

CLASSIFICATION OF BREAST CANCER USING THE NAIVE BAYES ALGORITHM AT RAA SOEWONDO PATI

KLASIFIKASI PENYAKIT KANKER PAYUDARA MENGGUNAKAN ALGORITMA NAIVE BAYES DI RAA SOEWONDO PATI

^{1a}Dini Fahrani, ^{1a}Bijanto M.Kom, ^{1a}Ellen Proborini M.Pd

^{1a}Program Studi Informatika, Sekolah Tinggi Teknik Pati

e-mail: dinifahrani57@gmail.com, bijanto@sttp.ac.id, ellena@sttp.ac.id

Abstract Breast cancer is a malignant disease that attacks breast tissue which can originate from glandular cells, glandular ducts, and breast supporting tissues. Based on this research, breast cancer is a very deadly disease for women. In this study classify the types of benign and malignant breast cancer. The method used in this research is Naive bayes. Naive bayes is one of the ten best algorithms in data mining. This study obtained the highest accuracy, namely 97.78% and an error of 2.22%. With a comparison of training and testing data 60:40.

Keywords: breast cancer, classification, Naive bayes algorithm.

Abstrak Kanker payudara merupakan suatu penyakit keganasan yang menyerang jaringan payudara yang bisa berasal dari sel kelenjar, saluran kelenjar, dan jaringan penunjang payudara. Berdasarkan penelitian tersebut kanker payudara merupakan penyakit yang sangat mematikan bagi wanita. Pada penelitian ini mengklasifikasikan jenis kanker payudara jinak dan ganas. Metode yang digunakan dalam penelitian ini adalah Naive bayes. Naive bayes merupakan salah satu dari sepuluh algoritma terbaik dalam data mining. Penelitian ini mendapatkan akurasi tertinggi yaitu 97.78% dan error 2,22%. Dengan perbandingan data training dan testing 60:40.

Kata Kunci: kanker payudara, klasifikasi, algoritma Naive bayes.

PENDAHULUAN

Angka penderita kanker di Indonesia dapat dikatakan meningkat secara fantastis setiap tahunnya (Lestari & Wulansari, 2018). Ditunjukkan dengan data pada *Global Cancer Observatory 2018* dari *World Health Organization (WHO)* menunjukkan kasus kanker yang paling banyak terjadi di Indonesia adalah kanker payudara, yakni 58.256 kasus atau 16,7% dari total 348.809 kasus kanker (Zega & Pangemanan, 2020). Kementerian Kesehatan (Kemenkes) menyatakan, angka kanker payudara di Indonesia mencapai 42,1 orang per 100 ribu penduduk. Rata-rata kematian akibat kanker ini mencapai 17 orang per 100 ribu penduduk (Globocan, 2020). Data Globocan tahun 2020, jumlah kasus baru kanker payudara mencapai 68.858 kasus (16,6%) dari total 396.914 kasus baru kanker di Indonesia. Sementara itu, untuk jumlah kematiannya mencapai lebih dari 22 ribu jiwa kasus (Rokom, n.d.). Maka dari itu salah satu kanker yang angka kematiannya tinggi adalah kanker payudara. Penyakit kanker sudah menjadi salah satu penyakit

paling mematikan bagi umat manusia di seluruh dunia. Penelitian sebelumnya yang dilakukan oleh Athalla mengenai Klasifikasi Penyakit Kanker Payudara dengan Metode K-NN untuk mengklasifikasi dataset. Penelitian ini menggunakan perhitungan jarak kemiripan menggunakan jarak minkowski yang telah ditentukan dari kedua jarak Euclidean dan jarak Manhattan. Penelitian ini mendapatkan hasil dengan tingkat akurasi sebesar 93%(Athalla et al., 2018). Penelitian lain yang dilakukan oleh Chazar mengenai diagnosis kanker payudara dengan menggunakan Algoritma Support Vector Machine yang hanya mendapatkan hasil tingkat akurasi sebesar 88,733%(Chazar, 2020). Penelitian yang lain dari Tiana dengan menggunakan Algoritma lain yaitu *Naive bayes* hanya mendapatkan hasil nilai akurasi sebesar 72,7%(Aisyah & Sulisty, 2016). Berdasarkan dari ketiga penelitian yang terkait dengan kanker payudara, pada penelitian ini akan menerapkan metode lain yang diharapkan dapat menghasilkan akurasi yang lebih baik dari penelitian sebelumnya. *Naive bayes Classifier* atau sering disebut Bayesian Classification adalah metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Naive bayes Classifier* (NBC) terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar(Murdifin Haming et al., 2019). Dengan berbagai kelebihan dari metode *Naive bayes* yang telah dijelaskan pada beberapa pernyataan sebelumnya, maka hal tersebut yang mendasari penelitian ini untuk mengaplikasikan algoritma *Naive bayes* untuk dijadikan metode dalam mengklasifikasi kanker payudara.

METODE PENELITIAN

1. Algoritma *Naive Bayes*

Penelitian ini menggunakan metode algoritma *Naive bayes* untuk mengklasifikasi metode *Naive bayes* itu sendiri merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Metode ini menggunakan probabilitas bersyarat sebagai dasarnya. Dalam ilmu statistik, probabilitas bersyarat dinyatakan probabilitas X di dalam Y adalah probabilitas interaksi X dan Y dari probabilitas Y, atau dengan bahasa lain $P(X|Y)$ adalah prosentase banyaknya X di dalam Y.

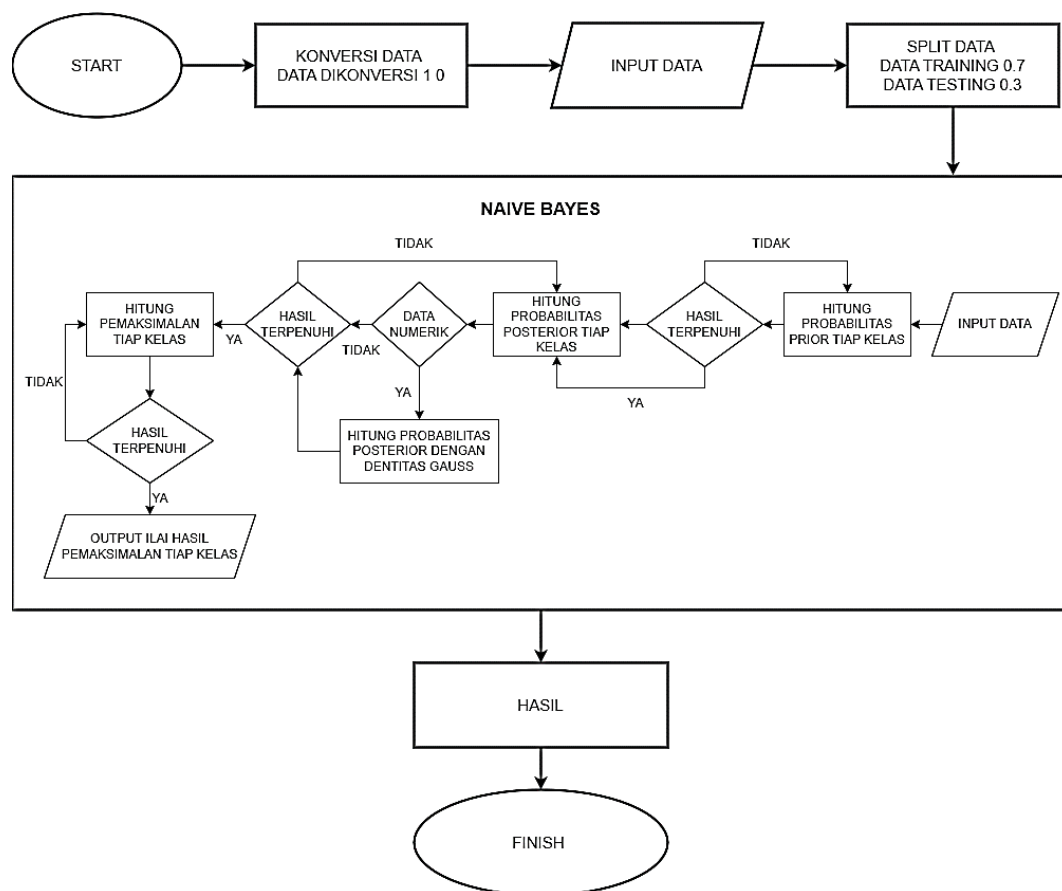
2. Pengumpulan Data

Penelitian ini menggunakan data sekunder, Data sekunder adalah informasi yang dikumpulkan oleh lembaga yang menyediakan informasi dan setelah itu informasi tersebut dibuka untuk umum secara luas(Athalla et al., 2018). Dalam penelitian ini data yang didapatkan dari RM RSUD RAA Soewondo Pati. Data yang diambil dalam penelitian ini termasuk dalam data yang dapat digunakan untuk klasifikasi. Data tersebut terdiri dari 114 record data dan 10 atribut.

3. Target/Sasaran Penelitian

Target dalam penelitian yang ingin dicapai yaitu algoritma yang diimplementasikan dalam hal ini algoritma *Naive bayes* dapat menghasilkan klasifikasi dengan akurasi yang tinggi. *Naive bayes Classifier* (NBC) terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar(Abdominal et al., 2022).

4. Teknik Analisis Data



Gambar 1. Teknis Analisis Data

a. *Start*

Tahap *start*, maka seluruh hal yang dibutuhkan guna penelitian akan dipersiapkan termasuk alat, bahan, dan dataset yang digunakan dalam penelitian ini

b. *Konversi Data*

Proses konversi data merupakan proses dimana data yang awalnya bernilai "Ya" dan "Tidak" dikonversi menjadi "1" dan "0" supaya data dapat diolah menggunakan algoritma.

c. *Input Data*

Proses input data merupakan proses dimana data akan diinput untuk kemudian data akan diolah dengan menggunakan algoritma *Naive bayes*

d. *Split Data*

Setelah data diinput langkah selanjutnya adalah split data yang akan membagi data menjadi 2 yaitu training dan testing. Proses split data berguna untuk meringankan komputasi karena data tidak diolah semua. Data yang terbagi menjadi data training akan dijadikan sebagai model pembelajaran sementara data yang tergolong dalam data testing akan menjadi data uji untuk mengetahui performa model klasifikasi yang

digunakan. Pada penelitian ini, rasio data training dan testing adalah 70% untuk data training dan 30% untuk data testing.

e. Klasifikasi Menggunakan *Naive Bayes*

Pada tahap ini, algoritma *Naive Bayes* akan diterapkan untuk mengklasifikasi data. Tahap ini akan mengetahui hasil dari akurasi dan error pada data. Proses dari klasifikasi menggunakan algoritma *Naive Bayes* meliputi:

1. *Start*
2. *Input data*
3. Hitung probabilitas prior tiap kelas
4. Jika hasil tidak terpenuhi kembali ke proses perhitungan probabilitas prior tiap kelas
5. Jika hasil terpenuhi maka lanjut ke proses hitung probabilitas posterior tiap kelas
6. Setelah perhitungan probabilitas posterior tiap kelas maka akan mendapatkan data numerik.
7. Jika tidak kembali ke proses perhitungan probabilitas tiap kelas
8. Jika hasil terpenuhi, maka hitung probabilitas dengan dentitas gaus
9. Jika perhitungan probabilitas terpenuhi maka lanjut ke proses pemaksimalan tiap kelas
10. Jika hasil tidak terpenuhi kembali ke proses perhitungan pemaksimalan tiap kelas
11. Jika *output* terpenuhi maka didapatkan *output* pemaksimalan tiap kelas

f. Hasil Klasifikasi

Setelah proses klasifikasi selesai didapatkan hasil dari perhitungan algoritma *Naive Bayes*. Pada penelitian ini parameter output yang digunakan adalah akurasi dan error.

HASIL DAN PEMBAHASAN

1. Hasil

a. Data Penelitian

Menunjukan sampel dari data penyakit kanker payudara yang digunakan dalam penelitian ini

Tabel 1. *Sample Data*

No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Hasil
1	Ya	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Mamae jinak
2	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Mamae jinak
3	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Mamae jinak
4	Ya	Ya	Ya	Tidak	Tidak	Tidak	Ya	Ya	Tidak	Tidak	Mamae ganas
5	Ya	Ya	Ya	Tidak	Tidak	Tidak	Ya	Ya	Tidak	Tidak	Mamae ganas
6	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Tidak	Mamahe ganas
7	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Mamae jinak
8	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Tidak	Mamae ganas
9	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Mamae jinak
10	Ya	Ya	Ya	Ya	Tidak	Ya	Ya	Ya	Ya	Tidak	Mamae ganas

...

112	Tidak	Ya	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Ya	Tidak	Mamae jinak
113	Ya	Ya	Ya	Tidak	Tidak	Tidak	Ya	Ya	Ya	Ya	Mamae ganas
114	Tidak	Ya	Tidak	Ya	Tidak	Ya	Ya	Ya	Tidak	Ya	Mamae jinak

Ket dari atribut table 1

A1 = nyeri payudara

A2 = benjolan

A3 = cairan

A4 = pusing

A5 = mual

A6 = muntah

A7 = perubahan berat badan

A8 = perubahan asupan makan

A9 = insomnia

A10 = kelancaran BAB dan BAK

A11 = Hasil

Data pada table 1 masih belum dapat diolah karena masih dalam bentuk huruf (YA/TIDAK), maka dari itu perlu langkah konversi kedalam bentuk angka. *Record* data yang bernilai 'YA' akan dikonversi dengan nilai "1" dan *record* data yang bernilai 'TIDAK' akan dikonversi dengan nilai "0".

b. Konversi data

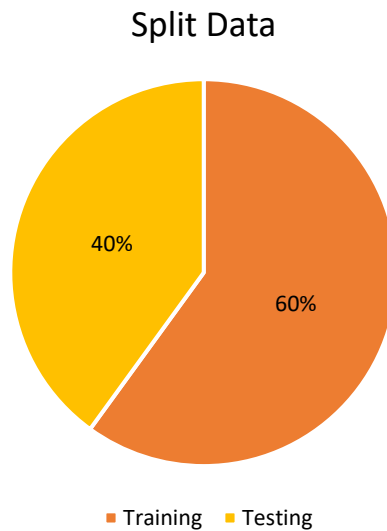
Tabel 2 merupakan hasil dari konversi data dari table 1. data hasil konversi sudah dapat diolah menggunakan algoritma. Selanjutnya data hasil konversi akan dibagi menjadi 2 yaitu testing dan training.

Tabel 2. Hasil Konversi Data

No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Hasil
1	1	0	1	0	0	0	0	0	0	1	Mamae jinak
2	1	1	0	0	0	0	0	0	0	1	Mamae jinak
3	1	1	1	0	0	0	0	0	0	1	Mamae jinak
4	1	1	1	0	0	0	1	1	0	0	Mamae ganas
5	1	1	1	0	0	0	1	1	0	0	Mamae ganas
...											
112	0	1	0	1	0	0	0	0	1	0	Mamae jinak
113	1	1	1	0	0	0	1	1	1	1	Mamae ganas
114	0	1	0	1	0	1	1	1	0	1	Mamae jinak

a. Split Data

Split data pada penelitian ini membagi menjadi data *training* dan *testing* menggunakan rasio 60:40. Artinya, data *training* yang digunakan sebesar 60 % dan data *testing* yang digunakan sebesar 40 %. Data *training* dan *testing* dengan rasio 60:40 dianjurkan dalam data mining karena dianggap menghasilkan kinerja yang bagus (Chazar, 2020). Ilustrasi pada *split* data dapat dilihat pada gambar 2.



Gambar 2. Ilustrasi *Split* Data
(sumber peneliti)

b. Klasifikasi data

Pada tahap ini data akan diklasifikasi menggunakan *Naive bayes* formula yang digunakan untuk perhitungan *Naive bayes* adalah sebagai berikut :

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (1)$$

Hasil klasifikasi menggunakan *Naive bayes* diklasifikasi tepat seperti pada data awal. Artinya, sama seperti data ID ke 2 bahwa ID ke 5 juga termasuk dalam salah satu data yang akan digunakan sebagai acuan dalam menghitung nilai dari akurasi data. Tabel 3 merupakan hasil klasifikasi menggunakan algoritma *Naive bayes*.

Table 3. Hasil Klasifikasi Menggunakan *Naive Bayes*

NO	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	HASIL
2	1	1	0	0	0	0	0	0	0	1	MAMAE JINAK
4	1	1	1	0	0	0	1	1	0	0	MAMAE GANAS
5	1	1	1	0	0	0	1	1	0	0	MAMAE GANAS
7	1	0	0	1	0	0	0	0	0	1	MAMAE JINAK
8	1	1	1	1	1	1	1	1	1	0	MAMAE GANAS
24	1	1	0	0	0	0	0	0	0	1	MAMAE JINAK
25	0	1	0	0	0	0	0	0	0	1	MAMAE JINAK
29	1	0	0	1	0	0	0	0	0	0	MAMAE JINAK
31	1	0	1	0	0	0	0	0	0	0	MAMAE JINAK
41	1	1	1	0	0	0	0	0	0	1	MAMAE GANAS
...											
100	1	1	0	0	0	0	0	0	0	1	MAMAE JINAK

104	1	1	1	0	0	0	1	1	1	0	MAMAE GANAS
114	0	1	0	1	0	1	1	1	0	1	MAMAE JINAK

(Sumber : Peneliti)

1. Pembahasan

Berdasarkan hasil pengolahan menggunakan algoritma *Naive bayes*, selanjutnya yaitu melakukan pengujian dari hasil klasifikasi menggunakan algoritma *Naive bayes*. Untuk menentukan hasil pengujian dari klasifikasi yaitu menggunakan *confussion matrix*. Tabel 3 berikut merupakan hasil dari *confussion matrix* dari hasil klasifikasi penyakit kanker payudara menggunakan algoritma *Naive bayes* pada penelitian ini.

a. Menggunakan rasio *training* 60% dan *testing* 40%

Table 4. Confussion Matrix Rasio Training 60% dan Testing 40%

	True MAMAE JINAK	True MAMAE GANAS
Pred. MAMAE JINAK	20	0
Pred. MAMAE GANAS	1	24

Diketahui dalam tabel yang disajikan dalam tabel data yang diblok warna hijau merupakan *true positif*, kemudian data yang diblok warna kuning merupakan *false negatif*, data yang diblok warna biru merupakan *false positif*, dan data yang diblok warna merah merupakan *true negatif*. Berdasarkan pernyataan tersebut, maka untuk menghitung akurasi dan *error* yaitu sebagai berikut:

- Akurasi data

$$\begin{aligned} \text{Akurasi} &= \frac{TP+TN}{TP+TN+FN+FP} \\ \text{Akurasi} &= \frac{20+24}{20+24+0+1} \\ \text{Akurasi} &= \frac{44}{45} \\ \text{Akurasi} &= 0,9778 \end{aligned}$$

Berdasarkan perhitungan tersebut, maka hasil klasifikasi menggunakan algoritma *Naive bayes* dengan rasio *training* 60% dan *testing* 40% menghasilkan akurasi sebesar 0,9778 atau 97,78%.

- Error

$$\begin{aligned} \text{Error} &= \frac{FP+FN}{TP+TN+FP+FN} \\ \text{Error} &= \frac{1+0}{20+24+0+1} \\ \text{Error} &= \frac{1}{45} \end{aligned}$$

$$\text{Error} = 0,0222$$

Berdasarkan perhitungan tersebut, maka hasil klasifikasi menggunakan algoritma *Naive bayes* dengan rasio *training* 60% dan *testing* 40% menghasilkan *error* sebesar 0,0222 atau 2,22%.

KESIMPULAN

Kesimpulan yang dapat diambil dari penelitian yang telah dilaksanakan yaitu sebagai berikut:

1. Pada metode sebelumnya mengenai klasifikasi penyakit kanker payudara oleh Athalla,dkk (2018) menggunakan algoritma K-NN, Chazar (2020) menggunakan SVM, dan Tiana,dkk (2020) menggunakan *Naive Bayes*. Pada penelitian ini mengklasifikasi kanker payudara menggunakan *Naive bayes* dari data yang didapatkan di RSUD RAA Soewondo Pati.
2. Pada penelitian Athalla,dkk (2018) menghasilkan akurasi sebesar 93% menggunakan nilai $k=3$, oleh Chazar menggunakan Algoritma Support Vector Machine mendapatkan hasil akurasi sebesar 88,733%, berikutnya dari Tiana dengan menggunakan Algoritma *Naive bayes* mendapatkan hasil nilai akurasi sebesar 72,7%. Sedangkan pada penelitian ini mendapatkan akurasi 97,78%, artinya penelitian ini mengalami peningkatan 4,78% dari penelitian Athalla, 9,05% dari penelitian Chazar, dan 25,08% dari penelitian Tiana.
3. Penelitian yang telah dilaksanakan berhasil mendapatkan akurasi tertinggi yaitu 97,78 dengan alokasi data *training* dan *testing* 60%:40% menggunakan algoritma *Naive bayes* dengan nilai *error* sebesar 2,22%.
4. Pada penelitian ini alokasi data *training* dan *testing* yang digunakan awalnya 70%:30%. Dan mendapatkan akurasi 97,06 dengan *error* 2,94.

DAFTAR PUSTAKA

- Abdominal, P., Exercise, S., Penurunan, T., & Nyeri, I. (2022). *Fakumi medical journal*. 2(5), 359–367.
- Aisyah, B., & Sulisty, Y. (2016). Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio. *Jurnal Teknik Elektro*, 8(2), 43–46.
- Atthalla, I. N., Jovandy, A., & Habibie, H. (2018). Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor. *Prosiding Annual Research Seminar*, 4(1), 148–151.
- Chazar, C. (2020). Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine. *INFORMASI (Jurnal Informatika Dan Sistem Informasi)*, 12(1), 67–80. <https://doi.org/10.37424/informasi.v12i1.48>
- Globocan. (2020). Cancer Incident in Indonesia. *International Agency for Research on Cancer*, 858, 1–2.
- Lestari, P., & Wulansari. (2018). Pentingnya Pemeriksaan Payudara Sendiri (SADARI) Sebagai Upaya Deteksi Dini Kanker Payudara. *Indonesian Journal of Community Empowerment (IJCE)*, 1161, 55–58.
- Murdifin Haming, A. Zulfikar Syaiful, Aditya Halim Perdana Kusuma Putra, & Imaduddin Murdifin. (2019). The Application of SERVQUAL Distribution In Measuring Customer Satisfaction of Retails Company. *Journal of Distribution Science*, 17(2), 25–31. <https://doi.org/10.15722/jds.17.2.201902.25>
- Rokom. (n.d.). *No Title*. Biro Komunikasi & Pelayanan Publik Kementerian Kesehatan RI.

Zega, W. S. H., & Pangemanan, A. (2020). Gambaran kualitas hidup holistik pasien kanker payudara yang menjalani kemoterapi di satu rumah sakit swasta di Indonesia. *Nursing Current*, 8(2), 194–204.