Jurnal Sains Teknologi dan Sistem Informasi
SATESI
journal.yp3a.org/index.php/satesi
Vol. 4 No. 2 (2024) 156 – 163 ISSN Media Elektronik: 2807–8152

Klasifikasi Kanker Payudara Menggunakan Algoritma *K-Nearest Neighbor* dan Metode *Naive Bayes*

Tresi Aprilia
Teknik Informatika, Fakultas Komputer dan Desain, Universitas Selamat Sri, Kendal, Indonesia
Email: tresiaprilia98@gmail.com

ABSTRACT
Breast cancer is one of the diseases that causes death and is one of the most frightening leading causes worldwide. This disease falls under the category of highly dangerous cancers, ranking second after lung cancer. Breast cancer cases occur in large numbers across various regions of the world, raising significant concerns globally. Breast cancer not only affects the quality of life of patients but also contributes significantly to the global cancer mortality rate. It ranks as the fifth leading cause of cancer-related deaths, accounting for approximately 16.6% of the total cancer deaths worldwide. In this study, a classification of blood sample data from breast cancer patients was conducted. Various classification techniques and methods were applied, including the K-Nearest Neighbor (KNN) method and Naïve Bayes (NB). To achieve accurate results, this study tested accuracy using Cross Validation techniques and a Confusion Matrix to evaluate the test data. Of the total 569 data points collected, 70% were used as training data, amounting to 398 data points, while the remaining 30%, or 171 data points, were used as test data. The results of this study showed that the Naïve Bayes method produced an accuracy rate of 96%, with a precision of 94% and a recall of 91%. On the other hand, the K-Nearest Neighbor method yielded a lower accuracy rate of 73%, with a precision of 74% and a recall of 66%, using K=7.

Keywords: Data Mining, Classification, Naïve Bayes, K-NN.

ABSTRAK
Penyakit kanker payudara merupakan salah satu penyakit yang menyebabkan kematian dan menjadi salah satu penyebab utama yang menakutkan di seluruh dunia. Penyakit ini tergolong sebagai salah satu kategori penyakit kanker yang sangat berbahaya dan menduduki peringkat kedua setelah kanker paru-paru. Kasus kanker payudara sangat banyak terjadi di berbagai belahan dunia, sehingga menimbulkan kekhawatiran yang mendalam bagi masyarakat global. Kanker payudara tidak hanya berdampak pada kualitas hidup pasien, tetapi juga berkontribusi besar terhadap angka kematian akibat kanker di dunia. Kanker payudara menjadi penyebab kematian kelima dari semua jenis kanker yang ada, dengan angka yang cukup signifikan, yaitu menyumbang sekitar 16,6% dari total kematian akibat kanker di seluruh dunia. Dalam penelitian ini, dilakukan klasifikasi terhadap data sampel darah pasien yang menderita kanker payudara. Berbagai teknik dan metode klasifikasi diterapkan, termasuk di dalamnya metode *K-Nearest Neighbor* (KNN) dan *Naïve Bayes* (NB). Untuk mendapatkan hasil yang akurat, penelitian ini melakukan pengujian akurasi dengan menggunakan teknik Cross Validation serta Confusion Matrix untuk mengevaluasi data uji. Dari total 569 data yang dikumpulkan, 70% digunakan sebagai data latih, yaitu sebanyak 398 data, sedangkan 30% sisanya digunakan sebagai data uji, yaitu sebanyak 171 data. Hasil penelitian menunjukkan bahwa metode *Naïve Bayes* memberikan nilai akurasi sebesar 96%, presisi 94%, dan recall 91%. Sebaliknya, metode *K-Nearest Neighbor* memberikan hasil akurasi yang lebih rendah, yaitu 73%, dengan presisi 74% dan recall 66% pada nilai K=7.

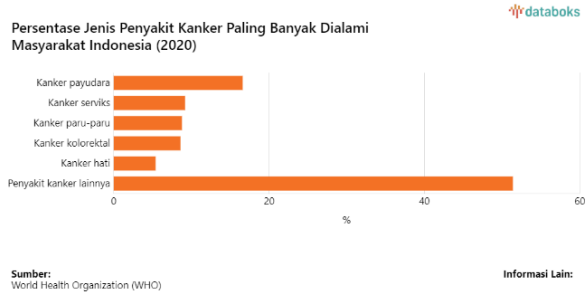
Kata Kunci: Data Mining, Klasifikasi, *Naïve Bayes*, K-NN.

1. Pendahuluan

Faktor utama yang menyebabkan kematian wanita adalah kanker payudara, yang berada di tingkat kedua tertinggi setalah kanker paru-paru. Kanker payudara adalah paling berbahaya pada wanita, penyeban penyakit ini disebabkan oleh beberapa faktor, meliputi faktor sel-sel darah dan saluran jaringan pendukung payudara. Kanker payudara bersifat jinak atau ganas. Pencegahan dan pengobatan kanker payudara yang sesuai dengan jenisnya akan dilakukan segera untuk

Diterima Redaksi: 21-09-2024 | Selesai Revisi: 15-10-2024 | Diterbitkan Online: 20-10-2024

mencegah efek samping dan mengatasi masalah kematian akibat kanker payudara. Dalam kasus penangananya untuk mendeteksi kanker payudara harus mulai disadari sejak awal. Laporan Global Burden of Cancer Study (Globocan) WHO menyatakan bahwa pada tahun 2020, terdapat 396.914 kasus [1].



Gambar 1. Prosentase Jenis Kanker Paling Banyak Dialami Masyarakat Indonesia

Beberapa jenis kanker yang paling umum di Indonesia adalah kanker payudara, sejumlah 65.858 kasus, atau 16,6% dari total kasus kanker di seluruh negeri. Dengan 36.633 kasus yang terajdi di Indonesia, atau 9,2% dari seluruh kanker di seluruh negeri diantaranya kanker serviks menempati peringkat kedua. Kanker paru-paru mendominasi dengan 30.843 kasus (13,2%), diikuti oleh kanker payudara dan kanker serviks, masing-masing dengan 22.430 kasus (9,6%) dan 21.003 kasus (9,6%). Kanker kolorektal menyusul dengan 34.189 kasus (8,6%) dan kanker hati dengan 21.392 kasus (5,4%). Secara keseluruhan, terdapat 204.059 kasus lainnya, yang mencakup 51,4% dari total kasus kanker nasional. Jumlah penderita pria mencapai 137.717.861, sedangkan wanita 135.805.760. Angka kematian akibat kanker di Indonesia mencapai 234.511 pada tahun 2020, dan tingginya kasus kanker serta kematian di negara ini perlu diwaspadai. Machine learning, sebagai subbidang kecerdasan buatan, berfokus pada penggunaan algoritma prediksi, pengenalan pola, dan metode klasifikasi, termasuk neural network, decision tree, naive bayes, dan k-nearest neighbor [2].

Dalam penelitian sebelumnya, algortiam *Naïve Bayes* digunakan untuk mengidentifikasi berbagai jenis kanker payudara. Tahapan yang dilakukan dalam proses kalsifikasi terdiri dari beberapa tahapan, yaitu menganalisa missing value, pemilihan atribut yang akan digunakan, menormalisasi data dan melakukan evaluasi pengujian model dengan menggunakan confusion matrix. Hasil penelitian ini memiliki nilai akurasi 69,12%. Dalam penelitian selanjutnya dengan melakukan klasifikasi data kanker payudara menggunakan algortima gain ratio. Proses yang dilakukan dalam tahapan klasifikasi diantara menentukan nilai internal node cabang satu sampai tiga pada simpul akar. Hasil yang didapatkan dalam menentukan klasifikasi kanker payudara diambil dari performa algoritma yang diukur berdasarkan nilai recall, precision, dan accuracy. Nilai akurasi yang didapatkan sebesar 92,5% [3]. Penelitian yang dilakukan untuk mendapatkan nilai klasifikasi meliputi tahapan split dataset, dan mengembangkan fungsi sigmoid untuk mengubah nilai continue ke dalam range. Dalam penelitian ini berhasil mereduksi fitur dari 30 fitur menjadi 5 fitur dengan tidak mengurangi akurasi dari model klasifikasi dengan menggunakan algortima BPSO dan KNN untuk mendiagnosis penyakit kanker payudara dan menghasilkan tingkat akurasi 95,32% [4].

Penelitian selanjutnya oleh Ilham Mubarog, dkk yang dilakukan pada tahun 2019 berfokus pada pengembangan model diagnosis kanker payudara dengan menggunakan dataset kanker payudara Coimbra. Dalam penelitian ini, metode *Naïve Bayes* digunakan untuk mengklasifikasikan data, dengan tujuan untuk memprediksi apakah seorang pasien menderita kanker payudara atau tidak. Hasil penelitian tersebut menunjukkan bahwa metode *Naïve Bayes* mampu mencapai akurasi terbaik sebesar 80%, dengan nilai presisi dan recall yang sama, yaitu 83%. Ini berarti bahwa model *Naïve Bayes* tidak hanya memberikan hasil prediksi yang akurat, tetapi juga memiliki kemampuan yang baik dalam mendeteksi kasus positif kanker payudara, serta mengurangi jumlah kesalahan dalam klasifikasi [5]. Penelitian lainnya yang dilakukan oleh Nur Ghaniaviyanto Ramadhan dan Faisal Dharma Adhinata menyoroti tantangan yang dihadapi dalam klasifikasi kanker payudara, terutama dalam menangani ketidakseimbangan data (imbalanced data), yaitu ketika jumlah sampel dari satu kelas jauh lebih banyak dibandingkan kelas lainnya. Mereka membandingkan dua model klasifikasi, yaitu *Naïve Bayes* dan Random Forest, untuk menangani masalah tersebut. Hasil dari penelitian mereka menunjukkan bahwa model yang digunakan mampu menghasilkan tingkat presisi yang sangat tinggi, mencapai 99%. Ini menunjukkan bahwa Random Forest dan *Naïve Bayes* dapat digunakan secara efektif untuk mengatasi tantangan ketidakseimbangan data dalam klasifikasi kanker payudara [6]. Kedua penelitian ini memberikan wawasan penting mengenai efektivitas metode *Naïve Bayes* dan Random Forest dalam mengklasifikasikan data kanker payudara, baik pada dataset yang seimbang maupun yang tidak seimbang.

Berdasarkan hasil dan pendekatan dari penelitian yang ada, peneliti dalam studi ini mengusulkan perbandingan model klasifikasi *Naïve Bayes* dan *K-Nearest Neighbor* (K-NN) pada dataset kanker payudara. Tujuan dari perbandingan ini adalah untuk melihat efektivitas kedua metode dalam mengklasifikasikan pasien dengan penyakit kanker payudara dan orang sehat, serta untuk mengidentifikasi metode mana yang lebih unggul dalam memberikan hasil yang akurat [7]. Lebih lanjut, untuk dataset penyakit kanker payudara Coimbra, sejauh ini belum banyak penelitian yang secara eksplisit membandingkan metode *Naïve Bayes* dan K-

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

NN. Ini membuka peluang baru untuk melihat bagaimana kedua metode ini dapat digunakan secara bersamaan untuk mengeksplorasi performa model pada dataset ini. Secara umum, metode klasifikasi digunakan untuk mengelompokkan objek ke dalam kategori yang telah ditentukan sebelumnya. Dalam konteks penelitian ini, tujuan utama dari penggunaan metode klasifikasi adalah untuk memprediksi kelas dari objek (pasien) yang belum memiliki label, yaitu untuk menentukan apakah pasien termasuk dalam kelompok yang menderita kanker payudara atau kelompok yang sehat [8]. Metode klasifikasi *Naïve Bayes* memiliki beberapa keunggulan yang menjadikannya pilihan dalam berbagai kasus, termasuk di bidang kesehatan seperti diagnosis kanker payudara. Metode ini fleksibel karena bisa digunakan untuk data kualitatif maupun kuantitatif, serta tidak memerlukan banyak data pelatihan. Selain itu, *Naïve Bayes* relatif mudah dipahami, efisien dalam perhitungan, dan cocok untuk klasifikasi biner maupun multikelas. Namun, kelemahannya, metode ini sangat bergantung pada asumsi independensi antar variabel, yang tidak selalu realistis. Selain itu, jika terdapat nilai probabilitas 0, maka prediksinya juga akan menjadi 0, yang dapat mempengaruhi hasil klasifikasi [9].

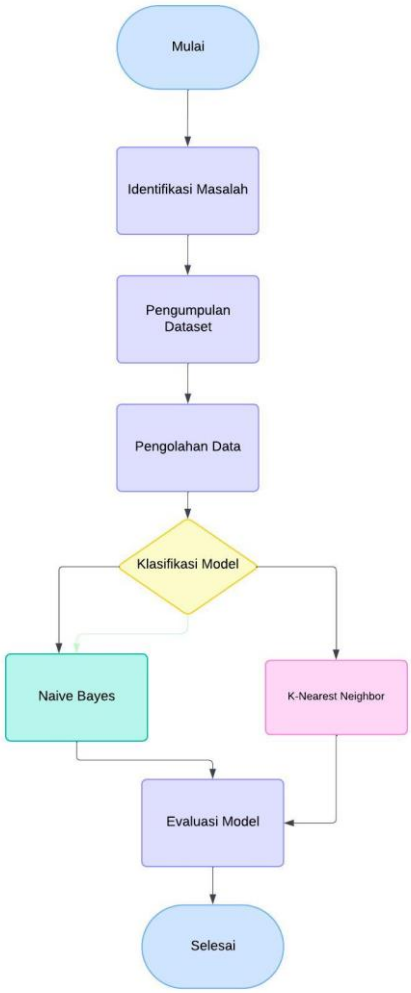
Di Indonesia, metode ini bisa menjadi pilihan dalam proyek-proyek teknologi kesehatan yang semakin berkembang, terutama seiring dengan meningkatnya perhatian terhadap solusi berbasis AI dan data dalam diagnosa penyakit di rumah sakit. Hal ini terlihat dari adanya upaya startup lokal yang mengembangkan solusi diagnosis berbasis machine learning untuk mendeteksi kanker di fase awal. Selain *Naïve Bayes*, metode klasifikasi K-Nearest Neighbour (KNN) juga memiliki keunggulan dalam menangani data dengan banyak noise dan efektif pada dataset besar. Namun, kelemahan KNN terletak pada penentuan parameter K yang optimal dan penentuan jarak yang kurang jelas, yang bisa mempengaruhi performanya pada kasus tertentu. Penggunaan kedua metode ini dalam analisis data kesehatan dapat mendukung perkembangan teknologi di sektor medis Indonesia, terutama dalam deteksi dini kanker payudara yang penting mengingat tingginya angka kejadian kanker payudara [10].

Berdasarkan penelitian yang sudah dilakukan dalam penelitian terdahulu, dapat diambil kesimpulan bahwa dalam menentukan jenis kanker payudara dapat menggunakan teknik metode sudah baik, akan tetapi dalam penelitian sebelumnya memiliki tingkat akurasi yang bisa masih bisa untuk di tingkatkan dengan menggunakan metode yang lain. Oleh sebab itu penelitian ini menggunakan perbandingan algoritma yaitu Naive Bayes dan K-NN dengan menggunakan data uji 30% dari data latihsehingga menghasilkan tingkat akurasi yang optimal. Berdasarkan permasalah yang ada dalam penelitian ini dalam klasifikasi jenis kanker payudara dengan menggunakan Naive Bayes dan K-Nearest Neighbor. Hasil dari penelitian ini menentukan tingkat akurasi yang tinggi sehingga membantu dalam menentukan jenis kanker payudara.

2. Metode Penelitian

2.1 Metode Penelitian

Tahapan pada penelitian ini dimulai dengan mengumpulkan sampel dataset yang didapatkan dari Breast Cancer Wisconsin (Diagnostic). Berikut merupakan alur penelitian yang terdapat pada gambar 2.



Gambar 2 Alur Penelitian

2.1. Dataset

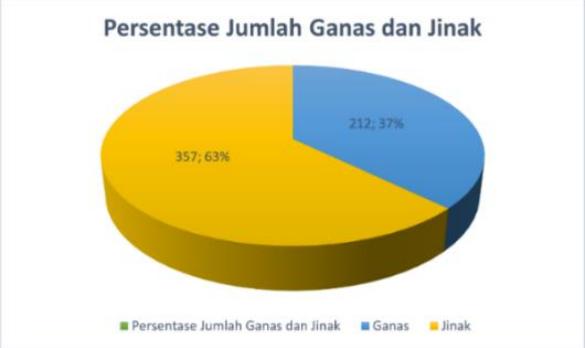
Penelitian ini menggunakan dataset Breast Cancer Wisconsin (Diagnostic) yang diambil dari UCI Machine Learning Repository, yang merupakan salah satu sumber data yang sering digunakan dalam penelitian machine learning. Dataset ini terdiri dari 31 atribut pendukung yang memberikan berbagai informasi terkait kondisi kesehatan pasien, serta 1 label utama, yaitu diagnosis yang menunjukkan status kanker. Label diagnosis ini memiliki dua kategori atau kelas, yaitu jinak dan ganas, yang digunakan untuk membedakan antara jenis kanker yang tidak berbahaya dan yang bersifat agresif. Total record yang

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

tersedia dalam dataset ini berjumlah 569, yang merupakan jumlah data pasien yang digunakan sebagai dasar klasifikasi dalam penelitian ini [11]. Penelitian ini membagi data sesuai pola standar, dengan 70% digunakan sebagai data latih untuk melatih model dan 30% sebagai data uji untuk mengevaluasi performa model. Metode klasifikasi yang digunakan adalah *Naïve Bayes* dan *K-Nearest Neighbor* (KNN), di mana nilai K ditetapkan pada 7, sehingga algoritma mempertimbangkan 7 tetangga terdekat dalam klasifikasi data uji. Hasil yang diharapkan dalam penelitian ini Dataset kanker payudara dapat dilihat pada Tabel 1.

Tabel 1. Dataset Sebelum Encoding Label

No.	ID	texture1	……..	Diagnosis
1	842302	10.38	……..	Ganas
2	842517	17.77	……..	Ganas
3	84300903	21.25	……..	Ganas
4	84348301	20.38	……..	Ganas
5	84358402	14.34	……..	Ganas
6	843786	15.7	……..	Ganas
7	844359	19.98	……..	Jinak
8	84458202	20.83	……..	Ganas
9	844981	21.82	……..	Jinak
10	84501001	24.04	……..	Ganas
11	845636	23.24	……..	Jinak
……..	……..	……..	……..	……..
569	92751	24.54	……..	Jinak



Gambar 3. Persentase Jumlah Kanker Ganas dan Jinak

Gambar 3 merupakan grafik perbandingan untuk jenis label dalam proses klasifikasi. Berdasarkan grafik visualisasi yang disajikan, dapat dilihat persentase dari dua kelas, yaitu jinak dan ganas cukup jauh. Dimana jinak sebanyak jinak sebanyak 63% dan ganas sebanyak 37%. Jadi dapat disimpulkan bahwa dataset yang digunakan memiliki kelas yang tidak seimbang.

2.2. Preprocessing

Dalam proses preprocessing transformasi data yang bersifat kategorik, transformasi data kategorik dapat dilakukan dengan cara Label Encoder. Dalam penelitian ini nilai 1 adalah ganas dan nilai 0 adalah jinak [12]. Dataset kanker payudara setelah dilakukan Label Encoder dapat dilihat pada Tabel 2.

Tabel 2. Dataset setelah Encoding Label

No.	ID	texture1	……..	Diagnosis
1	842302	10.38	……..	1
2	842517	17.77	……..	1
3	84300903	21.25	……..	1
4	84348301	20.38	……..	1
5	84358402	14.34	……..	1
6	843786	15.7	……..	1
7	844359	19.98	……..	0
8	84458202	20.83	……..	1
9	844981	21.82	……..	0
10	84501001	24.04	……..	1
11	845636	23.24	……..	0
……..	……..	……..	……..	……..
569	92751	24.54	……..	0

Selain mentransformasi data kategorik, hal yang juga dapat dilakukan dalam transformasi pada data numerik menggunakan Standard Scaler atau Min Max Scaler. Selain itu hal yang dapat dilakukan adalah menangani data yang tidak seimbang. Pada dataset yang ada tidak terdapat kekosongan data dan tipe data pada setiap features sudah sesuai.

2.3. Klasifikasi

Naïve Bayes adalah metode statistik untuk pengenalan pola yang mendasarkan keputusan klasifikasi pada perhitungan trade-off antara alternatif klasifikasi, menggunakan probabilitas dan biaya terkait. Selain melakukan klasifikasi, metode ini juga dapat memprediksi probabilitas keanggotaan suatu objek dalam sebuah kelas berdasarkan teorema Bayesian. Teorema ini menyatakan bahwa *Naïve Bayes* dapat memberikan hasil yang sangat akurat dan efisien, terutama ketika diterapkan pada dataset berukuran besar, di mana akurasi dan kecepatan pemrosesan menjadi faktor penting dalam analisis data [13]. Selain itu, metode Nearest Neighbor (K-NN) adalah pendekatan lain yang berfokus pada pencarian kasus serupa dengan membandingkan kedekatan antara kasus baru dengan kasus yang ada dalam basis data, berdasarkan bobot dari fitur-fitur tertentu yang digunakan untuk pencocokan. Dalam metode ini, kedekatan atau kesamaan antara kasus dihitung melalui jarak dalam ruang fitur, dan keputusan klasifikasi dibuat berdasarkan kasus-kasus terdekat (tetangga) yang paling mirip. K-NN secara umum sangat efektif untuk situasi di mana pola dapat dikenali melalui hubungan kedekatan antar data. Proses klasifikasi pada kedua metode ini biasanya terdiri dari beberapa tahapan yang terstruktur dengan baik, mulai dari pemrosesan awal data, seleksi fitur, hingga pengelompokan objek atau informasi ke dalam kelas-kelas tertentu secara sistematis. Tahapan-tahapan ini bertujuan untuk mempermudah identifikasi pola dan pengambilan keputusan yang lebih akurat dan efisien, sehingga memungkinkan peneliti untuk mengelompokkan data secara lebih optimal sesuai dengan tujuan analisis yang diinginkan [14].

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2.4. Pembagian Data

Data yang digunakan dalam proses machine learning secara umum terbagi menjadi dua kategori utama, yaitu data latih (training data) dan data uji (testing data). Data latih adalah kumpulan data yang digunakan untuk melatih model klasifikasi sehingga model tersebut dapat mengenali pola atau hubungan antar fitur yang ada dalam data tersebut. Dengan memanfaatkan data latih, model akan "belajar" dan menyesuaikan parameter-parameter yang digunakan untuk membuat prediksi di masa depan. Setelah model dilatih, data uji kemudian digunakan untuk mengukur performa model dan mengevaluasi seberapa baik model tersebut dapat menggeneralisasi terhadap data baru yang belum pernah dilihat sebelumnya. Dalam penelitian atau penerapan machine learning, pembagian data latih dan data uji sering kali dilakukan dengan persentase tertentu. Pada umumnya, sebanyak 70% dari total data digunakan sebagai data latih untuk memastikan bahwa model memiliki cukup informasi untuk belajar dari pola-pola yang ada. Sisanya, sebanyak 30% dari total data, dialokasikan sebagai data uji untuk menguji dan memvalidasi hasil dari model yang telah dilatih. Pembagian ini penting untuk memastikan bahwa model tidak hanya dapat melakukan klasifikasi dengan baik pada data yang sudah dilihat (overfitting), tetapi juga mampu melakukan prediksi yang akurat ketika dihadapkan pada data baru [15]. Proses ini adalah kunci dalam menghasilkan model machine learning yang andal dan dapat diimplementasikan di dunia nyata. Distribusi data latih dan uji dapat ditunjukkan pada Gambar 4.

Gambar 4 Pembagian Dataset

2.5. Metode *Naïve Bayes*

Meetode ini memakai prinsip probabilitas dalam menciptakan model prediksi klasifikas. Metode ini merupakan salah satu metode yang diawasi dan membutuhkan data latih untuk dapat mengambil keputusan. Nilai probabilitas dari setiap kelas yang dituju akan dihitung dengan mempertimbangkan input yang diberikan pada tahap klasifikasi. Kelas target yang memiliki probabilitas paling besarlah yang menjadi kelas pada data inputan tersebut. Keunggulan dari Naive Bayes adalah cepat dan efektif dalam mengolah data dalam jumlah besar. Dibawah ini merupakan formula persamaan Naive Bayes:

$$P(J \mid K) = \frac{P(K \mid J)P(J)}{P(K)} = \frac{P(J \cap K)}{P(K)}$$

Keterangan:

P(J|K) = Probabilitas J terjadi bila K terjadi

P(K|J) = Probabilitas K terjadi bila J terjadi

P(J) = Probabilitas J terjadi

P(K) = Probabilitas K terjadi

P(J∩K) = Probabilitas P(J) dan P(K) terjadi secara bersamaan

Menurut Zhan, meskipun asumsi independensi pada Naive Bayes seringkali tidak realistis dalam aplikasi dunia nyata, Naive Bayes tetap menjadi model yang efektif dan kompetitif, terutama pada dataset besar. Dalam beberapa kasus, bahkan ketika asumsi ini dilanggar, Naive Bayes masih menghasilkan akurasi yang baik. Salah satu kelemahan utama dari metode Naive Bayes adalah masalah ketika ada suatu probabilitas fitur bernilai 0. Hal ini dikenal sebagai zero probability problem, yang bisa menyebabkan hasil prediksi menjadi tidak akurat. Jika dalam data pelatihan tidak ada contoh untuk suatu kombinasi fitur dan kelas tertentu, maka Naive Bayes akan mengasumsikan probabilitas 0 untuk kelas tersebut, yang mengakibatkan keseluruhan probabilitas prediksi juga bernilai 0. Ini tentunya bisa menjadi masalah besar, terutama dalam kasus dengan dataset yang kecil atau tidak merata. Untuk mengatasi masalah ini, biasanya digunakan teknik yang disebut Laplace Smoothing atau Additive Smoothing. Teknik ini menambahkan konstanta kecil ke setiap hitungan frekuensi untuk menghindari nilai probabilitas 0. Dengan demikian, bahkan jika suatu fitur tidak pernah muncul dalam data pelatihan untuk kelas tertentu, algoritma tetap memberikan probabilitas yang kecil tetapi tidak nol [16]. Ini membantu meningkatkan performa dan stabilitas model. Naive Bayes memiliki sejumlah kelebihan yang membuatnya sering dipilih untuk berbagai aplikasi klasifikasi. Beberapa kelebihan utamanya dibandingkan metode klasifikasi lain adalah sebagai berikut: Bisa digunakan untuk data kualitatif maupun kuantitatif: Naive Bayes fleksibel dalam mengolah data yang bersifat kualitatif (kategori) dan kuantitatif (numerik). Ini membuatnya cocok untuk berbagai jenis dataset tanpa memerlukan banyak perubahan dalam pendekatan. Efektif dengan jumlah data yang sedikit: Berbeda dengan beberapa metode klasifikasi yang memerlukan jumlah data pelatihan besar untuk memberikan hasil yang baik, Naive Bayes dapat memberikan hasil yang cukup baik meskipun data latihannya sedikit. Hal ini dikarenakan penggunaan probabilitas bersyarat berdasarkan fitur-fitur individu. Efisien dalam perhitungan: Algoritma Naive Bayes sangat sederhana dalam hal perhitungan, terutama karena tidak ada banyak iterasi atau pemrosesan kompleks seperti dalam metode lain (misalnya, metode berbasis optimasi seperti Support Vector Machine atau Neural Network). Ini membuatnya efisien dari segi waktu dan sumber daya komputasi. Mudah dipahami dan dibua, implementasi Naive Bayes sederhana dan konsep probabilitas yang digunakannya relatif mudah dipahami, sehingga

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

metode ini cocok untuk pemula atau mereka yang baru mempelajari machine learning dan klasifikasi. Dapat disesuaikan untuk klasifikasi dokumen: Naive Bayes sangat efektif dalam pengolahan teks, terutama dalam tugas-tugas seperti klasifikasi dokumen, analisis sentimen, dan pengkategorian email (misalnya, spam filter). Algoritma ini bisa disesuaikan untuk berbagai kebutuhan pengelompokan teks. Mendukung klasifikasi biner maupun multikelas: Naive Bayes bisa digunakan baik untuk klasifikasi biner (dua kelas) maupun multikelas (lebih dari dua kelas). Fleksibilitas ini membuatnya berguna dalam berbagai konteks klasifikasi [17].

2.6. Algoritma K-Nearest Neighbor

Algoritma ini menemukan jumlah k di setiap pola yang terdekat pada pola masukan (di antara semua skema latih yang terdapat di setiap kelas), sehingga dapat ditentukankelas keputusan didasarkan pada jumlah pola yang paling banyak di antara k pola. Salah satu besaran jarak yang telah ditentukan dapat digunakan untuk mengetahui seberapa dekat atau jauh lokasinya (jarak). Namun, jarak geometris sering digunakan karena sangat akurat dan efisien [18].

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{p} (x_{ip} - x_{jp})^2}$$

D (xi, xj) adalah jarak geometris antara data uji dan data latihan, sedangkan xip dan x jp adalah data uji ke-i dan data pelatihan ke-j. Algoritma *K-Nearest Neighbor* (KNN) adalah metode klasifikasi yang didasarkan pada jarak terdekat antara suatu objek dan data pembelajaran. KNN merupakan algoritma supervised learning, di mana objek baru diklasifikasikan berdasarkan mayoritas kategori dari tetangga terdekatnya, dengan kelas yang paling sering muncul digunakan sebagai hasil klasifikasi. Kedekatan diukur menggunakan metrik jarak, seperti jarak Euclidean. Dalam penelitian ini, dua model KNN akan dibuat: satu tanpa seleksi fitur dan satu lagi dengan seleksi fitur menggunakan Binary PSO, untuk mengevaluasi pengaruh seleksi fitur terhadap kinerja algoritma. Nilai k, yang merupakan jumlah tetangga terdekat, diambil dari bilangan ganjil antara 1 hingga 21 untuk menghindari hasil seri dalam kasus data genap. Misalnya, jika k=11, maka 11 tetangga terdekat akan dipertimbangkan. Data uji dan data latih dibagi sesuai proporsi yang diinginkan, lalu jarak antara setiap tetangga dan objek dihitung dan diurutkan dari yang terkecil. Setelah itu, 11 tetangga terdekat diidentifikasi, dan jika mayoritas dari tetangga tersebut tergolong dalam kelas kanker jinak, maka objek tersebut diklasifikasikan sebagai kanker jinak, dan sebaliknya untuk kanker ganas.

2.6. Evaluasi Model

Analisis akan menampilkan hasil eksperimen dan mengevaluasi kinerja pengklasifikasi yang diusulkan untuk dataset kanker payudara. Metode evaluasi kinerja adalah confusion matrix. Confusion matrix berisi informasi detail tentang hasil pengenalan atau klasifikasi (prediksi) sistem terhadap data pengujian yang telah diketahui kelasnya (aktual). Confusion matrixbiasanya disusun dalam bentuk matrik [19]. Evaluasi kinerja didefinisikan sebagai berikut:

a. Akurasi

$$\frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (1)}$$

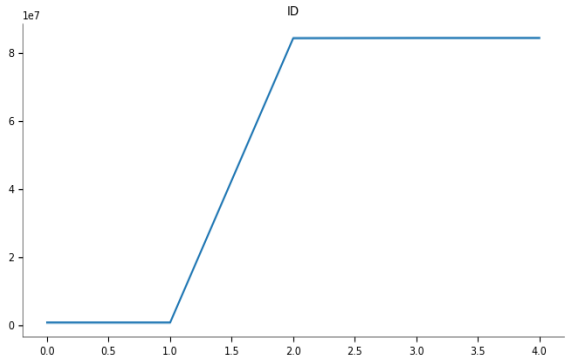
b. Presisi

$$\frac{TN}{TP+TN} \quad \text{..... (2)}$$

c. Recall

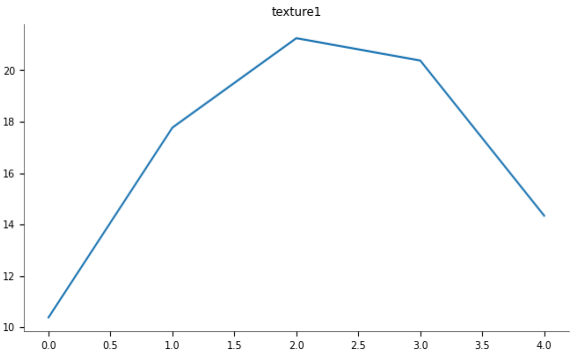
$$\frac{TN}{TN+FN} \quad \text{..... (3)}$$

Berdasarkan dari hasil klasifikasi dengan metode *Naïve Bayes* didapatkan hasil grafik distribution untuk atribut ID yang ditunjukan pada grafik gambar 5.



Gambar 5. Gambar Grafik ID

Berdasarkan dari hasil klasifikasi dengan metode *Naïve Bayes* didapatkan hasil grafik distribution untuk tribut texture1 yang ditunjukan pada grafik gambar 6.

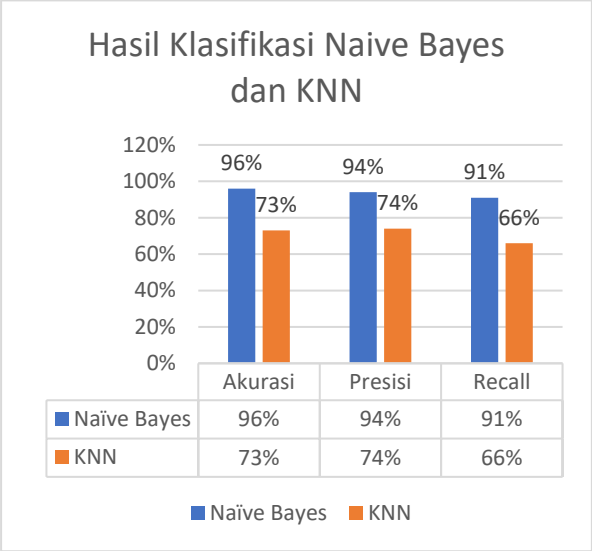


Gambar 6. Gambar Grafik Texture

3. Hasil dan Pembahasan

Pada metode *Naïve Bayes* didapatkan hasil dengan nilai akurasi 96%, presisi 94% dan recall 91%. Pada algoritma KNN mendapatkan hasil dengan nilai akurasi 73%, presisi 74% dan recall 66% dengan nilai K=7.

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Gambar 7. Grafik Hasil Klasifikasi

Dalam penelitian ini, dilakukan beberapa skenario pengujian untuk mengevaluasi performa model klasifikasi. Skenario-skenario pengujian tersebut mencakup beberapa aspek penting, yaitu pengujian berdasarkan data sebelum dan sesudah proses normalisasi, perbandingan performa antara data training dan data testing, variasi nilai k pada algoritma *K-Nearest Neighbor* (K-NN), serta pemilihan atribut terkuat menggunakan uji korelasi Pearson. Tujuan dari berbagai skenario ini adalah untuk memastikan model dapat mempelajari pola dengan optimal dan menghasilkan prediksi yang akurat. Pada skenario pertama, dilakukan pengujian untuk membandingkan performa model sebelum dan sesudah data dinormalisasi. Normalisasi data bertujuan untuk mengubah skala data sehingga setiap fitur memiliki bobot yang lebih seimbang, yang dapat membantu model memahami hubungan antar fitur secara lebih efektif. Hasil dari skenario ini menunjukkan bahwa

metode *Naïve Bayes* menampilkan kinerja yang sangat unggul, dengan akurasi mencapai 96%, yang berarti 96% dari total prediksi yang dibuat oleh model ini adalah benar. Selain itu, metode ini juga menghasilkan presisi sebesar 94%, yang mengindikasikan bahwa dari semua prediksi positif yang dihasilkan, 94% di antaranya benar-benar positif. Nilai recall sebesar 91% menunjukkan bahwa model *Naïve Bayes* mampu mendeteksi 91% dari seluruh data yang benar-benar positif, menandakan bahwa model ini andal dalam mengenali data yang relevan dan penting. Dengan hasil ini, *Naïve Bayes* terbukti sangat efektif dalam konteks penelitian ini, baik dalam hal akurasi, presisi, maupun recall. Di sisi lain, metode *K-Nearest Neighbor* (K-NN) dengan nilai k=7 menunjukkan kinerja yang lebih rendah dibandingkan *Naïve Bayes*. Model K-NN ini hanya mampu mencapai akurasi sebesar 73%, yang berarti 73% dari total prediksi yang dibuat oleh model ini adalah benar. Presisi model K-NN sebesar 74% menunjukkan bahwa dari semua prediksi positif yang dihasilkan, hanya 74% yang benar-benar positif, sementara recall sebesar 66% menunjukkan bahwa model ini hanya mampu mendeteksi 66% dari total data positif yang ada. Meskipun K-NN memberikan hasil yang cukup baik dalam melakukan klasifikasi, kinerjanya masih berada di bawah metode *Naïve Bayes*. Secara keseluruhan, hasil ini menegaskan bahwa *Naïve Bayes* lebih unggul dalam hal keakuratan dan keandalan dalam klasifikasi dataset yang digunakan dalam penelitian ini.

4. Kesimpulan

Kesimpulan yang dapat diambil dari pembahasan sebelumnya adalah bahwa penelitian ini menggunakan total 569 data, di mana 70% dari data tersebut, yaitu sebanyak 398 data, digunakan sebagai data latih (training data), sementara sisanya, yaitu 30% atau sebanyak 171 data, digunakan sebagai data uji (testing data). Pemisahan data ini dilakukan untuk memastikan bahwa model dapat dilatih dan diuji secara terpisah, sehingga hasil yang diperoleh dapat memberikan gambaran akurat mengenai performa model pada data yang belum pernah dilihat sebelumnya. Dari hasil penelitian, dapat dilihat bahwa metode *Naïve Bayes* menunjukkan kinerja yang sangat baik dengan menghasilkan akurasi sebesar 96%. Ini berarti, 96% dari total prediksi yang dibuat oleh model *Naïve Bayes* adalah benar. Selain itu, presisi dari metode ini mencapai 94%, yang menunjukkan bahwa dari seluruh prediksi positif yang dihasilkan model, 94% di antaranya benar-benar positif. Recall sebesar 91% mengindikasikan bahwa dari seluruh data yang benar-benar positif, 91% berhasil dideteksi oleh model ini. Dengan demikian, *Naïve Bayes* terbukti memiliki kinerja yang unggul dalam hal akurasi, presisi, dan recall dalam konteks dataset yang digunakan. Di sisi lain, metode *K-Nearest Neighbor* (K-NN) yang menggunakan nilai k=7 menunjukkan hasil yang lebih rendah dibandingkan *Naïve Bayes*. Model K-NN ini menghasilkan akurasi sebesar 73%, yang berarti 73% dari total prediksi yang dihasilkan adalah benar. Nilai presisi sebesar 74% menunjukkan bahwa dari seluruh prediksi positif yang dibuat oleh K-NN, 74% adalah benar, sementara recall sebesar 66% menunjukkan bahwa hanya 66% dari total data yang benar-benar positif berhasil dikenali oleh model ini. Dengan demikian, meskipun metode K-NN memberikan hasil yang cukup baik, performanya masih berada di bawah metode *Naïve Bayes* dalam konteks penelitian ini.

5. Saran

Pada kumpulan data yang digunakan dalam penelitian ini, tidak ditemukan adanya kesalahan atau inkonsistensi, sehingga data asli dapat dipertahankan dengan baik tanpa modifikasi atau pembersihan lebih lanjut. Hal ini memastikan bahwa data yang digunakan adalah representasi yang valid dari sumber aslinya. Meskipun demikian, masih terdapat potensi untuk melakukan pemrosesan ulang data tersebut

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

dengan tujuan mendapatkan hasil yang lebih optimal. Pemrosesan ulang ini dapat mencakup berbagai metode, seperti normalisasi tambahan, seleksi fitur, atau penerapan teknik-teknik pengolahan data lainnya yang dapat meningkatkan performa model. Selain itu, sejauh ini belum dilakukan perbandingan tingkat akurasi model yang dihasilkan dengan algoritma lain menggunakan dataset yang serupa. Langkah ini bisa memberikan perspektif lebih luas mengenai efektivitas algoritma yang digunakan dalam konteks yang berbeda. Untuk lebih meningkatkan akurasi model yang ada, salah satu strategi yang dapat diterapkan adalah melakukan penyesuaian parameter sampling secara linier, yang disesuaikan dengan karakteristik spesifik dari dataset yang digunakan. Penyesuaian ini bertujuan untuk mencocokkan model dengan pola data secara lebih akurat, sehingga meningkatkan performa keseluruhan model dalam melakukan prediksi atau klasifikasi.

SUMBER RUJUKAN
Referensi

[1] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.

[2] J. T. Wijaya, H. Oktavianto, H. Azizah, and A. Faruq, "Perbandingan Algoritma *K-Nearest Neighbor* (Knn) Dan Gaussian Naive Bayes (Gnb) Dalam Klasifikasi Breast Cancer Coimbra Comparison Between *K-Nearest Neighbor* (Knn) And Gaussian Naive Bayes (Gnb) Algorithm In The Coimbra Breast Cancer Classification," *J. Smart Teknol.*, vol. 3, no. 3, pp. 2774–1702, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[3] H. Harafani and H. A. Al-Kautsar, "Meningkatkan Kinerja K-Nn Untuk Klasifikasi Kanker Payudara Dengan Forward Selection," *J. Pendidik. Teknol. dan Kejuru.*, vol. 18, no. 1, p. 99, 2021, doi: 10.23887/jptk-undiksha.v18i1.29905.

[4] L. W. Astuti, I. Saluza, F. Faradilla, and M. F. Alie, "Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes," *J. Ilm. Inform. Glob.*, vol. 11, no. 2, 2021, doi: 10.36982/jiig.v11i2.1235.

[5] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 148–151, 2018.

[6] R. Hidayat, D. Kartini, M. I. Mazdadi, I. Budiman, and R. Ramadhani, "Implementasi Seleksi Fitur Binary Particle Swarm Optimization pada Algoritma K-NN untuk Klasifikasi Kanker Payudara," *J. Sist. dan Teknol. Inf.*, vol. 11, no. 1, p. 135, 2023, doi: 10.26418/justin.v11i1.53608.

[7] H. Oktavianto and R. P. Handri, "Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes," *INFORMAL Informatics J.*, vol. 4, no. 3, p. 117, 2020, doi: 10.19184/isj.v4i3.14170.

[8] B. Aisyah and Y. Sulistyo, "Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio," *J. Tek. Elektro*, vol. 8, no. 2, pp. 43–46, 2016.

[9] Fahrurrozi and Wasilah, "Deteksi Dini Kanker Payudara Menggunakan Algoritma *K-Nearest Neighbor* (KNN) Dan Decision Tree C-45," *Teknika*, vol. 17, no. 2, pp. 427–434, 2023, [Online]. Available: https://jurnal.polsri.ac.id/index.php/teknika/article/view/7565

[10] J. KUSUMA, B. H. HAYADI, W. WANAYUMINI, and R. ROSNELLY, "Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara," *MIND J.*, vol. 7, no. 1, pp. 51–60, 2022, doi: 10.26760/mindjournal.v7i1.51-60.

[11] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.

[12] A. Nugraheni, R. D. Ramadhani, A. B. Arifa, and A. Prasetiadi, "Perbandingan Performa Antara Algoritma Naive Bayes Dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 1, pp. 11–20, 2022, doi: 10.20895/dinda.v2i1.391.

[13] M. Abdul Jabbar, E. Hasmin, C. Susanto, W. Musu, and I. Artikel, "Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification," *Oktober*, vol. 14, no. 3, pp. 258–270, 2022, [Online]. Available: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

[14] T. A. Yoga and Prihandoko, "Penerapan Optimasi Berbasis Particle Swarm Optimization (Pso) Algoritma *Naïve Bayes* Dan *K-Nearest Neighbor* Sebagai Perbandingan Untuk Mencari Kinerja Terbaik Dalam Mendeteksi Kanker Payudara," *J. Bangkit Indones.*, vol. 7, no. 2, p. 1, 2018, [Online]. Available: http://journal.universitasmulia.ac.id/index.php/metik/article/view/62

[15] M. Alwohaibi, M. Alzaqebah, N. M. Alotaibi, A. M. Alzahrani, and M. Zouch, "A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5192–5203, 2022, doi: 10.1016/j.jksuci.2021.05.004.

[16] Y. Feng *et al.*, "Predicting breast cancer-specific survival in metaplastic breast cancer patients using machine learning algorithms," *J. Pathol. Inform.*, vol. 14, no. August, p. 100329, 2023, doi: 10.1016/j.jpi.2023.100329.

[17] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. November 2020, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[18] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. May, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.

[19] V. Jaiswal, P. Saurabh, U. K. Lilhore, M. Pathak, S. Simaiya, and S. Dalal, "A breast cancer risk predication and classification model with ensemble learning and big data fusion," *Decis. Anal. J.*, vol. 8, no. April, p. 100298, 2023, doi: 10.1016/j.dajour.2023.100298.

DOI: 10.54259/satesi.v4i2.3167
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)