JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

Implementasi RapidMiner untuk Klasifikasi Risiko Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Cross Validation

RapidMiner Implementation for Breast Cancer Risk Classification Using Naive Bayes Method Based on Cross Validation

M.Fery Ardiansyah[1*], Hasbi Firmansyah[2], Wahyu Asriyani[3], Rizki Prasetyo Tulodho[4]
Universitas Pancasakti Tegal
Email: mferyardiansyah43@gmail.com[1], hasbifirmansyah@upstegal.ac.id[2], asriyani1409@gmail.com[3]*
Rizki.prasetyo.tulodo@gmail.com[4]

Article Info

Article history :
Received : 19-12-2025
Revised : 20-12-2025
Accepted : 22-12-2025
Pulished : 25-12-2025

Abstract

Predicting breast cancer recurrence after mastectomy is a complex clinical challenge due to the interaction of various patient biological variables. This prognostic uncertainty demands an accurate medical decision support system to determine the urgency of further treatment. This study aims to develop a recurrence risk classification model using the Naive Bayes Classifier algorithm, chosen for its superiority in handling conditional probabilities on nominal attribute data. The dataset used is sourced from the UCI Machine Learning Repository (Institute of Oncology, Ljubljana) and consists of 286 medical records, including nine categorical predictor attributes: tumor-size, inv-nodes, and deg-malignancy. This dataset faces challenges in the form of missing values and class imbalance, with a proportion of 201 non-recurrence cases compared to 85 recurrence cases. This study applies the Knowledge Discovery in Database (KDD) methodology using RapidMiner Studio software. Pre-processing steps include manual attribute naming and imputation of missing data using statistical methods. Model validation was rigorously performed using the 10-Fold Cross-Validation method to minimize evaluation bias. Experimental results showed that the Naive Bayes model produced an Accuracy rate of [72.03%], Precision of [78.14%], and Recall of [83.58%]. The high accuracy values, but with variations in recall values, indicate the influence of data imbalance on the model's sensitivity in detecting positive cases. In conclusion, Naive Bayes has proven effective and computationally efficient for small-dimensional medical datasets with categorical features. However, data balancing techniques (resampling) are recommended for further research to improve detection of minority classes.

Keywords : Breast Cancer, Data Mining, Naive Bayes

Abstrak

Prediksi kekambuhan (*recurrence*) kanker payudara pasca-mastektomi merupakan tantangan klinis yang kompleks karena melibatkan interaksi berbagai variabel biologis pasien. Ketidakpastian prognosis ini menuntut adanya sistem pendukung keputusan medis yang akurat untuk menentukan urgensi pengobatan lanjutan. Penelitian ini bertujuan untuk membangun model klasifikasi risiko kekambuhan menggunakan algoritma *Naive Bayes Classifier*, yang dipilih karena keunggulannya dalam menangani probabilitas bersyarat pada data atribut nominal. Dataset yang digunakan bersumber dari *UCI Machine Learning Repository* (Institute of Oncology, Ljubljana) yang terdiri dari 286 rekam medis, mencakup 9 atribut prediktor kategorikal seperti *tumor-size*, *inv-nodes*, dan *deg-malig*. Dataset ini memiliki tantangan berupa *missing values* dan ketidakseimbangan kelas (*class imbalance*), dengan proporsi 201 kasus *no-recurrence* berbanding 85 kasus *recurrence*. Penelitian ini menerapkan metodologi *Knowledge Discovery in Database* (KDD) menggunakan perangkat lunak RapidMiner Studio. Tahapan pra-pemrosesan meliputi penamaan atribut manual dan imputasi data yang hilang menggunakan modus statistik. Validasi model dilakukan secara

ketat menggunakan metode *10-Fold Cross-Validation* untuk meminimalisir bias evaluasi. Hasil eksperimen menunjukkan bahwa model Naive Bayes menghasilkan tingkat Akurasi sebesar [72,03 %], Presisi sebesar [78,14 %], dan Recall sebesar [83,58 %]. Tingginya nilai akurasi namun dengan variasi pada nilai *recall* mengindikasikan pengaruh ketidakseimbangan data terhadap sensitivitas model dalam mendeteksi kasus positif. Kesimpulannya, Naive Bayes terbukti efektif dan komputasional efisien untuk dataset medis berdimensi kecil dengan fitur kategorikal, namun teknik penyeimbang data (*resampling*) disarankan untuk penelitian lanjutan guna meningkatkan deteksi pada kelas minoritas.

Kata Kunci: ***Breast Cancer, Data Mining, Naive Bayes***

PENDAHULUAN

Kanker payudara merupakan tantangan kesehatan global yang signifikan, tercatat sebagai jenis kanker yang paling sering didiagnosis pada wanita dan menjadi salah satu penyebab utama mortalitas di seluruh dunia. Berdasarkan estimasi statistik global, insidensi penyakit ini terus menunjukkan tren peningkatan setiap tahunnya di berbagai negara(Bray et al., 2018). Permasalahan klinis yang krusial dalam penanganan penyakit ini adalah ketidakpastian prognosis, khususnya dalam memprediksi risiko kekambuhan (*recurrence*) pasca-mastektomi. Prediksi ini menjadi kompleks karena melibatkan interaksi berbagai variabel biologis pasien yang sulit dianalisis secara manual, sehingga menuntut adanya sistem pendukung keputusan medis yang akurat untuk menentukan urgensi pengobatan lanjutan[Caballé-cervigón].

Dalam era informatika medis, akumulasi data rekam medis yang besar di rumah sakit menyimpan pola tersembunyi yang berharga untuk pengambilan keputusan klinis. Teknik *Data Mining* melalui metodologi *Knowledge Discovery in Database* (KDD) menawarkan solusi efektif untuk mengekstraksi pengetahuan dari basis data tersebut[Caballé-cervigón, 2018]. Di antara berbagai teknik klasifikasi, algoritma *Naive Bayes Classifier* dipilih karena keunggulannya dalam menangani probabilitas bersyarat, khususnya pada data dengan atribut nominal(Kumar & Goswami, 2024). Meskipun memiliki asumsi independensi antar fitur yang sederhana, algoritma ini terbukti memiliki kinerja yang kompetitif dan efisien secara komputasional dibandingkan algoritma yang lebih kompleks, serta efektif untuk dataset medis berdimensi kecil (Kharya, 2012).

Namun, penerapan model klasifikasi pada data medis sering kali dihadapkan pada tantangan kualitas data. Penelitian ini menggunakan dataset dari *UCI Machine Learning Repository* yang memiliki tantangan berupa adanya *missing values* dan ketidakseimbangan kelas (*class imbalance*), di mana proporsi kasus tidak kambuh (*no-recurrence*) jauh lebih dominan dibandingkan kasus kambuh (*recurrence*)(Bhandari et al., 2015). Kondisi keterbatasan data ini berpotensi menimbulkan bias dalam evaluasi model. Oleh karena itu, diperlukan validasi yang ketat menggunakan metode *10-Fold Cross-Validation* untuk meminimalisir bias pengukuran performa dan memastikan model yang dihasilkan valid serta terhindar dari *overfitting* .

Kegunaan penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi dini prognosis kanker yang membantu tenaga medis memetakan pasien ke dalam kelas risiko yang tepat berdasarkan atribut klinisnya. Berdasarkan uraian permasalahan dan potensi solusi tersebut, penelitian ini bertujuan untuk membangun model klasifikasi risiko kanker payudara menggunakan algoritma *Naive Bayes* yang diimplementasikan melalui perangkat lunak RapidMiner, serta mengukur performanya secara akurat melalui validasi berbasis *Cross Validation*(Tsamardinos et al., 2018).

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

METODE PENELITIAN

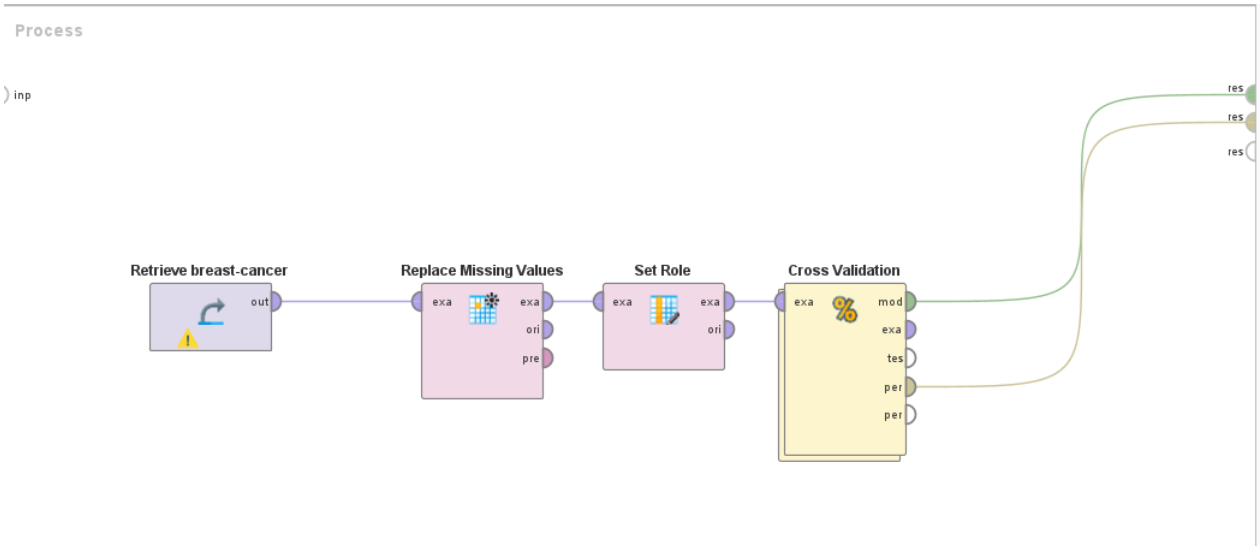
1. Uraian Masalah dan Alasan Penelitian

Prediksi kekambuhan (*recurrence*) kanker payudara pasca-mastektomi merupakan tantangan klinis yang kompleks karena melibatkan interaksi berbagai variabel biologis pasien. Ketidakpastian prognosis ini menuntut adanya sistem pendukung keputusan medis yang akurat untuk menentukan urgensi pengobatan lanjutan. Selain itu, data rekam medis yang tersedia memiliki tantangan tersendiri berupa adanya nilai yang hilang (*missing values*) dan ketidakseimbangan kelas (*class imbalance*), di mana proporsi kasus tidak kambuh jauh lebih banyak dibandingkan kasus kambuh. Penelitian ini bertujuan membangun model klasifikasi risiko kekambuhan menggunakan algoritma *Naive Bayes Classifier* karena keunggulannya dalam menangani probabilitas bersyarat pada data atribut nominal dan efisiensi komputasinya[M. M. Degree, 2012].

2. Bahan dan Alat

Penelitian ini menggunakan bahan berupa data sekunder yang terdiri dari 286 rekam medis pasien. Dataset ini mencakup 9 atribut prediktor kategorikal (seperti *tumor-size*, *inv-nodes*, *deg-malig*) dan 1 atribut kelas target (*recurrence-events* atau *no-recurrence-events*).

Alat bantu utama yang digunakan untuk pengolahan dan analisis data adalah perangkat lunak RapidMiner Studio. Perangkat lunak ini dipilih karena kemampuannya melakukan proses *Knowledge Discovery in Database* (KDD) secara visual dan terintegrasi.



gambar 1.dataset klasifikasi naive bayes di rapidminer

3. Lokasi Penelitian dan Sumber Data

Mengingat penelitian ini berbasis komputasi menggunakan data sekunder, lokasi sumber data berasal dari University Medical Centre, Institute of Oncology, Ljubljana, Slovenia. Dataset tersebut dipublikasikan secara global melalui repositori data publik.

4. Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini adalah studi dokumentasi dengan mengambil dataset sekunder bernama "Breast Cancer Data Set". Dataset ini diperoleh

dari UCI Machine Learning Repository yang dipublikasikan oleh Zwitter dan Soklic. Data yang dikumpulkan bersifat historis dan tidak melibatkan intervensi langsung kepada pasien (perancangan percobaan klinis), melainkan pemanfaatan data rekam medis yang sudah tersedia.

5. Analisis Data

Penelitian ini menerapkan metodologi *Knowledge Discovery in Database* (KDD) dengan tahapan analisis sebagai berikut:

a. Pra-pemrosesan Data (*Preprocessing*): Tahap ini meliputi penamaan atribut secara manual (seperti *Class, age, menopause*, dll) karena dataset asli tidak memiliki *header*. Selanjutnya, dilakukan penanganan data yang hilang (*missing values*) yang ditandai simbol "?" menggunakan operator *Replace Missing Values*, di mana nilai yang hilang diganti dengan modus (nilai yang paling sering muncul) statistik. Atribut "Class" kemudian ditetapkan sebagai label atau target prediksi.

b. Pemodelan (*Modeling*): Algoritma yang digunakan adalah *Naive Bayes Classifier*, sebuah metode klasifikasi statistik yang memprediksi probabilitas keanggotaan kelas dengan asumsi independensi antar atribut. Algoritma ini dipilih karena terbukti efektif untuk dataset medis berdimensi kecil dengan fitur kategorikal.

c. Validasi Model: Validasi dilakukan secara ketat menggunakan metode 10-Fold Cross-Validation. Data dibagi menjadi 10 bagian, di mana 9 bagian digunakan untuk pelatihan (*training*) dan 1 bagian untuk pengujian (*testing*) secara bergantian untuk meminimalisir bias evaluasi.

d. Evaluasi Kinerja: Performa model diukur menggunakan *Confusion Matrix* untuk menghitung nilai Akurasi, Presisi, dan *Recall* (Sensitivitas) guna mengetahui seberapa baik model mendeteksi kasus positif kekambuhan.

HASIL DAN PEMBAHASAN

1. Hasil Eksperimen

Eksperimen klasifikasi risiko kanker payudara dilakukan menggunakan dataset yang terdiri dari 286 rekam medis pasien. Pengujian model dilakukan menggunakan metode *10-Fold Cross Validation*, di mana data dibagi menjadi 10 bagian untuk memastikan validitas evaluasi dan mengurangi bias.

Hasil pengujian direpresentasikan dalam bentuk *Confusion Matrix* yang menggambarkan perbandingan antara prediksi model dengan kelas aktual data. Berikut adalah tabel *Confusion Matrix* yang dihasilkan dari proses pengujian:

accuracy: 72.03% +/- 8.97% (micro average: 72.03%)

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	168	47	78.14%
pred. recurrence-events	33	38	53.52%
class recall	83.58%	44.71%	

gambar 2 hasil dataset naive bayes di rapidminer

Berdasarkan gambar di atas, kinerja algoritma Naive Bayes secara keseluruhan menghasilkan metrik evaluasi sebagai berikut:

a. Akurasi (*Accuracy*): 72.03%.

b. Presisi (*Precision*): 78.14%.

c. Recall (*Sensitivity*): 83.58%.

2. Analisis Data

Berdasarkan hasil *Confusion Matrix*, terlihat bahwa model memiliki kemampuan prediksi yang tidak seimbang antara kedua kelas. Model mampu mengenali kelas mayoritas (*no-recurrence-events*) dengan sangat baik, ditunjukkan oleh nilai *Recall* sebesar 83.58%. Sebaliknya, kemampuan model dalam mendeteksi kelas minoritas (*recurrence-events* atau pasien yang kambuh) jauh lebih rendah, dengan *Recall* hanya sebesar 44.71%.

Ketimpangan performa ini mengindikasikan adanya pengaruh kuat dari ketidakseimbangan kelas (*class imbalance*) pada dataset asli. Data latih memiliki proporsi 201 kasus tidak kambuh berbanding 85 kasus kambuh, yang menyebabkan model cenderung bias ke arah kelas mayoritas. Tingginya nilai akurasi (72.03%) namun disertai variasi signifikan pada nilai *recall* menunjukkan bahwa meskipun model secara umum "benar", sensitivitasnya dalam mendeteksi kasus positif (kambuh) masih perlu ditingkatkan.

3. Pembahasan

Hasil penelitian ini menunjukkan bahwa algoritma Naive Bayes cukup efektif untuk klasifikasi awal risiko kanker payudara dengan akurasi mencapai 72.03%. Temuan ini sejalan dengan penelitian sejenis oleh Saritas & Yasar (2019, yang menyatakan bahwa Naive Bayes memiliki performa yang stabil pada dataset kanker payudara, khususnya jika dibandingkan dengan algoritma *Artificial Neural Network* (ANN) yang membutuhkan volume data jauh lebih besar untuk mencapai konvergensi[Ö. B. Güre, 2024].

Kelebihan utama penerapan Naive Bayes dalam penelitian ini terletak pada efisiensi komputasi dan kemampuannya menangani atribut kategorikal. Atribut klinis seperti *menopause*, *breast-quad*, dan *node-caps* dapat diproses langsung tanpa memerlukan konversi numerik yang rumit, yang menjadi keunggulan metode ini pada dataset berdimensi kecil namun kaya fitur nominal.

Namun, dari sisi teoritis, hasil *False Negative* (kesalahan prediksi pasien kambuh yang dianggap tidak kambuh) yang masih terjadi mengindikasikan keterbatasan asumsi dasar Naive Bayes. Algoritma ini mengasumsikan independensi antar fitur (bahwa satu gejala tidak

berhubungan dengan gejala lain). Pada kenyataannya, dalam data biologis kanker, variabel seperti ukuran tumor (*tumor-size*) seringkali memiliki korelasi dengan jumlah kelenjar getah bening yang terlibat (*inv-nodes*)[M. Zwitter et al, 2015]. Pelanggaran asumsi independensi inilah yang kemungkinan berkontribusi pada kesalahan klasifikasi kelas minoritas.

Sebagai implikasi lanjut, meskipun Naive Bayes terbukti komputasional efisien dan akurat secara umum, penggunaan teknik penyeimbang data (*resampling*) seperti SMOTE sangat disarankan untuk penelitian mendatang guna memperbaiki deteksi pada kasus risiko tinggi (*recurrence*).

KESIMPULAN

Berdasarkan implementasi dan analisis yang telah dilakukan, penelitian ini menyimpulkan bahwa algoritma *Naive Bayes Classifier* efektif digunakan sebagai model prediktif untuk menentukan risiko kekambuhan (*recurrence*) pada pasien kanker payudara. Keberhasilan model mencapai akurasi sebesar 72,03% menunjukkan bahwa pendekatan probabilitas yang didasarkan pada asumsi independensi fitur mampu memetakan pola pada dataset medis yang memiliki atribut dominan kategorikal, meskipun dengan jumlah data yang terbatas.

Penerapan metode evaluasi *10-Fold Cross Validation* terbukti memberikan estimasi kinerja yang valid dan objektif, sekaligus meminimalisir bias atau *overfitting* yang sering terjadi pada evaluasi model dengan data kecil. Selain itu, pemanfaatan perangkat lunak RapidMiner dikonfirmasi sangat membantu dalam menyederhanakan proses *Knowledge Discovery in Database* (KDD), terutama dalam tahapan *preprocessing* seperti penanganan *missing values* dan manajemen atribut yang tidak memiliki label standar.

Sebagai saran untuk penelitian lanjutan, fokus utama perlu diarahkan pada penanganan masalah ketidakseimbangan kelas (*class imbalance*) yang teridentifikasi mempengaruhi sensitivitas model. Peneliti menyarankan penerapan teknik *resampling*, khususnya *Synthetic Minority Over-sampling Technique* (SMOTE), guna menyeimbangkan proporsi data antara kasus kambuh dan tidak kambuh. Penerapan teknik ini diharapkan dapat meningkatkan kemampuan model dalam mendeteksi kelas minoritas (*recurrence-events*) secara lebih akurat tanpa mengorbankan presisi keseluruhan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Universitas Pancasakti Tegal atas dukungan fasilitas dan lingkungan akademik yang kondusif sehingga penelitian ini dapat terselesaikan dengan baik. Penghargaan juga disampaikan kepada para peneliti di Institute of Oncology, Ljubljana serta pengelola UCI Machine Learning Repository yang telah menyediakan dataset *Breast Cancer* secara publik, yang menjadi sumber data utama dalam eksperimen penelitian ini.

DAFTAR PUSTAKA

Bhandari, A., Gupta, A., & Das, D. (2015). Improvised apriori algorithm using frequent pattern tree for real time applications in data mining. *Procedia - Procedia Computer Science*, *46*(Icict 2014), 644–651. https://doi.org/10.1016/j.procs.2015.02.115

Bray, F., Ferlay, J., & Soerjomataram, I. (2018). *Global Cancer Statistics 2018 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. 394–424. https://doi.org/10.3322/caac.21492

Caballé-cervigón, N., Castillo-sequera, J. L., & Gómez-pulido, J. A. (n.d.). *applied sciences Machine Learning Applied to Diagnosis of Human Diseases : A Systematic Review*. 1–27. https://doi.org/10.3390/app10155135

Degree, M. M., Science, C., & Lecture, A. C. (2012). *Data Mining : Concepts and*.

Güre, Ö. B. (2024). *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*. 153–160. https://doi.org/10.17798/bitlisfen.1361016

Kharya, S. (2012). *U SING D ATA M ININ G T ECHNIQUES FOR DIAGNOSIS*. *2*(2), 55–66.

Kumar, R., & Goswami, B. K. (2024). *Naive Bayes in Focus : A Thorough Examination of its Algorithmic Foundations and Use Cases*. *9*(5).

Ruvald, R., Frank, M., Johansson, C., Larsson, T., Ruvald, R., Frank, M., & Johansson, C. (2018). Data Mining through Early Experience Prototyping through A step towards through through through through System A step step towards Design A step step towards Design step towards step towards ScienceDirect. *IFAC-PapersOnLine*, *51*(11), 1095–1100. https://doi.org/10.1016/j.ifacol.2018.08.458

Tsamardinos, I., Greasidou, E., & Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning*, *107*(12), 1895–1922. https://doi.org/10.1007/s10994-018-5714-4

Zwitter, M., & Soklic, M. (1988). *Breast Cancer*. UCI Machine Learning Repository. https://doi.org/10.24432/C51P4M