Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Klasifikasi Data Pasien Breast Cancer Menggunakan Metode Gaussian Naïve Bayes

Classification of Breast Cancer Patient Data Using the Gaussian Naïve Bayes Method

Rahmat Yulianto[1)], Deni Arifianto[2)]
[1]Mahasiswa Program Studi Teknik Informatika, Fakultaas Teknik, Universitas Muhammadiyah Jember
email: sahabatyulianto60@gmail.com
[2]Dosen Program Studi Teknik Informatika, Fakultaas Teknik, Universitas Muhammadiyah Jember
email: deniarifianto@unmuhjember@ac.id

Abstrak

Kanker payudara merupakan tumor ganas yang menyerang jaringan sel-sel payudara. Kanker payudara merupakan masalah paling besar bagi wanita di seluruh dunia dan menyebabkan kematian utama bagi penderita kanker payudara. penyakit kanker payudara di negara berkembang menunjukkan bahwa penyakit kanker dengan persentase kasus tertinggi, kurang lebih 43% kasus dan persentase kematian yaitu 12,9%. Pesatnya perkembangan teknologi sekarang ini memungkinkan untuk mendeteksi suatu penyakit kanker payudara dengan menggunakan teknik penalaran Soft computing. Salah satu teknik data mining yang dipakai dalam penelitian terhadap dunia kesehatan adalah klasifikasi. Penelitian ini berisi tentang pengukuran metode Gaussian Naive Bayes terhadap klasifikasi penyakit breast cancer terhadap pasien. Data yang digunakan berasal dari UCI Machine Learning dengan total data 116 pasien dengan partisi 80 data digunakan pada skenario uji dan 36 data digunakan pada uji validasi. Penelitian ini menggunakan skenario uji K Fold Cross Validation dengan nilai k = 2,4,5, 8 dan 10. Pada penelitian ini diperoleh hasil akurasi tertinggi yaitu 62,5% yang dihasilkan pada pengujian 10 Fold skenario 5 dan 9 dan presisi yang sama diperoleh yaitu 62,5%. Model pada 10 Fold skenario 5 dan 9 merupakan skenario terbaik maka keduanya akan diuji validasi. Hasil uji validasi menunjukkan pada 10 Fold skenario 5 dan 9 memperoleh hasil akurasi sebesar 61,11% dan presisi sebesar 56,66%

Kata Kunci : Data Mining, Klasifikasi, *Breast Cancer*, *Gaussian Naive Bayes.*

Abstract

Breast cancer is a malignant tumor that attacks the breast tissue cells. Breast cancer is the biggest problem for women around the world and the main cause of death for breast cancer sufferers. Breast cancer in developing countries shows that cancer has the highest percentage of cases, approximately 43% of cases and the percentage of deaths is 12.9%. The rapid development of technology today makes it possible to detect a breast cancer by using soft computing reasoning techniques. One of the data mining techniques used in research on the world of health is classification. This study contains the measurement of the Gaussian Naive Bayes method for the classification of breast cancer in patients. The data used comes from UCI Machine Learning with a total data of 116 patients with 80 data partitions used in the test scenario and 36 data used in the validation test. This study uses the K Fold Cross Validation test scenario with k values = 2, 4, 5, 8, and 10. In this study, the highest accuracy results were 62.5% which was produced in the 10 Fold test scenario 5 and 9 and the same precision was obtained. ie 62.5%. The model in 10 Fold scenarios 5 and 9 is the best scenario, so both will be validated. The results of the validation test show that in 10 Fold scenarios 5 and 9, the accuracy is 61.11% and the precision is 56.66%.

Keywords*: Data Mining, Classification, Breast Cancer, Gaussian Naive Bayes.*

Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

1. Pendahuluan

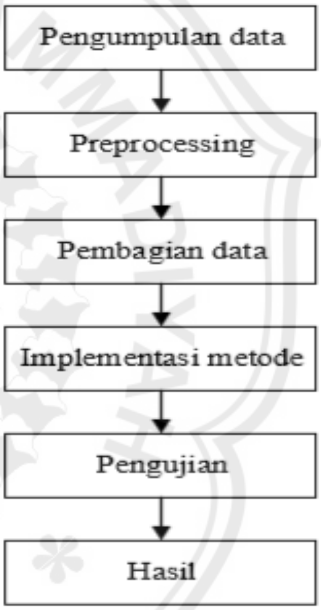
Kanker payudara merupakan tumor ganas yang menyerang jaringan sel-sel payudara. Kanker payudara merupakan masalah paling besar bagi wanita di seluruh dunia dan menyebabkan kematian utama bagi penderita kanker payudara. penyakit kanker payudara di negara berkembang menunjukkan bahwa penyakit kanker dengan persentase kasus tertinggi, kurang lebih 43% kasus dan persentase kematian yaitu 12,9%. Menurut WHO sekitar 8-9% wanita menderita penyakit kanker payudara. Kasus kanker payudara terus meningkat lebih dari 250,000 kasus baru, di Eropa dilakukan penelitian kanker payudara oleh American Cancer Society(ACS) hampir 178.000 wanita yang telah di diagnosis kanker payudara dan jumlah tersebut ditambah 2 juta wanita yang memiliki riwayat penyakit ini (American Cancer Society Inc, 2012). Penyakit kanker payudara masih menjadi masalah utama dalam dunia kesehatan, dibuktikan dari berbagai kasus komplikasi fisik fungsional dan dapat juga menyebabkan gangguan kualitas hidup. Penurunan kualitas hidup wanita penderita kanker payudara dapat dilihat dari sisi kesehatan fisik, status psikologi, hubungan sosial, kemandirian dan spiritual. Kualitas hidup merupakan persepsi individu dalam kemampuan, keterbatasan psikologi dalam konteks budaya dan system nilai untuk mengetahui peran dan fungsi (Rabin, E. G et al, 2008). Bahaya dari penyakit ini bukan hanya berdampak pada fisik pada penderita melainkan pada sisi psikologi juga. Keterlambatan mengetahui akan merambatnya penyakit ini adalah salah satu faktor yang paling sering dialami. Ketidaktahuan informasi penanganan sejak dini terhadap tanda-tanda kemunculan penyakit ini menyebabkan penderita harus mengalami operasi tahu bahkan meninggal dunia.

Pesatnya perkembangan teknologi sekarang ini memungkinkan untuk mendeteksi suatu penyakit kanker payudara dengan menggunakan teknik penalaran Soft computing. Soft computing adalah suatu model pendekatan untuk melakukan komputasi dengan meniru akal manusia dan memiliki kemampuan untuk menalar dan belajar pada lingkungan yang penuh dengan ketidakpastian (Kusumadewi S & Hartati S, 2006).

Salah satu teknik data mining yang dipakai dalam penelitian terhadap dunia kesehatan adalah klasifikasi. Diantaranya ada Decicion tree, K-NN, Bayesian dan lain-lain. Berikut ini beberapa penelitian terhadap kanker payudara menggunakan metode klasifikasi. Menurut Uma Ojha dalam penelitiannya A study on prediction of breast cancer recurrence using data mining techniques menggunakan metode

2. Tinjauan Pustaka

Tahapan penelitian di sini akan digambar dalam sebuah bagan yang mencakup dari awal sampai akhir penelitian.



Gambar 1 Diagram Aliran Penilitihan
Sumber : Penilitihan *Diagram Flowchat*

A. Pengumpulan Data

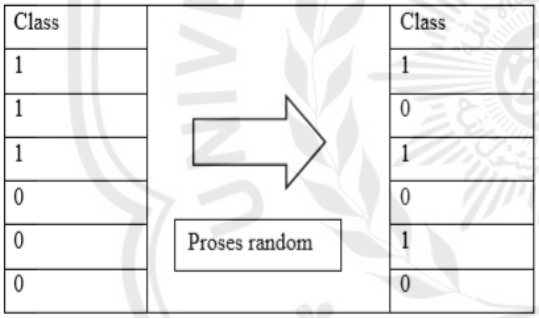
Tahap awal dalam penelitian ini pengumpulan data. Data yang digunakan dalam penelitian ini adalah data penderita kanker payudara. Peneliti menggunakan data yang berasal dari UCI Machine Learning. Dataset ini berasal dari Wisconsin Breast Cancer Dataset (WBCD), Wisconsin Diagnosis Breast Cancer (WDBC) dan the Wisconsin Prognosis Breast Cancer (WPBC) serta dipublikasikan oleh Wisconsin Breast Cancer Dataset (WBCD) pada tahun 2018. Atribut yang digunakan oleh

Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Wisconsin Breast Cancer Dataset (WBCD) dalam penelitiannya terhadap penderita kanker payudara adalah Age, Body mass indeks (BMI), Glucose, Insulin, Homeostasis Model Assessment (HOMA), Leptin, Adiponectin, Resistin dan MCP-1 dengan output class Healthy controls dan Patients.

B. Preprocessing

Tahap selanjutnya adalah Preprocessing, data sebelum diolah ke tahap implementasi metode diproses dengan preprocessing agar saat implementasi nantinya tidak ada hasil yang error. Preprocessing yang digunakan adalah Random data. Random dilakukan apabila data yang tersusun memiliki output yang rapi. Misal, pada sebuah dataset yang berjumlah 10 data dengan output 0 dan 1. Data 1 – 5 output 0 dan data 6 – 10 output 1, hal ini sangat tidak berimbang ketika data akan uji menggunakan skenario k-fold cross validation. Metode random data pada tahap preprocessing diolah menggunakan Ms.Excel dengan perintah formula RAND(). Berikut gambaran random data.



Gambar 2 Proses Random Data
Sumber : penilitan sendiri

C. Pembagian Data

Pembagian data di sini juga mencakup skenario pengujian, dimana pada penelitian ini penulis menggunakan k-fold cross validation untuk pembagian data uji dan data latih.

D. Implementasi Metode

Pada implementasi metode terdapat dua tahap yaitu pembentukan model dan pengujian model terhadap data uji. Berikut tahapannya.

Menentukan *prior probability* masing-masing *output* pada data latih dengan persamaan

$$Prior\ probability\ C1 = \frac{\sum kemunculan\ data\ A\ pada\ atribut\ X}{\sum data\ X}$$

Pada implementasi metode terdapat dua tahap yaitu pembentukan model dan pengujian model terhadap data uji. Berikut tahapannya.

$$mean(x) = \frac{1}{n}\ x\ \sum_{i=1}^{n} x_i$$

Dimana:
(x) = Rata rata
n = Jumlah seluruh frekuensi
$\sum_{i=1}^{n} x_i$ = Jumlah seluruh nilai data

$$\text{StandardDeviation(x)} = \sqrt{\frac{1}{\text{n}}\ \text{x}\ \sum_{\text{i=1}}^{\text{n}} (\text{x}_\text{i}\text{-mean(x)})^2}$$

Dimana:
(x) = Rata rata
n = Jumlah seluruh frekuensi
$\sum_{i=1}^{n} x_i$ = Jumlah seluruh nilai data

E. Pengujian Data Uji

Model yang telah didapatkan dari data latih akan diuji pada data uji untuk mengukur tingkat keberhasilan dan ketepatan. Menggunakan persamaan pada bab 2 yaitu.

$$pdf(x, mean.sd)\ \frac{1}{\sqrt{2\ x\ \pi}\ x\ sd}\ x\ e^{-(\frac{(x-mean)^2}{2xsd^2})}$$

Dimana:
(x) = Rata rata
n = Jumlah seluruh frekuensi
$\sum_{i=1}^{n} x_i$ = Jumlah seluruh nilai data
sd = Hasil dari *Standard Deviation*

F. Pengkreteriaan

Setelah pengujian terhadap data uji selanjutnya pengujian terhadap kriteria hasil perbandingan klasifikasi dan nilai asli. Dalam penelitian ini penulis menggunakan metode confusion matrix, merujuk pada persamaan pada dalam menentukan kriteria yaitu:

Aktual	Klasifikasi	
	Yes	No
Yes	*True positive*	*False negative*
No	*False positive*	*True negative*

Gambar 3 Klasifikasi
Sumber : Gorunescu F (2011)

Tahap selanjutnya adalah pengukuran hasil klasifikasi dengan menghitung kriteria yang dihasilkan. Dalam penelitian ini pengukuran difokuskan pada akurasi dan presisi. Merujuk pada persamaan dalam mengukur akurasi dan presisi yaitu.

Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

$$akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Dimana:
TP = *True positif*
TN = *True negatif*
FP = *False positif*
FN = *False negatif*

$$presisi = \frac{TP}{(TP+FP)}$$

Dimana:
TP = *True positif*
TN = *True negatif*
FP = *False positif*

3. Hasil Pembahasan

A. Random Data

Kondisi data pada saat diunduh atau data asli dari penelitian ini diketahui masih terklaster atau terkelompokkan berdasarkan *output class*. Hal ini memiliki kondisi yang tidak baik karena skenario uji pada penelitian ini menggunakan metode K *Fold Cross Validation.*

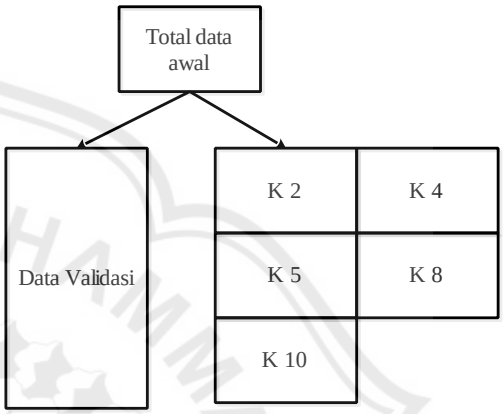
Tabel 1 Hasil Random Data

Sebelum *random* data		Setelah *random* data	
No.	*Class*ification	No.	*Class*ification
1	1	92	2
2	1	86	2
3	1	71	2
4	1	91	2
5	1	34	1
...	...	...	...
48	1	46	1
49	1	77	2
50	1	19	1
51	1	1	1
52	1	94	2
53	2	48	1
54	2	55	2
55	2	76	2
56	2	41	1
57	2	85	2
...	...	...	...
112	2	98	2
113	2	69	2
114	2	17	1
115	2	10	1
116	2	20	1

Sumber : Hasil Random

B. Partisi Data

Preprocessing selanjutnya adalah partisi data. Partisi data pada tahap ini adalah membagi data menjadi dua bagian yaitu data untuk K *Fold Cross Validation* data untuk Validasi.



Gambar 4 Hasil Partisi Data
Sumber : Hasil Penilitihan

Proses pemotongan data dilakukan menggunakan Ms.Excel dengan *range* pemotongan yaitu data 1 sampai 80 adalah data untuk K *Fold Cross Validation* dan data 81 sampai 116 adalah data *validation.*

C. Skenario Uji

Skenario uji pada penelitian ini meliputi 2, 4, 5, 8 dan 10 *Fold Cross Validation*.

Tahap selanjutnya adalah menghitung *prior* probabilitas *mean* dan standar deviasi tiap-tiap atribut.

a. *Prior Probability*

Untuk menghitung *prior* probabilitas adalah $P\ (X, A) = \frac{P|X}{P|A}$ dimana, X adalah jumlah *output* = 1 dan A =total data. Jadi $P\ (1) = \frac{5}{10}$ yaitu 0,5. Hal ini dilakukan juga pada probabilitas kemunculan ouput = 2 yaitu $P\ (2) = \frac{5}{10}$ yaitu 0,5. Pada *prior* probabilitas ini diketahui nilai *prior* probabilitas *output* 1 = 0,5 dan *output* 2 = 0,5.

b. *Mean*

Pertama hitung *mean* setiap atribut. Berikut dicontohkan menghitung *mean* pada atribut usia.

$mean(x) = \frac{1}{n}\ x\ \sum_{i=1}^{n} x_i$

Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

$mean\ (usia|1) = \frac{1}{5}\ x\ (48 + 83 + 82 + 68 + 86)$

$mean\ (usia|1) = 73{,}4$

Menghitung nilai *mean* juga dilakukan pada atribut usia dengan *output* = 2.

$mean\ (usia|2) = \frac{1}{5}\ x\ (45 + 45 + 49 + 34 + 42)$

$mean\ (usia|2) = 43$

Dari hasil hitung *mean* semua atribut didapat pada tabel di bawah ini.

Tabel 2 Hasil *Mean*

atribut	Output	*mean*
Age	1	73.4
	2	43
BMI	1	21.95876
	2	21.73858
Glucose	1	84.4
	2	91
Insulin	1	3.419
	2	11.083
HOMA	1	0.720414
	2	2.556566
Leptin	1	10.43446
	2	12.4918
Adiponectin	1	9.910505
	2	13.59857
Resistin	1	8.93588
	2	20.72511
MCP.1	1	628.5474
	2	462.5794

Sumber : Perhitungan

c. *Standar Deviastion*

Selanjutnya adalah menghitung nilai standar deviasi tiap atribut terhadap *output*. Berikut dicontohkan menghitung nilai standar deviasi pada atribut usia.

$StandardDeviation(x) = \sqrt{\frac{1}{n}\ x\ \sum_{i=1}^{n}(x_i - mean(x))^2}$

$StandarDeviastion\ (usia|1) = \sqrt{\frac{1}{5}\ x\ 999.2}$

$StandarDeviastion\ (usia|1) = 15.80506248$

999,2 didapat dari $\sum_{i=1}^{n}(x_i - mean(x))^2$ yaitu ((48-73,4)^2 + (83-73,4)^2 + (82-73,4)^2 + (68-73,4)^2 + (86-73,4)^2)

Penghitungan nilai standar deviasi dilakukan juga pada atribut usia dengan *output* = 2 (*Patients*).

$StandarDeviastion\ (usia|2) = \sqrt{\frac{1}{5}\ x\ 25.2}$

$StandarDeviastion\ (usia|2) = 5.61248608$

Tabel 3 Hasil perhitungan nilai standar deviasi

atribut	Output	stdev
Age	1	15.80506248
	2	5.61248608
BMI	1	1.26603088
	2	1.417615545
Glucose	1	10.26157883
	2	10.29563014
Insulin	1	0.674119796
	2	7.578929443
HOMA	1	0.204126017
	2	1.804204471
Leptin	1	4.351902904
	2	5.218593804
Adiponectin	1	7.251676322
	2	7.164004182
Resistin	1	3.244064202
	2	6.137141122
MCP.1	1	215.996834
	2	108.2918011

Sumber : Hasil Perhitungan

d. Data Uji

Tahap selanjutnya adalah pengujian terhadap data uji. Dalam pengujian digunakan persamaan *Gaussian Probability Density Function (PDF)*.

$pdf(x, meadnsd) = \frac{1}{\sqrt{2\ x\ \pi}\ x\ sd}\ x\ e^{-(\frac{(x-mean)^2}{2xsd^2})}$

Berikut contoh data uji

Tabel 4 Contoh data uji

No.	Age	…	MCP.1	Classification
1	78	…	209.749	1

Sumber : Hasil Perhitungan

Selanjutnya setiap atribut akan dihitung PDF. Berikut contoh untuk perhitungan PDF pada atribut usia.

$pdf(usia|1) = \frac{1}{\sqrt{2\ x\ \pi}\ x15.80506248}\ x\ e^{-(\frac{(78-73.4)^2}{2\ x\ 15.80506248})}$

$pdf(usia|1) = 0.024194675$

Lakukan penghitungan pada atribut usia = 0

$pdf(usia|2) = \frac{1}{\sqrt{2\ x\ \pi}\ x\ 5.61248608}\ x\ e^{-(\frac{(78-43)^2}{2\ x\ 5.61248608})}$

$pdf(usia|2) = 0.000000000255$

Penghitungan PDF dilakukan pada setiap atribut baik secara probabilitas 1 (*Healthy controls*) dan probabilitas 2 (*Patients*). Berikut hasil hitungnya.

P (*output*|1) = P PDF(age|1) x P PDF(BMI|1) x P PDF(Glucose|1) x P PDF(Insulin|1) x P PDF(HOMA|1) x P PDF(Leptin|1) x P PDF(Adiponectin|1) x P PDF(Resistin|1) x P PDF(MCP.1|1) x P PDF(*prior* prob|1)

P (*output*|1) = 0.024194675 x 0.009683117 x 0.00230122 x 0.586662103 x 1.202217195 x 0.062595151 x 0.054788622 x 0.051670573 x 0.00028192 x 0.5

P (*output*|1) = 0.000000000000000094980392857

Selanjutnya lakukan perhitungan PDF pada Probabilitas *output* = 2.

P (*output*|2) = P PDF(age|2) x P PDF(BMI|2) x P PDF(Glucose|2) x P PDF(Insulin|2) x P PDF(HOMA|2) x P PDF(Leptin|2) x P PDF(Adiponectin|2) x P PDF(Resistin|2) x P PDF(MCP.1|2) x P PDF(*prior* prob|2)

P (*output*|2) = 0.000000000255 x 0.011990441 x 0.000416478 x 0.031943321 x 0.116874881 x 0.040706424 x 0.050918676 x 0.002116864 x 0.000241362 x 0.5

P (*output*|2) = 0.0000000000000000000000000002521

Untuk menentukan hasil klasifikasi selanjutnya P (*output*|1) dibandingkan dengan P (*output*|2). Jika P (*output*|1) > P (*output*|2) = 1 (*Healthy controls*) tapi jika sebaliknya maka hasil klasifikasinya 2 (*Patients*).

Dari hasil penghitungan di atas P (*output*|1) lebih besar dari P (*output*|2) maka klasifikasi bernilai 1 (*Patients*).

Tabel 5 Nilai dari hasil Skenario Uji

ID	Uji Ke	AKurasi	Presisi
1	k 2.1	50	44.44444
2	k 2.2	55	50
3	k 4.1	40	33.33333
4	k 4.2	55	52.63158
5	k 4.3	45	37.5
6	k 4.4	55	55
7	k 5.1	37.5	28.57143
8	k.5.2	50	46.66667
9	k 5.3	50	50
10	k 5.4	50	42.85714
11	k 5.5	56.25	56.25
12	k 8.1	30	22.22222
13	k 8.2	50	44.44444
14	k 8.3	50	50
15	k 8.4	60	55.55556
16	k 8.5	30	25
17	k 8.6	60	50
18	k 8.7	50	50
19	k 8.8	60	60
20	k 10.1	25	25
21	k 10.2	50	33.33333
22	k 10.3	37.5	37.5
23	k 10.4	62.5	57.14286
24	**k 10.5**	**62.5**	**62.5**
25	k 10.6	37.5	33.33333
26	k 10.7	37.5	37.5
27	k 10.8	62.5	50
28	**k 10.9**	**62.5**	**62.5**
29	k 10.10	50	50

Sumber : Hasil Perhitungan

D. Uji Validasi

Model terbaik yang dihasilkan oleh skenario K *Fold Cross Validation* akan diuji kembali menggunakan data validasi. Berikut gambaran data validasi.

Tabel 6 Potongan Data Validasi

No.	Age	.	MCP.1	*Class*ification
1	66	.	864.968	1
2	73	.	788.902	2
3	69	.	263.499	1
4	24	.	632.22	1
5	44	.	63.61	1
...	...	.	...	...
32	29	.	174.8	1
33	85	.	1078.359	2

34	44	.	407.206	2
35	40	.	293.123	2
36	75	.	335.393	1

Sumber : Hasil Perhitungan

Data validasi ini berjumlah 36 dengankondisi *output class* 1 (healthty control) 18 data dan *output class* 2 (*Patients*) 18 data. Data ini dilakukan klasifikasi menggunakan persamaan *PDF* serta menggunakan model dari 10 *Fold* skenario 5 yang memiliki nilai akurasi tertinggi pada skenario K *Fold Cross Validation*. Berikut hasil potongan klasifikasi pada data validasi.

Tabel 7 Potongan hasil Klasifikasi Data Validasi Dari Model 10 *Fold* Skenario 5

ID	Age	.	GNB	Nilai Asli	Kriteria
1	66	.	2	1	FN
2	73	.	2	2	TN
3	69	.	2	1	FN
4	24	.	2	1	FN
5	44	.	2	1	FN
...	...	.	...	...	
32	29	.	2	1	FN
33	85	.	2	2	TN
34	44	.	2	2	TN
35	40	.	2	2	TN
36	75	.	1	1	TP

Sumber : Hasil Perhitungan

Berikut hasil kriteria menggunakan *confusion matrix* terhadap data uji.

Tabel 8 Hasil *confusion matrix* data validasi terhadap model 10 *Fold* skenario 5

Kriteria	Jumlah
TP	17
TN	5
FP	13
FN	1

Sumber : Hasil Perhitungan

Dari hasil kriteria menggunakan *confusion matrix*, diperoleh hasil pengukuran akurasi 61,11% dan presisi 56,66% pada data validasi.

Data validasi ini berjumlah 36 dengan kondisi *output class* 1 (healthty control) 18 data dan *output class* 2 (*Patients*) 18 data. Data ini dilakukan klasifikasi menggunakan persamaan *PDF* serta menggunakan model dari 10 *Fold* skenario 9 yang memiliki nilai akurasi tertinggi pada skenario K *Fold Cross Validation*. Berikut hasil klasifikasi pada data validasi.

Tabel 9 Potongan Hasil Klasifikasi Data Validasi Dari Model 10 *Fold* Skenario 9

ID	Age	.	GNB	Nilai Asli	Kriteria
1	66	.	1	1	TP
2	73	.	1	2	FP
3	69	.	2	1	FN
4	24	.	1	1	TP
5	44	.	1	1	TP
...	...	.	...	...	...
32	29	.	1	1	TP
33	85	.	2	2	TN
34	44	.	1	2	FP
35	40	.	1	2	FP
36	75	.	1	1	TP

Sumber : Hasil Perhittungan

Berikut hasil kriteria menggunakan *confusion matrix* terhadap data uji.

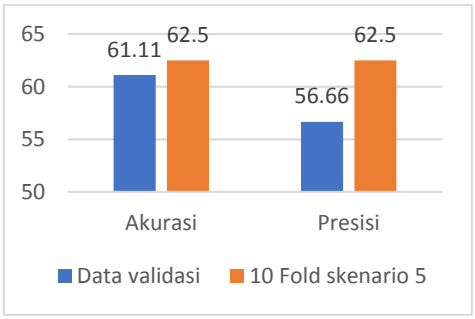
Tabel 10 Hasil *confusion matrix* data validasi terhadap model 10 *Fold* skenario 9

Kriteria	Jumlah
TP	17
TN	5
FP	13
FN	1

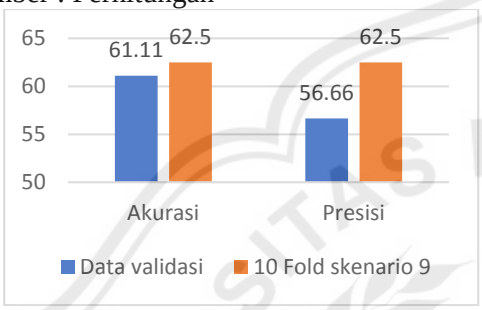
Sumber : Hasil Perhitungan

Dari hasil kriteria menggunakan *confusion matrix*, diperoleh hasil pengukuran akurasi 61,11% dan presisi 56,66% pada data validasi

Hasil pengukuran dari data validasi akan dibandingkan dengan hasil pengukuran yang diperoleh dari data uji 10 *Fold* skenario 5 dan hasil dari 10 *Fold* skenario 9 yang pada penelitian ini memperoleh nilai akurasi dan presisi yang sama, untuk mengetahui perubahan nilai akurasi dan presisi. Berikut hasil perbandingan nilai akurasi dari data validasi, data uji 10 *Fold* skenario 5 dan 10 *Fold* skenario 9.



Gambar 5 Grafik Akurasi dan Presisi Data Validasi 10 *Fold* Skenario 5
Sumber : Perhitungan



Gambar 6 Grafik Akurasi dan Presisi Data Validasi 10 *Fold* Skenario 5
Sumber : Perhitungan

Pada tingkat presisi terdapat penurunan yaitu sebesar 5,84% dari hasil 10 *Fold* skenario 9 terhadap hasil presisi data validasi. Nilai yang sama juga diperoleh pada percobaan hasil perbandingan data validasi terhadap 10 *Fold* skenario 9.

4. Kesimpulan

Hasil dari penelitian dari klasifikasi data penderita breast cancer menggunakan metode Gaussian Bayes dengan total 116 data dan skenario uji Cross Fold Validation dengan nilai k = 2, 4, 5, 8 dan 10 diperoleh beberapa kesimpulan yaitu:

a) Akurasi tertinggi dalam penelitian ini adalah 62,5% diperoleh pada skenario uji 10 fold skenario 5 dan skenario uji 10 fold skenario 9 dengan nilai akurasi yang sama. Jumlah data latih pada skenario ini adalah 72 data dan jumlah data uji 8 data.
b) Presisi tertinggi dalam penelitian ini adalah 62,5% diperoleh pada skenario uji 10 fold skenario 5 dan skenario uji 10 fold skenario 9 dengan nilai akurasi yang sama. Jumlah data latih pada skenario ini adalah 72 data dan jumlah data uji 8 data.
c) Hasil uji validasi menunjukkan tingkat akurasi yang diperoleh adalah 61,11% dan tingkat presisi yang diperoleh adalah 56,66%. Hal ini menunjukkan terdapat penurunan tingkat akurasi dari hasil uji 10 Fold skenario 5 sebesar 1,39% dan hasil uji validasi menunjukkan dari tingkat presisi yang diperoleh terdapat penurunan presisi sebesar 5,84% dari hasil uji 10 fold skenario 5 dan 10 fold skenario 9 terhadap hasil presisi data validasi.

5. Refrensi

Adi, I. L. T. (2007). Terapi herbal berdasarkan golongan darah. AgroMedia

Ahmad Zaki Hakimi. (2019). Klasifikasi Sel Darah Putih Menggunakan Gaussian NaïveBayes. Universitas Gajah Mada: Yogyakarta. Pada laman http://etd.repository.ugm.ac.id/index.php?mod=penelitian_detail&sub=PenelitianDetail&act=view&typ=html&buku_id=168410&is_local=1. Diakses pada tanggal 23 Oktober 2019, sekitar pukul 12.21 WIB.

Dalimartha, S. (2004). Deteksi dini kanker dan simplisia antikanker. Penebar Swadaya.

Dunston . T., & Yager N. (2009). Biometric System and Data Analysis Design, Evaluation, and Data Mining. New York: Springer.

Facts & Figures 2011-2012. American Cancer Society, Inc : Atlanta.

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

Han, J., & Kamber, M. (2012). Data Mining: C d h Concepts and Techniques.

Huang, J., Sun, H., Han, J., & Feng, B. (2011). Density-based shrinkage for revealing hierarchical and overlapping community structure in networks. Physica A: Statistical Mechanics and its Applications, 390(11), 2160-2171.

Jason Brownlee. (2016). Master Machine Learning Algorithms : Discover How They Work and Implement Them From Scratch.

Keating, B. (2008). Data Mining: What is it and how is it used? The Journal of Business Forecasting, 1, 33-35.

Kusumadewi, S., & Hartati, S. (2006). Neuro-Fuzzy: Integrasi Sistem Fuzzy dan Jaringan Syaraf. Yogyakarta: Graha Ilmu.

Leidiyana, H. (2013). Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor. Penelitian Ilmu Komputer Sistem Embedded dan Logic, 1(1).

Muhammad Firman Saputra. (2018). Klasifikasi Tingkat Buta Huruf Menggunakan Algoritma Kmeans-Naive Bayes. Universitas Negeri Malang: Malang. Pada laman http://karya-ilmiah.um.ac.id/index.php/TE/article/view/75325#. Diakses pada tanggal 23 Oktober 2019, sekitar pukul 12.21 WIB.

National Home Office: American Cancer Society Inc. (2012). Breast Cancer

Rabin, E. G., Heldt, E., Hirakata, V. N., & Fleck, M. P. (2008). Quality of life predictors in breast cancer women. European Journal of Oncology Nursing, 12(1), 53-57.

Uma Ojha. (2017). A study on prediction of breast cancer recurrence using data mining techniques. India: Delhi University.

Xie, T., Thummalapenta, S., Lo, D., & Liu, C. (2009). Data mining for software engineering. Computer, 42(8), 55-6

Jurnal Smart Teknologi
Vol. 1, No. 1, Maret 2021, Halaman 100 – 102
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST