Jurnal Sistem Komputer dan Informatika (JSON)
Volume 1, Nomor 2, Januari 2020
e-ISSN 2685-998X
DOI 10.30865/json.v1i2.1766

Hasil Analisis Teknik Data Mining dengan Metode Naive Bayes untuk Mendiagnosa Penyakit Kanker Payudara

Elma Tiana, Sri Wahyuni

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Darwan Ali, Sampit, Indonesia
Email: [1]hyemasinara@gmail.com, [2]sri.wahyuni19738@gmail.com

Abstrak−Kanker payudara atau *Carsinoma Mammae* adalah pertumbuhan sel yang tidak terkendali pada kelenjar penghasil susu (*lobular*), saluran kelenjar dari lobular ke puting payudara (*duktus*), dan jaringan penunjang payudara yang mengelilingi lobular, duktus, pembuluh darah dan pembuluh limfe, tetapi tidak termasuk kulit payudara. Penelitian dimulai dengan melakukan tahap preprocessing, untuk menghilangkan missing values. Setelah itu dilakukan proses imputasi untuk menghilangkan missing values. Kemudian dilakukan seleksi fitur untuk melihat atribut mana yang memiliki pengaruh besar terhadap data. Tahap terakhir dilakukan klasifikasi dengan dua metode, yaitu Naïve Bayes. Di akhir penelitian dilakukan perbandingan metode yang paling baik untuk mengklasifikasi data kekambuhan pasien kanker payudara.

Kata Kunci: Kanker Payudara, Data Mining, Naive Baiyes, Gain Ratio, Weka

Abstract−Breast cancer or Mammae Carsinoma is an uncontrolled cell growth in the milk-producing glands (lobular), the gland tract from the lobular to the Breast nipple (ductus), and the breast support tissues that surround the lobular, ductus, vessels Blood and limfe vessels, but does not include breast skin. Research begins by conducting a preprocessing stage, to eliminate missing values. After that the process is imputasi to remove missing values. It then performed a feature selection to see which attribute had a major impact on the data. The last stage is classification with two methods, namely Naïve Bayes. At the end of the study, the method is best to classify the recurrence data of breast cancer patients.

Keywords: Breast Cancer, Data Mining, Naive Baiyes, Gain Ratio, Weka

1. PENDAHULUAN

Kanker payudara atau *Carsinoma Mammae* adalah pertumbuhan sel yang tidak terkendali pada kelenjar penghasil susu (*lobular*), saluran kelenjar dari lobular ke puting payudara (*duktus*), dan jaringan penunjang payudara yang mengelilingi lobular, duktus, pembuluh darah dan pembuluh limfe, tetapi tidak termasuk kulit payudara[1]. Kanker payudara menjadi pembunuh wanita terbanyak di dunia. Namun begitu, laki-laki juga bisa terkena penyakit ini, tetapi kemungkinan pada wanita 100 kali lipat dibandingkan pada laki-laki. Sebagian besar kanker payudara berasal dari sel-sel duktus (86%), kemudian lobular (12%), dan sisanya berasal dari jaringan lain (Keitel dan Kopala, 2000).

Gejala dan pertumbuhan kanker payudara tidak mudah dideteksi karena awal pertumbuhan sel kanker payudara tidak dapat diketahui dengan gejala umumnya baru diketahui setelah stadium kanker berkembang agak lanjut, karena pada tahap dini biasanya tidak menimbukan keluhan. Penderita merasa sehat, tidak merasa nyeri, dan tidak mengganggu aktivitas. Gejala-gejala kanker payudara yang tidak disadari dan tidak dirasakan pada stadium dini menyebabkan banyak penderita yang berobat dalam kondisi kanker stadium lanjut[2].

Pada penelitian terdahulu, banyak metode data mining yang telah digunakan untuk mendiagnosis penyakit Peningkatan performa algoritma naive bayes dengan gain ratio untuk klasifikasi kanker payudara oleh Muhammad Faizal Kurniawan dan Jusak Nugraha Irawan [3], Ivandari[4]. Kemudian Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio oleh Balqis Aisyah Farahdiba dan Yusuf Sulistyo Nugroho[5]. Berdasarkan penelitian terdahulu Hasil klasifikasi kanker payudara menggunakan algoritma gain ratio ini dapat diambil kesimpulan bahwa performa algoritma yang diukur berdasarkan tingkat recall, accuracy dan precision yang masing-masing memiliki nilai 92,55%, 95,17% dan 93,76% menunjukkan bahwa algoritma gain ratio sangat baik digunakan dalam penelitian ini. Sementara itu, berdasarkan decision tree yang dihasilkan dapat dilihat bahwa variabel keseragaman ukuran sel merupakan variabel yang paling berpengaruh dalam klasifikasi kanker payudara. Hal ini ditunjukkan dari variabel tersebut menempati sebagai simpul akar (root node). Kemudian Hasil penelitian klasifikasi menggunakan algoritma naive bayes untuk dataset breast cancer wisconsin memiliki tingkat akurasi sebesar 92.7%. Hasil ini sudah dianggap baik dan dengan menggunakan keseluruhan atribut data yang ada. Dengan melakukan pre processing yaitu seleksi fitur menggunakan algoritma gain ratio akurasi algoritma naive bayes naik menjadi 96,71%. Hal ini membuktikan bahwa algoritma gain ratio dapat meningkatkan performa dari algoritma naive bayes untuk klasifikasi data breast cancer wisconsin. Kenaikan tingkat akurasi yang didapatkan adalah 4,1%.

2. METODOLOGI PENELITIAN

Submitted: **20/12/2019**; Accepted: **04/01/2020**; Published: **25/01/2020**

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 1, Nomor 2, Januari 2020
e-ISSN 2685-998X
DOI 10.30865/json.v1i2.1766

Penelitian ini menggunakan algoritma *naïve bayes* untuk mempercepat proses pelatihan dalam mendiagnosis penyakit kanker payudara yang terdiri dari 10 atribut sebagai berikut:

Tabel 1. Atribut data kanker payudara

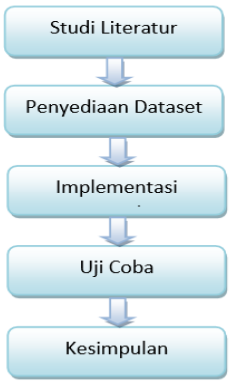
Atribut	Definisi	Record
Class	Kelas Kanker Payudara	no-recurrence-events, recurrence-events
Age	Umur pasien pengidap kanker payudara	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopouse	Masa berakhirnya siklus menstruasi	lt40, ge40, premeno
Tumor-size	Ukuran tumor payudara dalam ukuran(mm)	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 , 50-54, 55-59
Inv-nodes	Kelenjar getah bening aksila(ketiak)	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29 ,30-32, 33-35, 36-39
Node-caps	Penyebaran ke kelenjar getah bening	yes, no
Deg-malig	Tingkat keganasan	1, 2, 3
Breast	Lokasi kanker payudara	left, right
Breast-quad	Kuadran lokasi kanker	left-up, left-low, right-up , right-low, central
Irradiat	Riwayat terapi menggunakan radiasi	yes, no

Penelitian dimulai dengan melakukan tahap preprocessing, untuk menghilangkan missing values. Setelah itu dilakukan proses imputasi untuk menghilangkan missing values. Kemudian dilakukan seleksi fitur untuk melihat atribut mana yang memiliki pengaruh besar terhadap data. Tahap terakhir dilakukan klasifikasi dengan dua metode, yaitu Naïve Bayes. Di akhir penelitian dilakukan perbandingan metode yang paling baik untuk mengklasifikasi data kekambuhan pasien kanker payudara.

Untuk mendapatkan informasi dan data-data pendukung, maka metode pengumpulan data yang diterapkan adalah sebagai berikut:

1. Studi literatur, yaitu dengan literatur review naratif. Kami melakukan kajian komprehensif dari literatur untuk mengidentifikasi buku, file atau dokumen kanker payudara. Kajian literatur dilakukan pada penelitian yang terpublikasi secara online.
2. Observasi sebagai bahan penelitian menggunakan dataset dari https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Berikut adalah alur yang dilakukan dalam menganalisis penelitian ini :



Gambar 1. Tahapan Penelitian

Selanjutnya implementasi metode dan pengujian dilakukan dengan tool Weka versi 3.6.9 untuk melakukan skema klasifikasi dengan Naïve Bayes. Parameter uji yang akan digunakan antara lain adalah: TP Rate, FP Rate, Precision, Recall dan F Measure. Sedangkan dua macam scenario uji yang akan digunakan adalah training tanpa cross validasi dan training dengan cross validasi.

3. HASIL DAN PEMBAHASAN

Tabel 2. Missing Values

Nilai	Nilai	Kelas	Frekuensi	Jumlah

Submitted: **20/12/2019**; Accepted: **04/01/2020**; Published: **25/01/2020**

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 1, Nomor 2, Januari 2020
e-ISSN 2685-998X
DOI 10.30865/json.v1i2.1766

	Atribut			
Deg-malig	1	'norecurrenceevents'	59	71
		'recurrenceevents'	12	
	2	'norecurrenceevents'	102	130
		'recurrenceevents'	28	
	3	'norecurrenceevents'	40	85
		'recurrenceevents'	45	
Total			286	286

Missing Values merupakan tahap dalam preprocessing, tahap ini digunakan untuk mencari apakah terdapat value (nilai) data atribut yang kosong dalam setiap baris data. Hal ini dapat terjadi baik pada data bertipe numerik dan kategorik. Dalam hal ini penelitian dilakukan pada data bertipe kategorik. Pada data kekambuhan pasien kanker payudara ini terdapat beberapa atribut yang memiliki missing values, sehingga harus dilakukan imputation[6]. Berikut daftar missing values (nilai yang hilang) setiap atribut:

Tabel 3. Daftar missing values

Atribut	Jumlah *Missing Value*
Node-caps	8
Breast-quad	1

Untuk mengisi nilai yang hilang tersebut dapat menggunakan means/ median/ modus. Jika berupa data kategorikal maka dapat diterapkan menggunakan modus. Setelah data telah dilakukan imputasi missing values, maka tahap selanjutnya adalah mengukur atribut-atribut mana saja yang memiliki kontribusi besar pada data dengan cara meranking atribut menggunakan metode tertentu. Pada penelitian ini, ranking atribut menggunakan metode Information Gain. Information Gain merupakan salah satu metode yang digunakan untuk mencari seberapa potensial informasi pada atribut masing-masing dalam data. Output dari metode tersebut adalah berupa score setiap atribut yang merepresentasikan urutan atribut berkontribusi besar sampai terkecil[6]. Berikut ini adalah cara mengevaluasi atribut yakni berdasarkan perolehan frekuensi “tiap nilai atribut”.

3.1 Implementasi

Berikut ini merupakan hasil klasifikasi data kekambuhan pasien kanker payudaya dengan algoritma Naïve Bayes menggunakan tools Weka.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         208               72.7273 %
Incorrectly Classified Instances        78               27.2727 %
Kappa statistic                          0.2996
Mean absolute error                      0.3268
Root mean squared error                  0.4534
Relative absolute error                 77.8451 %
Root relative squared error             98.8513 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.856     0.576      0.778     0.856     0.815      0.699    no-recurrence-events
                 0.424     0.144      0.554     0.424     0.48       0.699    recurrence-events
Weighted Avg.    0.727     0.448      0.712     0.727     0.716      0.699

=== Confusion Matrix ===

   a   b   <-- classified as
 172  29 |   a = no-recurrence-events
  49  36 |   b = recurrence-events
```

Gambar 3. Hasil Klasifikasi dengan algoritma Naïve Bayes

4. KESIMPULAN

Berdasarkan uraian-uraian diatas pada pembahasan sebelumnya, dapat ditarik kesimpulan bahwa klasifikasi kekambuhan pasien kanker payudara menggunakan algoritma Naïve Bayes 72,7%. Proses imputasi berdampak baik pada algoritma Naïve Bayes yaitu penambahan nilai akurasi sebesar 1, 02%.

Submitted: **20/12/2019**; Accepted: **04/01/2020**; Published: **25/01/2020**

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 1, Nomor 2, Januari 2020
e-ISSN 2685-998X
DOI 10.30865/json.v1i2.1766

REFERENCES

[1] R. DAMAYANTI, PENGARUH PELAKSANAAN PEMERIKSAAN PAYUDARA SENDIRI (SADARI) TERHADAP PENGETAHUAN DAN KEMAMPUAN SISWI DALAM UPAYA DETEKSI DINI KANKER PAYUDARA DI SMP NEG.1 SIBULUE KAB. BONE. MAKASSAR, 2017.
[2] A. K. Omega Memed, KORELASI ANTARA C-reactive Protein DAN PARAMETER TROMBOSIT PADA PASIEN KANKER PAYUDARA. SURAKARTA, 2019.
[3] M. F. Kurniawan and J. Nugraha Irawan, "Peningkatan Performa Algoritma Naive Bayes dengan Gain Ratio untuk Klasifikasi Keganasan Kanker Payudara," 2018.
[4] M. F. Kurniawa and Ivandari (last), "KOMPARASI ALGORITMA DATA MINING UNTUK KLASIFIKASI PENYAKIT KANKER PAYUDARA," 2017.
[5] Balqis Aisyah Farahdiba and Yusuf Sulistyo Nugroho, "Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio," 2016.
[6] Ai Rita Rizqiah and Agus Subekti, "PREDIKSI KEKAMBUHAN KANKER PAYUDARA DENGAN ALGORITMA C4.5," 2018.

Submitted: **20/12/2019**; Accepted: **04/01/2020**; Published: **25/01/2020**