Diagnosa Penderita Penyakit Kanker Payudara Menggunakan Metode *Naïve Bayes*

Taufik Frissetyo [1,*], Heri Kuswara [2]

[1] Sistem Informasi; Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri); Jl. Damai No.8 Warung Jati Barat (Margasatwa) Jakarta Selatan; e-mail: taufik11180817@nusamandiri.ac.id

[2] Sistem Informasi; Universitas Bina Sarana Informatika; Jl. Kamal Raya No.18, Cengkareng, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta; e-mail: heri.hrk@bsi.ac.id

* Korespondensi: e-mail: heri.hrk@bsi.ac.id

Diterima: 05 Oktober 2019 Direview: 12 Oktober 2019; Disetujui: 19 Oktober 2019

Cara sitasi: Prasetyo T, Kuswara H. 2019. Diagnosa Penderita Penyakit Kanker Payudara Menggunakan Metode Naïve Bayes. Information Management For Educators and Professionals. 4(1): 51-62.

Abstrak: Perkembangan teknologi informasi yang begitu cepat mempengaruhi lini kehidupan manusia. Hal ini dibuktikan dengan diciptakannya sistem berbasis komputer yang dapat mengatasi persoalan disegala lini kehidupan. Salah satu bidang yang saat ini dipengaruhi oleh teknologi informasi adalah bidang kesehatan dan kedokteran. Penyakit kanker semakin hari semakin banyak diderita oleh sebagian besar orang bahkan tanpa mereka sadari. Hal ini diakibatkan karena pola hidup yang kurang sehat dan ada sebagian orang yang sudah tidak terlalu peduli akan kesehatannya padahal kesehatan merupakan hal yang amat sangat penting bagi kehidupan manusia. Salah satu Kanker yang paling banyak diderita olah masyarakat khususnya kaum perempuan dan sangat mematikan adalah Kanker Payudara. Maka perlunya deteksi dini kanker sangatlah penting demi keselamatan penderita. Metode yang digunakan dalam penelitian ini adalah metode *Naïve Bayes* karena metode ini sangat sederhana tetapi memiliki tingkat akurasi yang baik. Hasil perhitungan yang dilakukan baik secara manual dan menggunakan *Rapidminer* diketahui akurasi untuk *data training* sebesar 81.00% dan *AUC* sebesar 0,922 dengan kategori *Excellent Classification* dan hasil akurasi *data testing* sebesar 62.50% dan *AUC* sebesar 0,719 dengan kategori *Fair Classification*.

Kata kunci: Diagnosa, Kanker, Payudara, *Naïve Bayes*

Abstract: *The development of information technology is so fast influencing the lines of human life. This is evidenced by the creation of computer-based systems that can overcome problems in all lines of life. One of the fields currently affected by information technology is health and medicine. Cancer is increasingly suffering by most people even without them knowing it. This is caused by unhealthy lifestyles and there are some people who are not too concerned about their health even though health is very very important for human life. One of the most common cancers suffered by the community, especially women and is very deadly is Breast Cancer. So the need for early detection of cancer is very important for patient safety. The method used by the authors in this study is the Naïve Bayes method because this method is very simple but has a good level of accuracy. The results of calculations that the authors do both manually and using Rapidminer known accuracy for training data of 81.00% and AUC of 0.922 with the category of Excellent Classification and the results of data testing accuracy of 62.50% and AUC of 0.719 with the Fair Classification category.*

Keywords*: Breast, Cancer, Diagnosis, Naïve Bayes*

E-ISSN: 2548-3331 ; 51 – 62

1. Pendahuluan

Penyakit kanker saat ini menjadi salah satu penyakit yang menyebabkan kematian diseluruh dunia hal ini dibuktikan dengan data dari *International Agency for Reasearch on Cancer* yang merilis data di tahun 2012 tentang jenis penyakit kanker yang paling banyak diderita oleh perempuan adalah kanker payudara, kanker kolorektal dan kanker serviks [Praningki and Budi, 2018]. Penyakit ini memang semakin hari semakin banyak diderita oleh sebagian besar orang bahkan tanpa mereka sadari. Hal ini diakibatkan karena pola hidup yang kurang sehat dan ada sebagian orang yang sudah tidak terlalu peduli akan kesehatannya padahal kesehatan merupakan hal yang amat sangat penting bagi kehidupan manusia.

Perkembangan teknologi informasi dan komunikasi yang begitu cepat sudah mempengaruhi segala lini kehidupan manusia. Hal ini dibuktikan dengan mulai diciptakannya sistem berbasis komputer yang dapat mengatasi berbagai persoalan disegala lini kehidupan. Salah satu bidang yang saat ini sangat dipengaruhi oleh teknologi informasi dan komunikasi adalah bidang kesehatan dan kedokteran. Mulai diciptakannya alat-alat penunjang kesehatan berbasis komputer menjadi salah satu bukti bahwa bidang teknologi informasi dan komunikasi sudah menjadi bagian dari ilmu kesehatan dan kedokteran [Via, Nugroho, and Syafrizal, 2015].

Penyakit kanker sudah menjadi salah satu penyakit paling mematikan bagi umat manusia di seluruh dunia. Hal tersebut dibuktikan di tahun 2012 ada kurang lebih 8,2 juta meninggal dunia akibat penyakit ini. Adapun lima kanker yang harus diwaspadai karena lima jenis kanker ini adalah kanker yang menyebabkan kematian terbanyak tiap tahunnya yaitu kanker paru, kanker hati, kanker perut, kanker kolorektal dan kanker payudara [Ma'arif and Arifin, 2017].

Kanker pada umumnya dibedakan kedalam dua jenis yaitu kanker jinak dan kanker ganas. Untuk kanker jinak sendiri adalah kanker yang kondisinya masih dalam tahap awal sehingga terkadang jenis kanker ini masih bisa ditangani oleh tenaga medis dan bahkan sangat besar potensi penyembuhannya. Sedangkan untuk kanker ganas merupakan jenis kanker yang sangat berbahaya dan apabila tidak ditangani dengan baik akan berpotensi pada kematian. Maka dari itu perlunya deteksi dini kanker sangatlah penting demi keselamatan penderita [Farahdiba and Nugroho, 2016].

Berdasarkan beberapa permasalahan yang terjadi, tertarik untuk melakukan penelitian untuk mendiagnosa penderita penyakit kanker payudara. Adapun metode yang akan digunakan dalam penelitian ini adalah metode *Naïve Bayes*, karena disamping metode ini sangat sederhana tetapi juga memiliki tingkat akurasi yang baik. Menurut Xindong dan Kumar (tahun) *Naïve Bayes* merupakan salah satu dari sepuluh algoritma terbaik dalam data mining [Praningki and Budi, 2018].

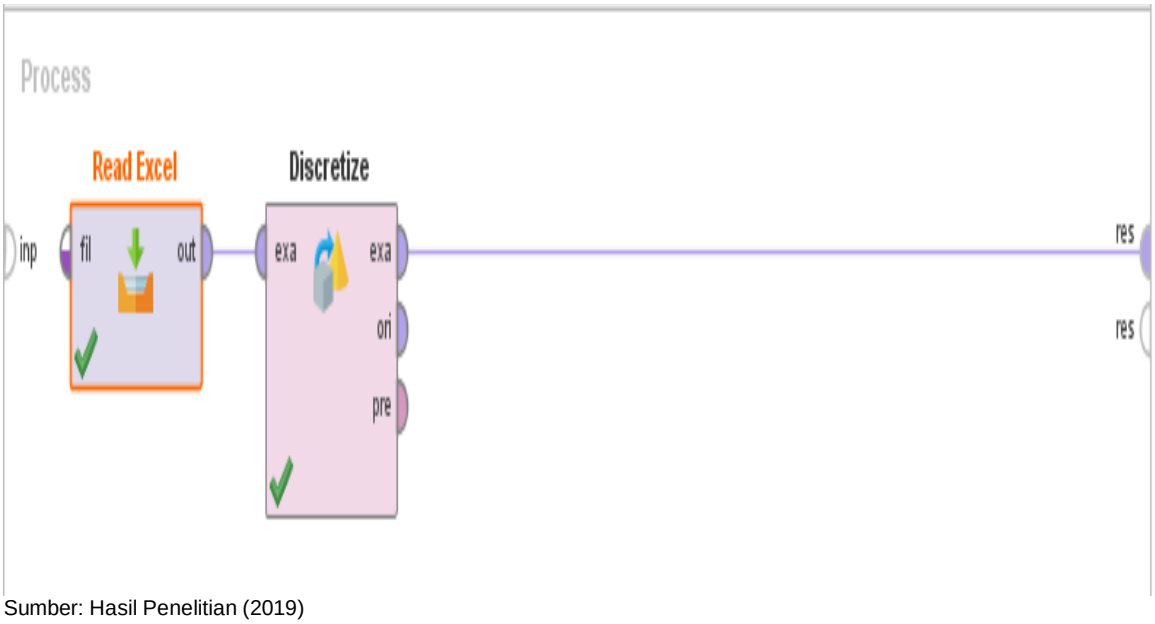
Teori *Bayes* adalah kondisi probabilitas suatu kejadian (hipotesis) bergantung pada kejadian lain (bukti). Pada dasarnya, teorema tersebut mengataksan bahwa kejadian di masa depan dapat diprediksi dengan syarat kejadian sebelumnya telah terjadi [Moriesta, Selviani, and Ibrahim, 2017]. *Naïve Bayes Classifier* atau sering disebut *Bayesian Classification* adalah metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Naïve Bayes Classifier (NBC)* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar Kusrini., Luthfi dan Emha dalam [Zulfikar and Lukman, 2017].

Penelitian tentang diagnosa, klasifikasi dan juga prediksi tentang penyakit kanker payudara ini sudah dilakukan oleh beberapa peneliti sebelumnya diantaranya penelitian yang dilakukan oleh [Ma'arif and Arifin, 2017] dengan penggabungan seleksi fitur *Backward Elemination* dan *Support Vector Machine* yang menghasilkan akurasi 97.14% dengan data yang digunakan adalah data *Wisconsin Breast Cancer* dari *UCI Machine Learning Repository*.

Penelitian selanjutnya dilakukan oleh [Via, Nugroho, and Syafrizal, 2015] dengan menggunakan metode *Naïve Bayes*, peneliti membuat sistem penunjang keputusan untuk menentukan keganasan kanker payudara dan diperoleh hasil akurasi sebesar 97.82%. Peneliti menggunakan data dari *Wisconsin Breast Cancer Original* yang diperoleh dari *UCI Machine Learning Repository*.

2. Metode Penelitian

Penelitian ini dimulai dengan melakukan pengumpulan data. Data yang digunakan dalam penelitian ini adalah data yang diperoleh dari *website* penyedia *dataset* yang terkenal dan sudah menjadi rujukan banyak peneliti untuk melakukan penelitan yaitu *UCI Machine Learning Repository*. Adapun data yang digunakan adalah *Breast Cancer Coimbra Data Set* yang terdiri dari 116 *record* data dan 10 atribut yang terdiri dari *Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1*, Label Kelas. dan dibagi kedalam dua kelas yaitu *patients* berjumlah 64 *record* dan *healthy controls* berjumlah 52 *record*. Pada tahap pengolahan data, data dibersihkan dari noise yang ada seperti data tidak lengkap, mencari apakah ada data yang duplikat, mencari data yang inkonsisten dan juga memperbaiki data-data tersebut apabila ada dan ditemukan kesalahan atau *noise* dari data tersebut.



Sumber: Hasil Penelitian (2019)

Gambar 1. Transformasi Data Menggunakan *RapidMiner*

Setelah data dipastikan tidak bermasalah atau *noise*, melakukan perhitungan untuk mengkategorikan *value* dari data tersebut dikarenakan data yang ada berbentuk numerik sehingga data jika dalam bentuk numerik akan sangat sulit diolah. Sehingga memutuskan untuk mengkategorikan semua *value* atribut tersebut kedalam 8 kategori atau interval dengan rumus 1+3,3 *log* n dimana n merupakan jumlah *record* [Rifai and Wijayanti, 2017]. Untuk transformasi data dari numerik ke kategori menggunakan bantuan aplikasi *RapidMiner* dengan membagi data tersebut kedalam 8 kategori yaitu *range1, range2, range3, range4, range5, range6, range7 dan range8.*

Tabel 1. Pembagian Kelas Interval *Age, BMI, Glucose, Insulin* dan *HOMA*

Age	BMI	Glucose	Insulin	HOMA
range1 [< 32.125]	range1 [< 20.896]	range1 [< 77.625]	range1 [< 9.435]	range1 [< 3.540]
range2 [32.125 - 40.250]	range2 [20.896 - 23.422]	range2 [77.625 - 95.250]	range2 [9.435 - 16.439]	range2 [3.540 - 6.613]
range3 [40.250 - 48.375]	range3 [23.422 - 25.948]	range3 [95.250 - 112.875]	range3 [16.439 - 23.442]	range3 [6.613 - 9.686]
range4 [48.375 - 56.500]	range4 [25.948 - 28.474]	range4 [112.875 - 130.500]	range4 [23.442 - 30.446]	range4 [9.686 - 12.759]
range5 [56.500 - 64.625]	range5 [28.474 - 31.000]	range5 [130.500 - 148.125]	range5 [30.446 - 37.450]	range5 [12.759 - 15.832]
range6 [64.625 - 72.750]	range6 [31.000 - 33.527]	range6 [148.125 - 165.750]	range6 [37.450 - 44.453]	range6 [15.832 - 18.905]
range7 [72.750 - 80.875]	range7 [33.527 - 36.053]	range7 [165.750 - 183.375]	range7 [44.453 - 51.456]	range7 [18.905 - 21.977]
range8 [80.875 >]	range8 [36.053 >]	range8 [183.375 >]	range8 [51.456 >]	range8 [21.977 >]

Sumber : Hasil Penelitian (2019)

E-ISSN: 2548-3331 ; 51 – 62

Tabel 2. Pembagian Kelas Interval *Leptin, Adiponectin, Resistin* dan *MCP.1*

Leptin	Adiponectin	Resistin	MCP.1
range1 [< 15.057]	range1 [< 6.204]	range1 [< 13.071]	range1 [< 252.418]
range2 [15.057 - 25.803]	range2 [6.204 - 10.752]	range2 [13.071 - 22.933]	range2 [252.418 - 458.992]
range3 [25.803 - 36.549]	range3 [10.752 - 15.300]	range3 [22.933 - 32.794]	range3 [458.992 - 665.567]
range4 [36.549 - 47.295]	range4 [15.300 - 19.848]	range4 [32.794 - 42.655]	range4 [665.567 - 872.141]
range5 [47.295 - 58.042]	range5 [19.848 - 24.396]	range5 [42.655 - 52.516]	range5 [872.141 - 1078.716]
range6 [58.042 - 68.788]	range6 [24.396 - 28.944]	range6 [52.516 - 62.378]	range6 [1078.716 - 1285.291]
range7 [68.788 - 79.534]	range7 [28.944 - 33.492]	range7 [62.378 - 72.239]	range7 [1285.291 - 1491.865]
range8 [79.534 >]	range8 [33.492 >]	range8 [72.239 >]	range8 [1491.865 >]

Sumber : Hasil Penelitian (2019)

Setelah melakukan pengumpulan dan pengolahan data awal, tahap selanjutnya adalah memilih metode yang akan digunakan pada Tahap Analisis. Dalam tahap ini juga dilakukan pembagian untuk *data training* sebanyak 100 data dan *data testing* sebanyak 16 data dari 116 data yang ada. Adapun metode yang digunakan dalam penelitian ini adalah metode *Naïve Bayes*, karena disamping sederhana, metode ini juga menghasilkan akurasi yang cukup tinggi. Berikut adalah rumus dari algoritma *Naïve Bayes*.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad \text{..... (1)}$$

Keterangan
X = Data dengan *class* yang belum diketahui
C = Hipotesis data merupakan suatu *class* spesifik
P(C|X) = Probabilitas hipotesis berdasar kondisi (*posteriori probability*)
P(C) = Probabilitas hipotesis *(prior probability*)
P(X|C) = Probabilitas berdasarkan kondisi pada hipotesis
P(X) = Probabilitas *C*

Menjelaskan Eksperimen dan Pengujian Metode tentang bagaimana eksperimen dilakukan sampai terbentunya sebuah keputusan. Melakukan perhitungan berdasarkan metode yang dipilih sampai nilai dari akurasi diperoleh. Tahap evaluasi merupakan tahap akhir dari penelitian ini. Setelah melakukan tahap analisis maka akan menghasilkan nilai akurasi. Kemudian dari hasil akurasi tersebut dievaluasi dan ditarik kesimpulan.

3. Hasil dan Pembahasan

Tahap pertama yang dilakukan adalah menghitung probabilitas masing-masing kelas berdasarkan kriteria yang ada menggunakan *data training* sebagai acuan. Berikut adalah perhitungan probabilitas berdasarkan data yang akan digunakan dalam penelitian ini.

Menghitung probabilitas kelas / *P(Ci)*, Nilai 100 merupakan jumlah data training keseluruhan sedangkan 44 merupakan jumlah data training yang memiliki kelas *Healthy Controls* lalu 56 merupakan jumlah data *training* yang memiliki kelas *Patients*.

P (Class = Healthy Controls) = 44/100 = 0,44
P (Class = Patients) = 56/100 = 0,56

Menghitung probabilitas *Age* terhadap kelas / *P(Age|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 6 merupakan jumlah *data training* atribut *Age range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 0 merupakan jumlah *data training* atribut *Age range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan porbabilitas *Age* terhadap masing-masing kelas.

P (Age = range1 | Class = Healthy Controls) = 6/44 = 0,136363636
P (Age = range1 | Class = Patients) = 0/56 = 0

Tabel 3. Probabilitas *Age* terhadap Kelas

	HC	P	P(Age|Class=HC)	P(Age|Class=P)
range1	6	0	0,136363636	0
range2	7	4	0,159090909	0,07142857
range3	4	18	0,090909091	0,32142857
range4	4	8	0,090909091	0,14285714
range5	3	6	0,068181818	0,10714286
range6	9	12	0,204545455	0,21428571
range7	10	3	0,227272727	0,05357143
range8	1	5	0,022727273	0,08928571

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *BMI* terhadap kelas / *P(BMI|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 2 merupakan jumlah *data training* atribut *BMI range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 5 merupakan jumlah *data training* atribut *BMI range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *BMI* terhadap masing-masing kelas.

P (BMI = range1 | Class = Healthy Controls) = 2/44 = 0,045454545
P (BMI = range1 | Class = Patients) = 5/56 = 0,089285714

Tabel 4. Probabilitas *BMI* terhadap Kelas

	HC	P	P(BMI|Class=HC)	P(BMI|C=P)
range1	2	5	0,045454545	0,08928571
range2	7	7	0,159090909	0,125
range3	3	5	0,068181818	0,08928571
range4	7	13	0,159090909	0,23214286
range5	7	12	0,159090909	0,21428571
range6	6	10	0,136363636	0,17857143
range7	7	3	0,159090909	0,05357143
range8	5	1	0,113636364	0,01785714

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *Glucose* terhadap Kelas / *P(Glucose|Ci)*, nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 4 merupakan jumlah *data training* atribut *Glucose range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 3 merupakan jumlah *data training* atribut *Glucose range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *Glucose* terhadap masing-masing kelas.

P (Glucose = range1 | Class = Healthy Controls) = 4/44 = 0,090909091
P (Glucose = range1 | Class = Patients) = 3/56 = 0,053571429

Tabel 5. Probabilitas *Glucose* terhadap Kelas

	HC	P	P(Glucose|Class=HC)	P(Glucose|Class=P)
range1	4	3	0,090909091	0,05357143
range2	31	20	0,704545455	0,35714286
range3	9	20	0,204545455	0,35714286
range4	0	3	0	0,05357143
range5	0	6	0	0,10714286
range6	0	1	0	0,01785714
range7	0	0	0	0
range8	0	3	0	0,05357143

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *Insulin* terhadap Kelas / *P(Insulin|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 36 merupakan jumlah *data training* atribut *Insulin range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 28 merupakan jumlah *data training* atribut *Insulin range1* yang terdapat di

E-ISSN: 2548-3331 ; 51 – 62

kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *Insulin* terhadap masing-masing kelas.

P (Insulin = range1 | Class = Healthy Controls) = 36/44 = 0,818181818
P (Insulin = range1 | Class = Patients) = 28/56 = 0,5

Tabel 6. Probabilitas *Insulin* terhadap Kelas

	HC	P	P(Insulin|Class=HC)	P(Insulin|Class=P)
range1	36	28	0,818181818	0,5
range2	5	12	0,113636364	0,21428571
range3	2	7	0,045454545	0,125
range4	1	4	0,022727273	0,07142857
range5	0	1	0	0,01785714
range6	0	2	0	0,03571429
range7	0	0	0	0
range8	0	2	0	0,03571429

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *HOMA* terhadap Kelas / *P(Homa|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 40 merupakan jumlah *data training* atribut *HOMA range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 38 merupakan jumlah *data training* atribut *HOMA range1* yang terdapat di kelas *Patients.* Berikut adalah daftar lengkap hasil perhitungan probabilitas *HOMA* terhadap masing-masing kelas.

P (HOMA = range1 | Class = Healthy Controls) = 40/44 = 0,909090909
P (HOMA = range1 | Class = Patients) = 38/56 = 0,678571429

Tabel 7. Probabilitas *HOMA* terhadap Kelas

	HC	P	P(HOMA|Class=HC)	P(HOMA|Class=P)
range1	40	38	0,909090909	0,67857143
range2	3	9	0,068181818	0,16071429
range3	1	4	0,022727273	0,07142857
range4	0	1	0	0,01785714
range5	0	2	0	0,03571429
range6	0	0	0	0
range7	0	1	0	0,01785714
range8	0	1	0	0,01785714

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *Leptin* terhadap Kelas / *P(Leptin|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 10 merupakan jumlah *data training* atribut *Leptin range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 18 merupakan jumlah *data training* atribut *Leptin range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *Leptin* terhadap masing-masing kelas.

P (Leptin = range1 | Class = Healthy Controls) = 10/44 = 0,227272727
P (Leptin = range1 | Class = Patients) = 18/56 = 0,321428571

Tabel 8. Probabilitas *Leptin* terhadap Kelas

	HC	P	P(Leptin|Class=HC)	P(Leptin|Class=P)
range1	10	18	0,227272727	0,32142857
range2	13	14	0,295454545	0,25
range3	8	7	0,181818182	0,125
range4	6	8	0,136363636	0,14285714
range5	3	5	0,068181818	0,08928571
range6	2	1	0,045454545	0,01785714
range7	1	1	0,022727273	0,01785714
range8	1	2	0,022727273	0,03571429

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *Adiponectin* terhadap Kelas / *P(Adiponectin|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 12 merupakan jumlah *data training* atribut *Adiponectin range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah

data training kelas *Patients* dan 13 merupakan jumlah *data training* atribut *Adiponectin range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *Adiponectin* terhadap masing-masing kelas.

P (Adiponectin = range1 | Class = Healthy Controls) = 12/44 = 0,272727273
P (Adiponectin = range1 | Class = Patients) = 13/56 = 0,232142857

Tabel 9. Probabilitas *Adiponectin* terhadap Kelas

	HC	P	P(Ad|C=HC)	P(Ad|C=P)
range1	12	13	0,272727273	0,23214286
range2	22	23	0,5	0,41071429
range3	3	9	0,068181818	0,16071429
range4	1	4	0,022727273	0,07142857
range5	3	6	0,068181818	0,10714286
range6	1	0	0,022727273	0
range7	0	0	0	0
range8	2	1	0,045454545	0,01785714

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *Resistin* terhadap Kelas / *P(Resistin|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 32 merupakan jumlah *data training* atribut *Resistin range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 24 merupakan jumlah *data training* atribut *Resistin range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *Resistin* terhadap masing-masing kelas.

P (Resistin = range1 | Class = Healthy Controls) = 32/44 = 0,727272727
P (Resistin = range1 | Class = Patients) = 24/56 = 0,428571429

Tabel 10. Probabilitas *Resistin* terhadap Kelas

	HC	P	P(Resistin|Class=HC)	P(Resistin|Class=P)
range1	32	24	0,727272727	0,42857143
range2	9	16	0,204545455	0,28571429
range3	2	10	0,045454545	0,17857143
range4	0	1	0	0,01785714
range5	0	2	0	0,03571429
range6	0	3	0	0,05357143
range7	0	0	0	0
range8	1	0	0,022727273	0

Sumber: Hasil Penelitian (2019)

Menghitung Probabilitas *MCP-1* terhadap Kelas / *P(MCP-1|Ci)*, Nilai 44 merupakan jumlah *data training* kelas *Healthy Controls* dan 12 merupakan jumlah *data training* atribut *MCP-1 range1* yang terdapat di kelas *Healthy Controls*. Nilai 56 merupakan jumlah *data training* kelas *Patients* dan 12 merupakan jumlah *data training* atribut *MCP-1 range1* yang terdapat di kelas *Patients*. Berikut adalah daftar lengkap hasil perhitungan probabilitas *MCP-1* terhadap masing-masing kelas.

P (MCP-1 = range1 | Class = Healthy Controls) = 12/44 = 0,272727273
P (MCP-1 = range1 | Class = Patients) = 12/56 = 0,214285714

Tabel 11. Probabilitas *MCP-1* terhadap Kelas

	HC	P	P(MCP-1|Class=HC)	P(MCP-1|Class=P)
range1	12	12	0,272727273	0,21428571
range2	11	16	0,25	0,28571429
range3	11	11	0,25	0,19642857
range4	5	8	0,113636364	0,14285714
range5	3	5	0,068181818	0,08928571
range6	2	0	0,045454545	0
range7	0	0	0	0
range8	0	4	0	0,07142857

Sumber: Hasil Penelitian (2019)

E-ISSN: 2548-3331 ; 51 – 62

Setelah probabilitas dari semua kriteria diperoleh maka tahap selanjutnya adalah melakukan perhitungan untuk menentukan kelas masing-masing *record*. Untuk pengklasifikasian pertama akan menggunakan *data training* terlebih dahulu.

Tabel 12. Salah Satu *Record* Dari *Data Training*

No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Asiponectin	Resistin	MCP.1	Class
9	range 7	range 2	range3	range1	range1	range 1	range2	range1	range 1	Healthy Controls

Sumber : Hasil Penelitian (2019)

Pada tabel 12 merupakan salah satu *record* dari 100 *record Data Training* yang akan digunakan dalam penelitian ini. Kemudian akan menghitung apakah hasil perhitungan dengan metode *Naïve Bayes* ini benar atau tidak klasifikasi yang dilakukan berdasarkan probabilitas yang telah dihitung sebelumnya. Pertama akan melakukan perhitungan dengan menggunakan Probabilitas kelas *Healthy Controls* dengan mengkalikan seluruh probabilitas dari atribut yang ada berdasarkan kelas *Healthy Controls*.

*(P(X|Class=Healthy Controls)=P(Age=range7, BMI=range2, Glucose=range3, Insulin=range1, HOMA=range1, Leptin=range1, Adiponectin=range2, Resistin=range1, MCP-1=range1 | Class=Healthy Controls))*P(Class=Healthy Controls)* (0,227272727 x 0,159090909 x 0,204545455 x 0,818181818 x 0,909090909 x 0,227272727 x 0,5 x 0,727272727 x 0,272727273) x 0,44 = 0,0000545551.

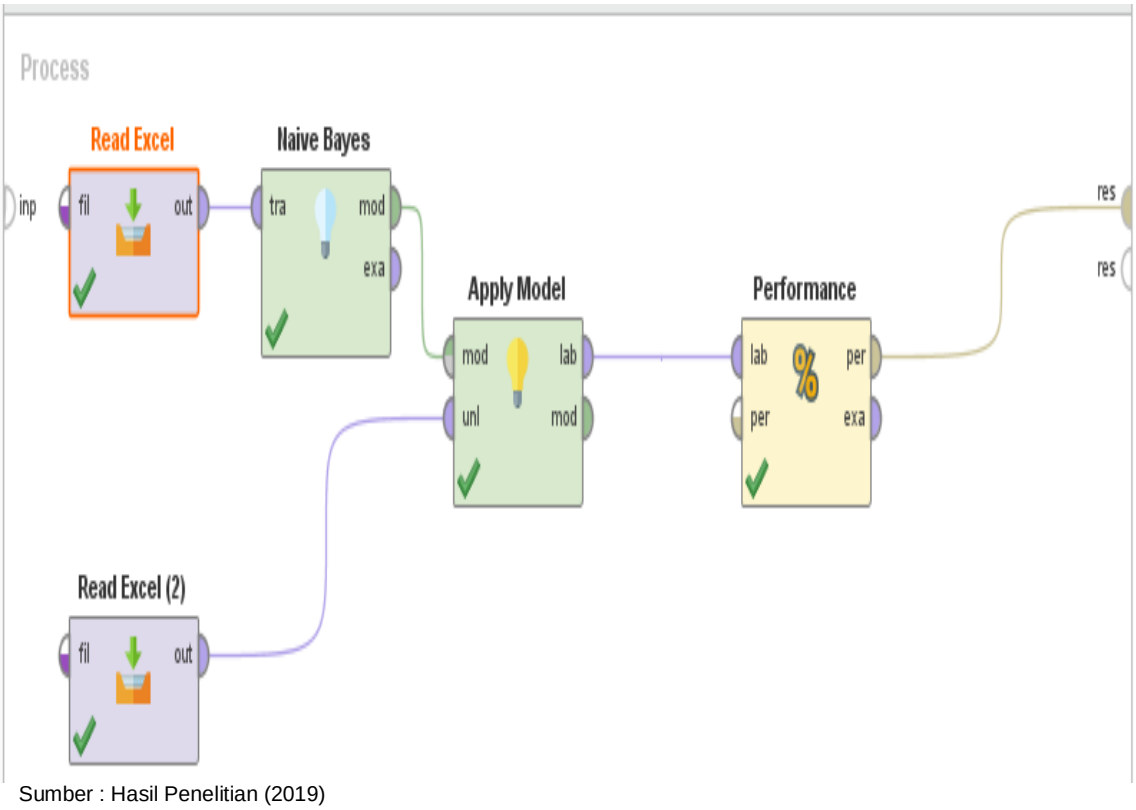
Selanjutnya juga akan melakukan perhitungan dengan menggunakan Probabilitas kelas *Patients*. *(P(X|Class=Patients)=P(Age=range7, BMI=range2, Glucose=range3, Insulin=range1, HOMA=range1, Leptin=range1, Adiponectin=range2, Resistin=range1, MCP-1=range1 | Class=(Patients))*P(Class=Patients)* (0,053571429 x 0,125 x 0,357142857 x 0,5 x 0,678571429 x 0,321428571 x 0,410714286 x 0,428571429 x 0,214285714) x 0,56 = 0,0000055091.

Setelah mengetahui hasil perhitungan dari kedua kelas yang ada yaitu *Healthy Controls* dan *Patients*, maka tahap selanjutnya adalah melakukan perbandingan dari kedua hasil tersebut untuk menentukan mana yang paling besar. Nilai yang paling besar tersebut yang akan menjadi kelas yang terpilih dari kedua kelas yang ada. Berdasarkan hasil perhitunga diatas maka nilai terbesar adalah 0,0000545551 yang merupakan kelas dari *Healthy Controls*. Jika dibandingkan dengan data diatas yang *record* tersebut memiliki kelas *Healthy Controls*. Maka dapat disimpulkan bahwa proses klasifikasi salah satu *record* ini adalah benar. Jika salah maka hasil dari perhitungan dengan data yang ada memiliki kelas berbeda.

Berdasarkan perhitungan dan perbandingan menggunakan *data training* maka diperoleh data yang berhasil diklasifikasikan atau cocok sesuai dengan data kelas sesungguhnya adalah 81 *record* dan yang gagal diklasifikasikan atau tidak sesuai dengan data kelas sesungguhnya sebesar 19 *record*. Setelah menyiapkan *data testing* yang akan menguji hasil dari perhitungan algoritma menggunakan metode *Naïve Bayes* untuk mengetahui hasil akurasi jika menggunakan data uji. Berdasarkan hasil perhitungan atau pengklasifikasian menggunakan *data testing*. Diperoleh hasil dari 16 *record* yang ada, berhasil diklasifikasikan dengan benar 10 *record* dan 6 *record* gagal diklasifikasikan dengan benar.

Pada tahap evaluasi ini, akan dilakukan beberapa perhitungan terutama perhitungan untuk menentukan akurasi dari metode *Naïve Bayes* ini. Berdasarkan hasil perhitungan sebelumnya diketahui bahwa dari 100 *data training* yang ada. Total ada 81 data yang berhasil diklasifikasikan dengan benar dan 19 data yang gagal diklasifikasikan dengan benar. Maka dari itu perhitungan akurasinya adalah 81/100 x 100%=81.00%.

Berdasarkan perhitungan diatas diperoleh hasil akurasi *data training* adalah 81.00% yang artinya bahwa klasifikasi data training pasien kanker payudara menggunakan metode Naïve Bayes ini bisa digunakan karena tingkat akurasi yang sudah cukup baik dari sisi akurasi. Pada penelitian ini, selanjutnya mencoba melakukan perhitungan secara otomatis menggunakan aplikasi *Rapidminer* untuk melihat dan membandingkan hasil perhitungan yang akan dilakukan secara manual dengan hasil otomatis yang dilakukan oleh aplikasi *Rapidminer*.



Sumber : Hasil Penelitian (2019)

Gambar 2. Alur Proses *Naïve Bayes* Menggunakan *RapidMiner*

Berikut gambar 3 adalah hasil dari perhitungan menggunakan aplikasi *Rapidminer* berdasarkan *data training* yang sama dalam penelitian ini.

accuracy: 81.00%

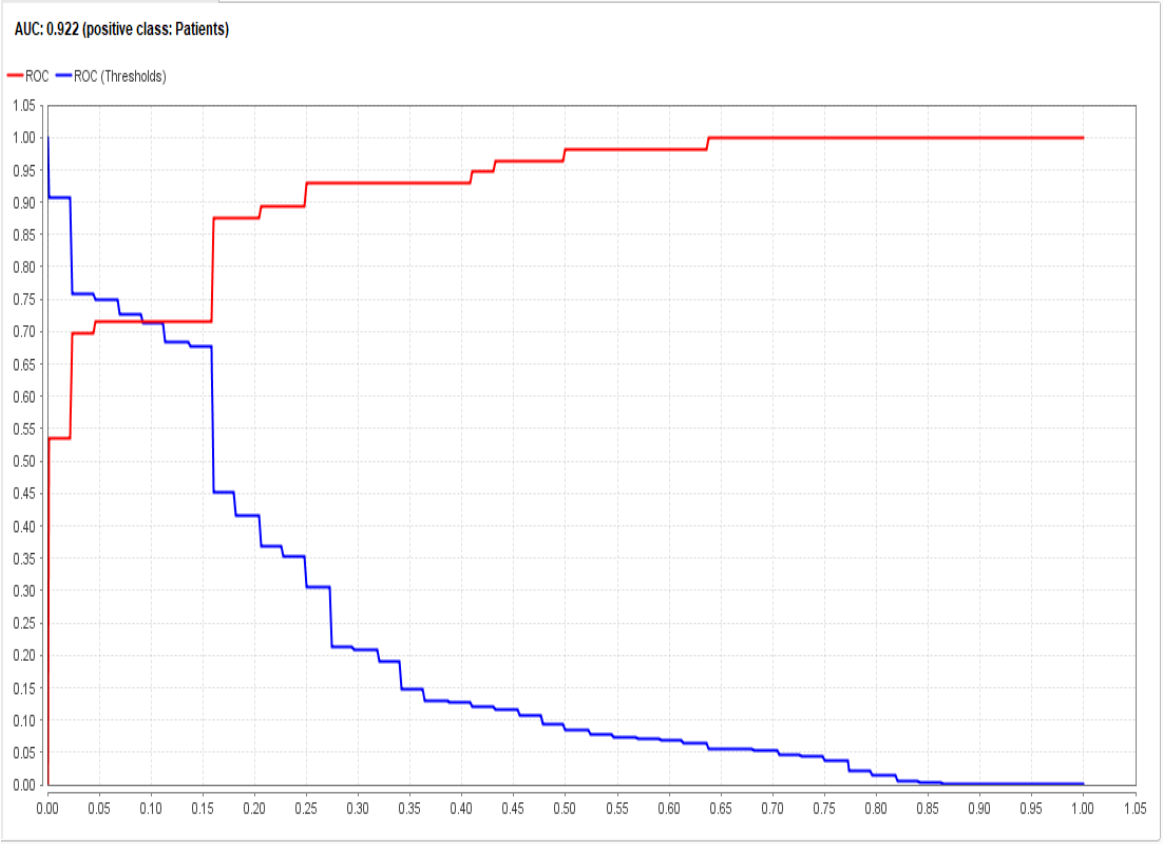
	true Healthy Controls	true Patients	class precision
pred. Healthy Controls	37	12	75.51%
pred. Patients	7	44	86.27%
class recall	84.09%	78.57%	

Sumber : Hasil Penelitian (2019)

Gambar 3. Hasil Akurasi *Data Training* Menggunakan *Rapidminer*

Dapat dilihat untuk nilai *AUC (Area Under Cover)* dari *data training* diperoleh hasil 0,922, nilai tersebut menurut Gorunescu masuk kedalam kategori *Excellent Classification* [Septiani, 2017].

E-ISSN: 2548-3331 ; 51 – 62



Sumber : Hasil Penelitian (2019)

Gambar 4. Hasil *AUC Data Training* Menggunakan *Rapidminer*

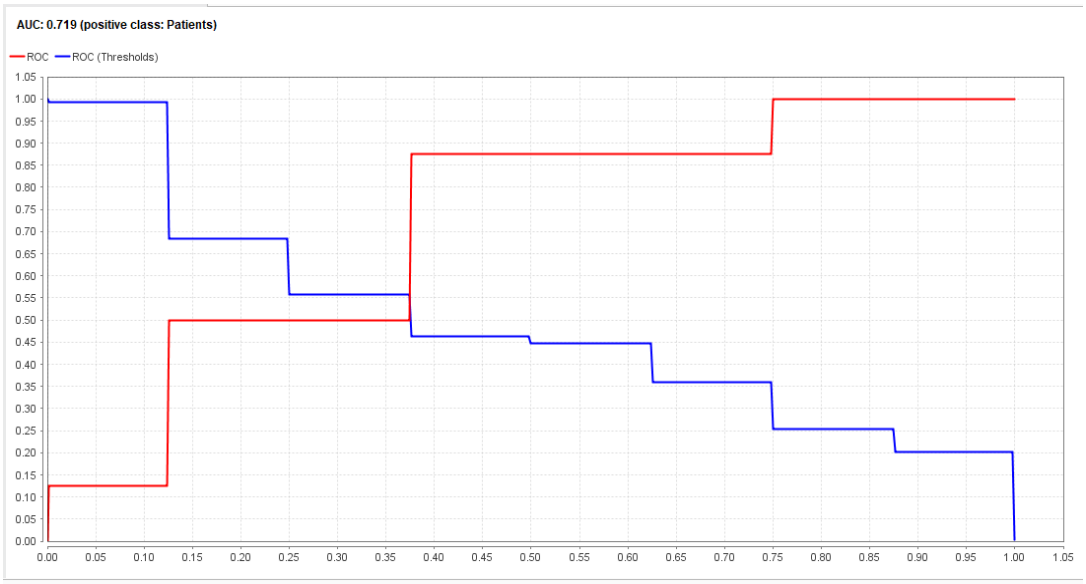
Sementara untuk hasil dari pengkalisifikasian menggunakan *data testing* diperoleh hasil dari 16 *record* yang ada 10 *record* berhasil diklasifikasikan dengan benar dan 6 *record* gagal diklasifikasikan dengan benar. Berikut adalah hasil perhitungan akurasinya sebesar 10/16 x 100%=62.50%. Dapat dilihat untuk akurasi dari peengklaisifikasian menggunakan *data testing* jauh lebih rendah dibanding dengan akurasi menggunakan *data training.* Selanjutnya kita akan bandingkan dengan hasil perhitungan menggunakan aplikasi *Rapidminer*.

accuracy: 62.50%

	true Healthy Controls	true Patients	class precision
pred. Healthy Controls	5	3	62.50%
pred. Patients	3	5	62.50%
class recall	62.50%	62.50%	

Sumber : Hasil Penelitian (2019)

Gambar 5. Hasil Akurasi *Data Testing* Menggunakan *Rapidminer*



Sumber : Hasil Penelitian (2019)

Gambar 6. Hasil *AUC Data Testing* Menggunakan *Rapidminer*

Dapat dilihat untuk nilai *AUC (Area Under Cover)* dari *data testing* diperoleh hasil 0,719, nilai tersebut menurut Gorunescu masuk kedalam kategori *Fair Classification* [Septiani, 2017].

4. Kesimpulan

Berdasarkan hasil dari perhitungan yang telah lakukan sebelumnya, dapat menyimpulkan bahwa metode *Naïve Bayes* cukup bisa digunakan untuk mengklasifikasikan dan diagnosa penderita penyakit Kanker Payudara. Hal ini didapat berdasarkan tingkat akurasi yang diperoleh melalui perhitungan baik manual maupun menggunakan aplikasi *Rapidminer*. Untuk perhitungan manual sendiri, hasil akurasi untuk *data training* sebesar 81.00% sama dengan hasil pengklasifikasian menggunakan *Rapidminer* sementara hasil akurasi *data testing* sebesar 62.50% sama persis dengan hasil perhitungan menggunakan *Rapidminer*. Hasil *AUC (Area Under Cover)* untuk *data training* sendiri menghasilkan nilai 0,992 dan masuk dalam kategori *Excellent Classification*. Untuk hasil *AUC* data testing menghasilkan nilai 0,719 dengan kategori *Fair Classification*. Pada penelitian ini menggunakan *Dataset* dari *Beast Cancer Coimbra* yang terdapat di *website UCI Machine Learning Repository*.

Referensi

Farahdiba BA, Nugroho YS. 2016. Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio. Jurnal Teknik Elektro. 8(2): 43–46.

Ma'arif F, Arifin T. 2017. Optimasi Fitur Menggunakan Backward Elimination Dan Algoritma SVM Untuk Klasifikasi Kanker Payudara. Jurnal Informatika (JI) UBSI. 4(1): 46–53. https://doi.org/10.31311/JI.V4I1.1548.

Moriesta E, Selviani, Ibrahim A. 2017. Analisis Penyaringan Email Spam Menggunakan Metode Naive Bayes. Prosiding Annual Research Seminar 2017. 3(1): 45–48.

Praningki T, Budi I .2018. Sistem Prediksi Penyakit Kanker Serviks Menggunakan Cart, Naive Bayes, dan k-NN. Creative Information Technology Journal. 4(2): 83–93. https://doi.org/10.24076/citec.2017v4i2.100.

Rifai MH, Wijayanti A. 2017. Pemanfaatan Media Pembelajaran Geografi SMA Di Kabupaten Karanganyar. Jurnal Edudikara. 2(3): 210–216.

Septiani WD. 2017. Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes

E-ISSN: 2548-3331 ; 51 – 62

Untuk Prediksi Penyakit Hepatitis. Jurnal Pilar Nusa Mandiri. 13(1): 76–84.

Via YV, Nugroho B, Syafrizal A. 2015. Sistem Pendukung Keputusan Klasifikasi Tingkat Keganasan Kanker Payudara Dengan Metode Naive Bayes Classifier. SCAN-Jurnal Teknologi Informasi Dan Komunikasi. 10(2): 63–68.

Zulfikar WB, Lukman N. 2017. Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata. Jurnal Online Informatika. 1(2): 82–86. https://doi.org/10.15575/join.v1i2.33