-1)^2 + (1-5)^2}$$
$$= 6.8557$$

$$d(x10,C3) = \sqrt{(3-10)^2 + (7-1)^2 + (1-7)^2 + (8-8)^2 + (2-7)^2 + (2-1)^2 + (1-1)^2 + (3-1)^2 + (1-3)^2}$$
$$= 12.4499$$

Berikut adalah tabel 3.4 hasil keanggotaan tiap klaster dari perhitungan jarak *euclidian* dari iterasi ke-1 menggunakan 3 pusat klaster.

Tabel 3.4 Hasil perhitungan jarak *eucledian* iterasi ke-1.

Data	C1	C2	C3	Klaster
X1	8.6603	8.5440	9.6437	C2
X2	11.2250	8.7178	0.0000	C3
X3	12.0830	12.4900	13.8564	C1
X4	10.1489	8.1854	9.8489	C2
X5	7.8102	9.5394	12.5300	C1
X6	0.0000	6.7823	11.2250	C1
X7	9.2195	9.4340	15.0000	C1
X8	7.5498	8.7750	11.7047	C1
X9	6.7823	0.0000	8.7178	C2
X10	7.5498	6.8557	12.4499	C2

Dari tabel 3.4 didapatkan jumlah anggota C1 = 5 (x3,x5,x6,x7,x8), C2 = 4 (x1,x4,x9,x10), dan C3 = 1 (x2). Penentuan untuk keanggotaan klaster diketahui dari membandingkan hasil perhitungan jarak di setiap klaster dan diambil dengan jarak terkecil.Misalnya :

jarak x ke C1 = 10,232

jarak x ke C2 = 11,892

jarak x ke C3 = 9,254 (terkecil)

x= C3

maka x masuk dalam keanggotaan klaster 3 (C3).

Setelah mendapatkan seluruh anggota tiap klaster, maka perhitungan untuk pusat klaster baru dari iterasi ke-1 dilakukan. Dengan cara menghitung mean data anggota klaster dari setiap klaster. Misal data yang dipilih adalah atribut *clump thickness*, berikut adalah contoh perhitungannya :

$$C1=\frac{4+5+5+1+3}{5} = 3,6$$

$$C2=\frac{8+4+5+3}{2} = 5,25$$

$$C3=\frac{5}{1} = 5$$

Hasil perhitungan pusat klaster baru ditunjukkan dengan tabel 3.5.

Tabel 3.5 Klaster baru iterasi ke-2

Centroid baru	A	B	C	D	E	F	G	H	I
C1	3.60	2.00	3.80	1.80	3.00	2.80	5.80	1.60	2.00
C2	5.25	5.25	3.50	6.75	3.25	1.75	4.00	1.50	2.00
C3	10.00	1.00	7.00	8.00	7.00	1.00	1.00	1.00	3.00

Selanjutnya melakukan iterasi ke-2 dengan pusat klaster baru.Berikut adalah hasil perhitungan jarak *euclidian* yang ditujukkan dalam tabel 3.6.

Tabel 3.6 Hasil perhitungan jarak *eucledian* iterasi ke-2

Data	C1	C2	C3	Klaster
X1	8.6603	8.5440	9.6437	C2
X2	11.2250	8.7178	0.0000	C3
X3	12.0830	12.4900	13.8564	C1
X4	10.1489	8.1854	9.8489	C2
X5	7.8102	9.5394	12.5300	C1
X6	0.0000	6.7823	11.2250	C1
X7	9.2195	9.4340	15.0000	C1
X8	7.5498	8.7750	11.7047	C1
X9	6.7823	0.0000	8.7178	C2
X10	7.5498	6.8557	12.4499	C2

Dari tabel 3.6 didapatkan jumlah anggota C1 = 5 (x3,x5,x6,x7,x8), C2 = 4 (x1,x4,x9,x10), dan C3 = 1 (x2). *Centroid* iterasi 1 dan 2 belum sama maka dilakukan iterasi ke-3. Dengan menggunakan rumus rata-rata dilakukan *update* terhadap *centroid.* Tabel hasil *update centroid* dapat dilihat pada tabel 3.7.

Tabel 3.7 Klaster baru iterasi ke-3

Centroid baru	A	B	C	D	E	F	G	H	I
C1	3.60	2.00	3.80	1.80	3.00	2.80	5.80	1.60	2.00
C2	5.25	5.25	3.50	6.75	3.25	1.75	4.00	1.50	2.00
C3	10.00	1.00	7.00	8.00	7.00	1.00	1.00	1.00	3.00

Selanjutnya melakukan iterasi ke-3 dengan pusat klaster baru.Berikut adalah hasil perhitungan jarak *euclidian* yang ditujukkan dalam tabel 3.8.

Tabel 3.8 Hasil perhitungan jarak *eucledian* iterasi ke-3

Data	C1	C2	C3	Klaster
X1	8.6603	8.5440	9.6437	C2
X2	11.2250	8.7178	0.0000	C3
X3	12.0830	12.4900	13.8564	C1
X4	10.1489	8.1854	9.8489	C2
X5	7.8102	9.5394	12.5300	C1
X6	0.0000	6.7823	11.2250	C1
X7	9.2195	9.4340	15.0000	C1
X8	7.5498	8.7750	11.7047	C1
X9	6.7823	0.0000	8.7178	C2
X10	7.5498	6.8557	12.4499	C2

Tabel 3.8 menunjukkan iterasi ke-3 dengan pusat klaster yang telah diupdate tidak mengalami perubahan (*konvergen*), maka dengan kata lain keanggotaan klaster pada iterasi ke-2 dan iterasi ke-3 adalah sama. Karena tidak mengalami perubahan anggota klaster, maka perhitungan tidak perlu dilakukan lagi dan proses iterasi telah berakhir.

3.3.3.2 Naïve Bayes

Cara kerja dari proses perhitungan Naïve Bayes adalah sebagai berikut, tahapan diawali dengan melakukan pengambilan data sampel atau data latih dari data yang telah melalui proses perhitungan dengan K-Means. Disini diasumsikan akan dihitung data C2 yang didefinisikan merupakan data ‘mungkin’. Sehingga data yang akan digunakan dalam pengklasifikasian dapat dilihat pada tabel 3.9.

Tabel 3.9 Data C2

No.	A	B	C	D	E	F	G	H	I	Kelas
x1	8	7	4	4	5	3	5	1	1	4
x2	10	1	7	8	7	1	1	1	3	4
x3	5	3	3	7	2	1	2	1	5	2
x4	3	7	1	8	2	2	1	3	1	2

Kemudian dari tabel 3.9 dapat dihitung nilai rata-rata dan varian dari masing-masing kelas. Contoh perhitungannya adalah sebagai berikut :

1. Perhitungan rata-rata masing-masing kelas.

$\mu A,C1 = \frac{8+10}{2} = 9$

$\mu A,C2 = \frac{5+3}{2} = 4$

…… ……

$\mu I,C1 = \frac{1+3}{2} = 2$

$\mu I,C2 = \frac{5+1}{2} = 3$

Berikut adalah tabel 3.10 hasil rata-rata.

Tabel 3.10 Rata-rata Naïve Bayes

Klaster	**Rata-rata**								
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
C1	9.0	4.0	5.5	6.0	6.0	2.0	3.0	1.0	2.0
C2	4.0	5.0	2.0	7.5	2.0	1.5	1.5	2.0	3.0

2. Perhitungan varian masing-masing kelas.

$\sigma^2 A,C1 = \frac{(8-9)^2+(10-9)^2}{2} = 2$

$\sigma^2 A,C2 = \frac{(8-4)^2+(10-4)^2}{2} = 1$

… …

$\sigma^2 I,C1 = \frac{(1-9)^2+(3-9)^2}{2} = 0,1$

$\sigma^2 I,C2 = \frac{(5-4)^2+(1-4)^2}{2} = 4$

Berikut adalah tabel 3.11 hasil varian.

Tabel 3.11 varian Naïve Bayes

Klaster	**Varian**								
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
C1	2	0,1	0,1	0,1	0,2	0,1	3,1	0,1	0,1
C2	1.0	4.0	1.0	0.3	0.0	0.3	0.3	1.0	4.0

3. Menghitung nilai *prior probability* untuk masing-masing kelas dengan menggunakan rumus :

$$P(Ci) = \frac{\text{jumlah data pada masing-masing kelas}}{\text{total jumlah data}}$$

$P(C1) = \frac{2}{4} = 0.5$ $P(C2) = \frac{2}{4} = 0.5$

4. Menghitung nilai *probability* dengan menggunakan rumus distribusi gauss. Perhitungan manual distribusi gauss adalah sebagai berikut :

$$P(Ax1|C1) = \frac{1}{\sqrt{2(3{,}14)(9)}} e^{-\frac{(8-9)}{2(1)}} = 0.282166$$

$$P(Ax1|C2) = \frac{1}{\sqrt{2(3{,}14)(4)}} e^{-\frac{(8-4)}{2(1)}} = 0.103803$$

… …

$$P(Ix6|C1) = \frac{1}{\sqrt{2(3{,}14)(9)}} e^{-\frac{(3-9)}{2(1)}} = 0.103803$$

$$P(Ix6|C2) = \frac{1}{\sqrt{2(3{,}14)(4)}} e^{-\frac{(3-4)}{2(1)}} = 0.103803$$

Berikut adalah tabel 3.12 dan tabel 3.13 hasil perhitungan distribusi gauss dari data yang akan dicari klasternya.

Tabel 3.12 Perhitungan distribusi gauss C1

data	P(Ai|C1)	P(Bi|C1)	P(Ci|C1)	P(Di|C1)	P(Ei|C1)	P(Fi|C1)	P(Gi|C1)	P(Hi|C1)	P(Ii|C1)	P(Ai|C1)
x1	2.42E-01	5.40E-02	1.49E-06	1.49E-06	1.34E-04	6.08E-09	1.34E-04	5.05E-15	5.05E-15	2.42E-01
x2	2.42E-01	5.05E-15	5.40E-02	2.42E-01	5.40E-02	5.05E-15	5.05E-15	5.05E-15	6.08E-09	2.42E-01
x3	1.34E-04	6.08E-09	6.08E-09	5.40E-02	9.14E-12	5.05E-15	9.14E-12	5.05E-15	1.34E-04	1.34E-04
x4	6.08E-09	5.40E-02	5.05E-15	2.42E-01	9.14E-12	9.14E-12	5.05E-15	6.08E-09	5.05E-15	6.08E-09

Tabel 3.13 Perhitungan distribusi gauss C2

data	P(Ai|C1)	P(Bi|C1)	P(Ci|C1)	P(Di|C1)	P(Ei|C1)	P(Fi|C1)	P(Gi|C1)	P(Hi|C1)	P(Ii|C1)	P(Ai|C1)
x1	1.34E-04	4.43E-03	3.99E-01	3.99E-01	2.42E-01	2.42E-01	2.42E-01	4.43E-03	4.43E-03	1.34E-04
x2	6.08E-09	4.43E-03	4.43E-03	1.34E-04	4.43E-03	4.43E-03	4.43E-03	4.43E-03	2.42E-01	6.08E-09
x3	2.42E-01	2.42E-01	2.42E-01	4.43E-03	5.40E-02	4.43E-03	5.40E-02	4.43E-03	2.42E-01	2.42E-01
x4	2.42E-01	4.43E-03	4.43E-03	1.34E-04	5.40E-02	5.40E-02	4.43E-03	2.42E-01	4.43E-03	2.42E-01

5. Menghitung nilai *likelihood* dengan cara mengalikan nilai atribut yang didapatkan dari perhitungan distribusi gauss.

P(C1|x1A,..,x1I) = 2.42E-01*5.40E-02*1.49E-06*1.49E-06*1.34E-04 *6.08E-09*1.34E-04*5.05E-15*5.05E-15
= 8.03941E-59

P(C2|x1A,..,x1I) = 2.42E-01*5.05E-15*5.40E-02*2.42E-01*5.40E-02 *5.05E-15*5.05E-15*5.05E-15*6.08E-09
= 2.63274E-14

……. ….....

P(C1|x4A,..,x4I) = 6.08E-09*5.40E-02*5.05E-15*2.42E-01* 9.14E-12*9.14E-12*5.05E-15*6.08E-09*5.05E-15
= 5.20169E-84

P(C2|x4A,..,x4I) = 2.42E-01*4.43E-03*4.43E-03*1.34E-04*5.40E-02*5.40E-02*4.43E-03*2.42E-01*4.43E-03
= 8.83187E-18

6. Mendapatkan nilai maksimum.

   Dari hasil perhitungan *likelihood* akan didapatkan nilai yang akan dibandingkan antar klaster dan diambil nilai tertinggi dari semua peluang klaster. Berikut adalah tabel yang didapatkan dari perhitungan *likelihood* serta penentuan anggota klaster .

Tabel 3.14 Mendapatkan nilai maksimum

Data	P(C1)	P(C2)	Klaster
x1	8.03941E-59	2.63274E-14	C2
x2	6.77197E-70	1.49426E-27	C2
x3	7.62084E-77	8.71845E-13	C2
x4	5.20169E-84	8.83187E-18	C2

Dari tabel 3.14 dapat dilihat data telah menjadi anggota klaster C1.Dengan masuknya data‘mungkin’ke dalam suatu klaster, maka *clustering* untuk seluruh data telah selesai.

3.4 Perancangan Antarmuka (*interface*)

Graphical User Interface atau yang sering disebut sebagai GUI adalah tampilan dari program yang bisa dinikmati oleh user. Perancangan *User Interface* harus dibuat semenarik dan seindah mungkin dengan tetap mengutamakan kenyamanan dalam mengoperasikan program (*user friendly*). Tampilan User *Interface* dituangkan dalam sebuah web site yang dibangun dengan menggunakan bahasa pemrograman php. Dalam proses perancangan ini, tampilan awal web berisi menu Home, Deskripsi sistem, Tentang K-Means dan Naïve Bayes, Galeri Kanker Payudara, dan Aplikasi. Pada menu Aplikasi terdapat beberapa sub menu yaitu, Pelatihan KMNB, Pengujian KMNB, K-Means dan Keluar aplikasi. Berikut adalah contoh tampilan dari web aplikasi yang akan dibangun.

a. Tampilan utama sistem.

Tampilan utama sistem dapat dilihat pada gambar 3.5.

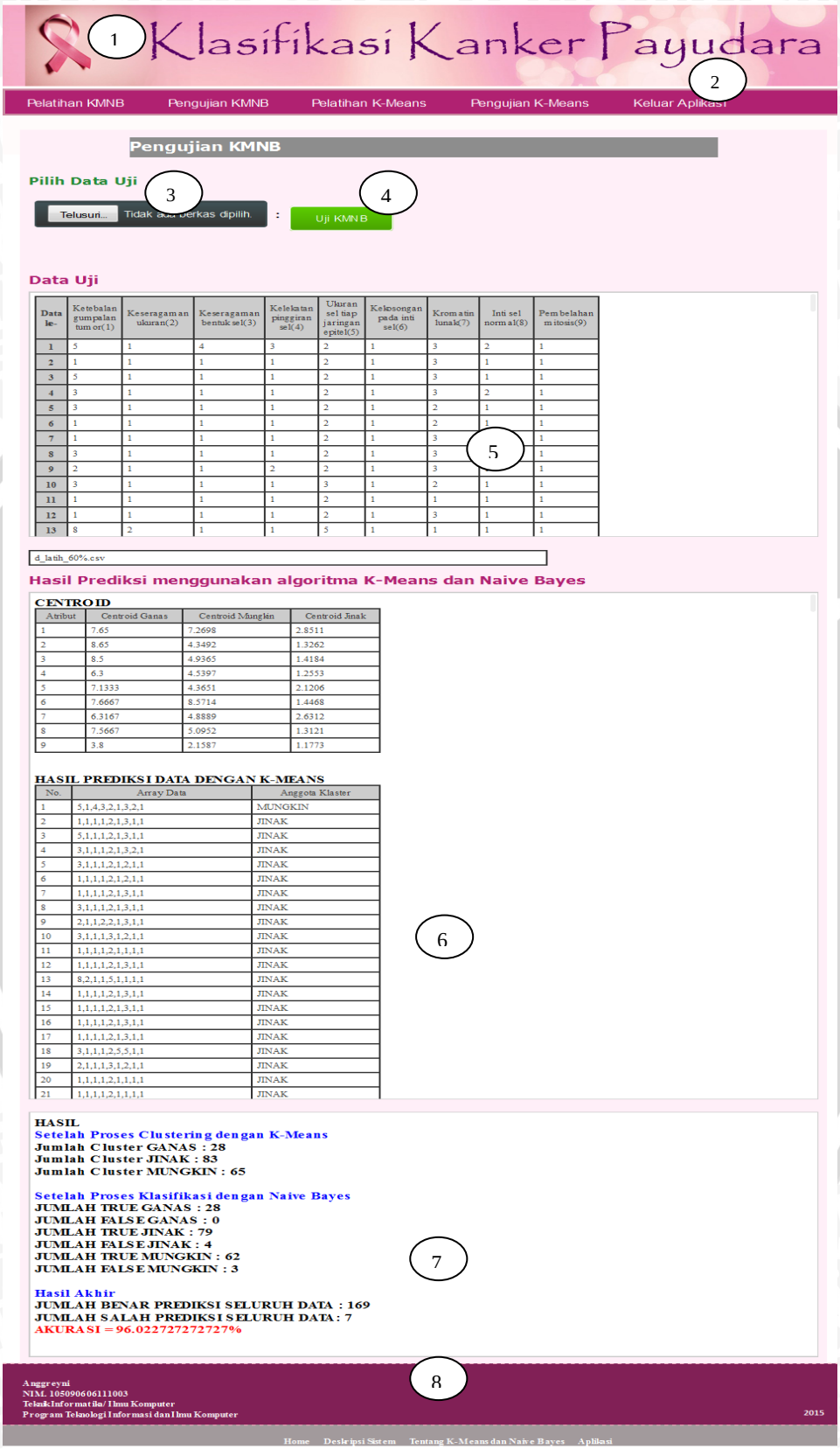
Gambar 3.5 Tampilan awal sistem.

Keterangan :

1. *Header* sebagai judul mengenai sistem klasifikasi kanker payudara.
2. Tab menu untuk navigasi memilih *Home*, deskripsi sistem, tentang KMNB dan aplikasi K-Means dan Naïve Bayes.
3. Merupakan informasi umum tentang penelitian.
4. Merupakan informasi singkat mengenai kanker payudara dan penggunaan metode gabungan K-Means dan Naïve Bayes dalam upaya menambah informasi.
5. *Footer* yang berisi sedikit informasi mengenai peneliti.

b. Tampilan halaman menu Aplikasi.

Tampilan halaman aplikasi K-Means dan Naïve Bayes dapat dilihat pada gambar 3.6.



Data ke-	Ketebalan gumpalan tumor(1)	Keseragaman ukuran(2)	Keseragaman bentuk sel(3)	Kelekatan panggiran sel(4)	Ukuran sel tiap jaringan epitel(5)	Kelosongan pada inti sel(6)	Kromatin lunak(7)	Inti sel normal(8)	Pembelahan mitosis(9)
1	5	1	4	3	2	1	3	2	1
2	1	1	1	1	2	1	3	1	1
3	5	1	1	1	2	1	3	1	1
4	3	1	1	1	2	1	3	2	1
5	3	1	1	1	2	1	2	1	1
6	1	1	1	1	2	1	2	1	1
7	1	1	1	1	2	1	3	1	1
8	3	1	1	1	2	1	3	1	1
9	2	1	1	2	2	1	3	1	1
10	3	1	1	1	3	1	2	1	1
11	1	1	1	1	2	1	1	1	1
12	1	1	1	1	2	1	3	1	1
13	8	2	1	1	5	1	1	1	1

Atribut	Centroid Ganas	Centroid Mungkin	Centroid Jinak
1	7.65	7.2698	2.8511
2	8.65	4.3492	1.3262
3	8.5	4.9365	1.4184
4	6.3	4.5397	1.2553
5	7.1333	4.3651	2.1206
6	7.6667	8.5714	1.4468
7	6.3167	4.8889	2.6312
8	7.5667	5.0952	1.3121
9	3.8	2.1587	1.1773

No.	Array Data	Anggota Klaster
1	5,1,4,3,2,1,3,2,1	MUNGKIN
2	1,1,1,1,2,1,3,1,1	JINAK
3	5,1,1,1,2,1,3,1,1	JINAK
4	3,1,1,1,2,1,3,2,1	JINAK
5	3,1,1,1,2,1,2,1,1	JINAK
6	1,1,1,1,2,1,2,1,1	JINAK
7	1,1,1,1,2,1,3,1,1	JINAK
8	3,1,1,1,2,1,3,1,1	JINAK
9	2,1,1,2,2,1,3,1,1	JINAK
10	3,1,1,1,3,1,2,1,1	JINAK
11	1,1,1,1,2,1,1,1,1	JINAK
12	1,1,1,1,2,1,3,1,1	JINAK
13	8,2,1,1,5,1,1,1,1	JINAK
14	1,1,1,1,2,1,3,1,1	JINAK
15	1,1,1,1,2,1,3,1,1	JINAK
16	1,1,1,1,2,1,3,1,1	JINAK
17	1,1,1,1,2,1,3,1,1	JINAK
18	3,1,1,1,2,5,5,1,1	JINAK
19	2,1,1,1,3,1,2,1,1	JINAK
20	1,1,1,1,2,1,1,1,1	JINAK
21	1,1,1,1,2,1,1,1,1	JINAK

Gambar 3.6 Tampilan aplikasi K-Means dan Naïve Bayes.

Keterangan :

1. *Header* sebagai judul mengenai sistem klasifikasi kanker payudara.
2. Tab Tab menu untuk navigasi memilih pelatihan KMNB, Pengujian KMNB, K-Means dan Keluar aplikasi.
3. *Text field* dan tombol *choose file* untuk menampilkan nama file dan memilih data latih yang di akan digunakan.
4. Tombol untuk memulai proses pelatihan/pengujian.
5. Text area untuk tampilan dari data inputan yang telah dipilih sebelumnya.
6. Text area untuk tampilan dari hasil *clustering* K-Means dan klasifikasi Naïve Bayes.
7. Text area untuk tampilan dari hasil akhir akurasi *clustering* K-Means dan klasifikasi Naïve Bayes.
8. *Footer* yang berisi sedikit informasi mengenai peneliti.

3.5 Perancangan Pengujian

Pengujian pada penelitian ini dilakukan untuk menunjukkan tingkat keberhasilan dari metode klasifikasi yang diuji. Proses pengukuran keberhasilan sistem dilakukan dengan menggunakan akurasi sistem. Selanjutnya, dilakukan beberapa skenario uji coba.

Skenario pengujian awal, dimana pada tahap ini data diklasifikasikan menggunakan algoritma K-Means dan Naïve Bayes. Pada tahap ini nantinya K-Means akan menghasilkan tiga klaster yaitu : jinak, ganas, dan mungkin. Selanjutnya yang belum masuk kategori (mungkin) akan dihitung menggunakan Naïve Bayes. Selanjutnya pada tahap kedua adalah pengujian menggunakan data uji. Pada tahap ini pengujian menggunakan rule yang terbentuk dari pengujian pada tahap pertama. Pengujian ini dilakukan untuk mengetahui ketetapan rule yang dibuat dalam menganalisa data uji, sehingga pengklasifikasian terhadap data uji dapat dilakukan dengan benar. Pengujian dilakukan untuk mengukur tingkat akurasi algoritma yang digunakan pada proses klasifikasi terhadap data penyakit kanker payudara. Perancangan uji coba sistem dapat dilihat pada tabel 3.15.

Tabel 3.15 Perancangan uji coba sistem

Jumlah data latih (%)	Jumlah data latih	Jumlah data uji (%)	Rata-rata Akurasi dari Percobaan 1 sampai 5	Akurasi Tertinggi dari Percobaan 1 sampai 5
60%		40%		
70%		30%		
80%		20%		

Dapat dilihat pada tabel 3.15 Tahap awal perhitungan menggunakan data latih sebesar 60% dari jumlah seluruh data. Rule yang terbentuk dari data latih tersebut akan diujikan terhadap 40% data uji. Tahap selanjutnya, melakukan perhitungan yang sama seperti dengan tahap awal, namun di tahap kedua menggunakan 70% data latih dan diujikan terhadap 30% data uji. Selanjutnya, dilakukan perhitungan dengan data latih 80% dan data uji 20%.

Setelah melakukan perhitungan akan didapatkan nilai akurasi masing-masing perhitungan, kemudian dihitung rata-rata dari seluruh nilai akurasi dan didapatkan nilai akurasi tertinggi.