Gudang Jurnal Multidisiplin Ilmu E-ISSN : 2988-5760
Volume 2 ; Nomor 12 ; Desember 2024 ; Page 778-785
Doi : https://doi.org/10.59435/gjmi.v2i12.1206
Website : https://gudangjurnal.com/index.php/gjmi

Komparasi Algoritma Naive Bayes Dan Random Forest Pada Klasifikasi Kanker Payudara

Ika Indah Lestari[1*], Ahmad Homaidi[2]

[1] Ilmu Komputer, Universitas Ibrahimy, Indonesia
[2] Teknologi Informasi, Universitas Ibrahimy, Indonesia
[1*] ikaindah445@gmail.com

Abstrak

Kanker payudara merupakan salah satu jenis kanker yang paling umum ditemukan pada wanita dan menjadi penyebab utama kematian akibat kanker di seluruh dunia. Ketepatan dalam diagnosis kanker payudara menjadi sangat krusial untuk penanganan yang tepat. Penelitian ini bertujuan untuk membandingkan performa algoritma Naive Bayes dan Random Forest dalam mengklasifikasikan kanker payudara menggunakan dataset Breast Cancer Wisconsin. Metodologi penelitian dimulai dengan pengumpulan data dari dataset Breast Cancer Wisconsin yang terdiri dari 569 sampel dengan 32 atribut. Proses preprocessing data meliputi konversi data dari format nominal ke binominal untuk atribut diagnosis. Implementasi algoritma menggunakan tools RapidMiner dengan pendekatan cross validation (k=10) untuk evaluasi model yang lebih robust. Performa kedua algoritma dibandingkan menggunakan berbagai metrik evaluasi termasuk accuracy, precision, recall, dan analisis confusion matrix. Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan performa yang lebih unggul dengan tingkat akurasi 94,91% (±5,06%), precision 95,33%, dan recall 93,90%. Sementara itu, Naive Bayes mencapai akurasi 93,51% (±5,30%), precision 93,68%, dan recall 92,67%. Random Forest juga menunjukkan keunggulan dalam mengurangi false positive, dengan hanya 8 kasus dibandingkan 15 kasus pada Naive Bayes. Analisis confusion matrix menunjukkan bahwa kedua algoritma memiliki kemampuan yang baik dalam mengklasifikasikan kasus kanker payudara, meskipun Random Forest menunjukkan performa yang lebih stabil dan akurat. Kesimpulan dari penelitian ini menunjukkan bahwa kedua algoritma efektif untuk klasifikasi kanker payudara, dengan Random Forest menunjukkan keunggulan dalam hal akurasi dan presisi. Hasil ini dapat menjadi pertimbangan dalam pengembangan sistem pendukung keputusan untuk diagnosis kanker payudara, dimana Random Forest dapat menjadi pilihan utama ketika akurasi menjadi prioritas, sementara Naive Bayes tetap menjadi alternatif yang valid ketika kesederhanaan implementasi dan efisiensi komputasi diperlukan.

Kata Kunci: Kanker Payudara, Naive Bayes, Random Forest, Klasifikasi, Data Mining

PENDAHULUAN

Kanker payudara merupakan jenis kanker yang paling umum ditemukan pada wanita dan menjadi penyebab utama kematian akibat kanker di seluruh dunia. Data dari World Health Organization (WHO) menunjukkan bahwa pada tahun 2015, lebih dari 2,1 juta wanita terdiagnosis kanker payudara, dengan kontribusi sebesar 25% dari seluruh kasus kanker(*Tri Cita Pelima*, 2021). Penyakit ini dimulai ketika sel-sel di payudara tumbuh secara tidak terkendali dan membentuk tumor yang dapat terdeteksi melalui pemeriksaan sinar-X atau terasa sebagai benjolan. Angka ini terus menunjukkan peningkatan yang signifikan setiap tahunnya, menjadikan kanker payudara sebagai masalah kesehatan global yang membutuhkan perhatian serius(Nugraheni et al., 2022). Dalam dunia medis, diagnosis kanker payudara menjadi tantangan tersendiri bagi para tenaga kesehatan, terutama dalam mengklasifikasikan tumor sebagai ganas (malignant) atau jinak (benign). Ketepatan diagnosis ini sangat krusial karena berkaitan langsung dengan penentuan tindakan pengobatan dan tingkat keberhasilan terapi(Shidqi et al., 2022). Kesalahan dalam klasifikasi dapat berakibat fatal, baik berupa keterlambatan penanganan untuk kasus tumor ganas maupun pengobatan yang tidak perlu untuk tumor jinak. Kompleksitas dalam proses diagnosis ini semakin meningkat karena banyaknya variabel yang harus dipertimbangkan dalam menentukan sifat tumor(Muntiari & Hanif, 2022).

Seiring dengan perkembangan teknologi informasi, pendekatan berbasis data mining dan machine learning telah membuka peluang baru dalam meningkatkan akurasi diagnosis kanker payudara. Teknologi ini memungkinkan analisis yang lebih cepat dan akurat terhadap berbagai parameter tumor, seperti ukuran, bentuk, dan karakteristik sel-sel kanker yang diperoleh dari hasil pemeriksaan biopsi(A'yunan et al., 2023). Data mining telah terbukti efektif dalam mengidentifikasi pola-pola tersembunyi dalam data medis yang kompleks, yang seringkali sulit dideteksi melalui metode konvensional. Di antara berbagai algoritma machine learning yang ada, Naive Bayes dan Random Forest muncul sebagai dua pendekatan yang menjanjikan dalam klasifikasi kanker payudara. Algoritma Naive Bayes, yang didasarkan pada teorema probabilitas Bayes, dikenal dengan kemampuannya dalam menangani data kategoris dan numerik, serta efisiensinya dalam pemrosesan dataset besar(Ismai, 2017). Sementara itu, Random Forest, yang

merupakan pengembangan dari metode decision tree, unggul dalam menangani data yang kompleks dan mampu mengatasi masalah overfitting yang sering terjadi pada algoritma klasifikasi lainnya.

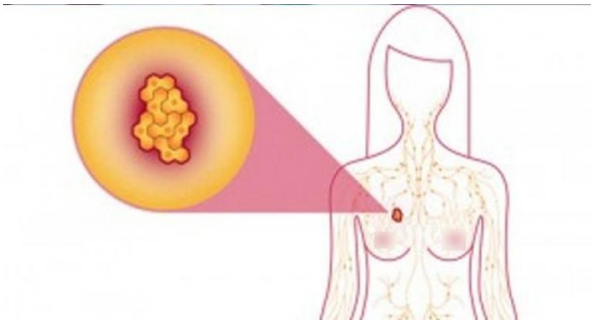
Beberapa penelitian terdahulu telah menunjukkan hasil yang beragam dalam penggunaan kedua algoritma ini. Penelitian yang dilakukan oleh Vincent Angkasa dan Jefri Junifer Pangaribuan menunjukkan bahwa Random Forest dapat mencapai akurasi hingga 99,51% dalam klasifikasi kanker payudara. Di sisi lain, penelitian oleh Novita Ranti Muntiari dan Kharis Hudaiby Hanif mengungkapkan bahwa algoritma Naive Bayes juga mampu memberikan hasil yang kompetitif dengan tingkat akurasi mencapai 95%(Angkasa & Pangaribuan, 2022). Namun, masih terdapat kesenjangan dalam pemahaman mengenai performa relatif kedua algoritma ini ketika diaplikasikan pada dataset yang sama dengan karakteristik yang spesifik. Dataset Wisconsin Breast Cancer, yang digunakan dalam penelitian ini, menyediakan data yang komprehensif tentang berbagai karakteristik sel tumor payudara. Dataset ini mencakup 30 fitur yang diukur dari gambar digital hasil *Fine Needle Aspirate* (FNA), termasuk radius, tekstur, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, dan fractal dimension. Setiap fitur ini memiliki nilai mean, standard error, dan "worst" atau rata-rata dari tiga nilai terbesar, yang memberikan gambaran detail tentang variasi karakteristik sel tumor.

Penggunaan RapidMiner sebagai tools analisis dalam penelitian ini memungkinkan implementasi dan evaluasi yang lebih terstruktur dari kedua algoritma. RapidMiner menyediakan lingkungan yang terintegrasi untuk seluruh proses data mining, mulai dari preprocessing data hingga evaluasi model, dengan berbagai metrik performa yang dapat digunakan untuk membandingkan efektivitas algoritma. Platform ini juga memungkinkan visualisasi hasil yang lebih komprehensif, memudahkan interpretasi dan analisis perbandingan kedua algoritma(Faid et al., 2019).

METODE

Kanker Payudara

Kanker payudara adalah pertumbuhan sel abnormal yang terjadi pada jaringan payudara, dimana sel-sel ini dapat berkembang secara tidak terkendali dan berpotensi menyebar ke bagian tubuh lainnya. Kanker ini umumnya bermula dari sel-sel pada saluran yang membawa air susu ke puting (duktus) atau dari sel-sel yang membentuk kelenjar penghasil air susu (lobulus). Kondisi ini dapat terjadi pada pria maupun wanita, namun kasus pada wanita jauh lebih umum ditemukan. Perkembangan kanker payudara umumnya terjadi melalui beberapa tahapan. Pada tahap awal, sel-sel abnormal mulai berkembang di dalam saluran susu atau kelenjar susu(Risiko & Payudara, 2013). Ketika sel-sel ini mulai menyebar ke jaringan sekitarnya, kondisi ini dikenal sebagai karsinoma invasif. Stadium kanker payudara ditentukan berdasarkan seberapa jauh sel kanker telah menyebar, dimulai dari stadium 0 (tahap paling awal) hingga stadium IV (tahap paling lanjut dimana kanker telah menyebar ke organ lain).

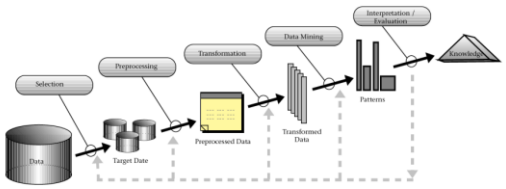


Gambar 1. Ilustrasi Kanker Payudara

Faktor risiko kanker payudara sangat beragam dan mencakup aspek genetik, hormonal, dan lingkungan. Beberapa faktor risiko utama meliputi usia, riwayat keluarga dengan kanker payudara, paparan hormon estrogen yang berkepanjangan, obesitas, dan gaya hidup tidak sehat. Pemahaman tentang faktor-faktor risiko ini penting dalam upaya pencegahan dan deteksi dini kanker payudara. Diagnosis kanker payudara melibatkan beberapa metode pemeriksaan, termasuk pemeriksaan fisik, mamografi, ultrasonografi, dan biopsi. Biopsi, khususnya Fine Needle Aspiration (FNA), merupakan metode yang sangat penting dalam menentukan sifat tumor - apakah jinak (benign) atau ganas (malignant)(Rahmadini et al., 2022). Hasil pemeriksaan FNA menghasilkan data-data penting seperti ukuran sel, bentuk sel, tekstur, dan karakteristik lainnya yang menjadi dasar dalam penentuan diagnosis.

Data Mining

Data mining merupakan proses ekstraksi atau "penggalian" pengetahuan dan wawasan yang bermanfaat dari kumpulan data dalam jumlah besar. Proses ini melibatkan penggunaan metode statistik, matematika, dan machine learning untuk mengidentifikasi pola dan hubungan tersembunyi dalam data yang tidak dapat dideteksi dengan mudah melalui analisis manual biasa. Dalam praktiknya, data mining memiliki beberapa tahapan utama yang saling berkaitan(Sigit & Yuita, 2018). Tahap pertama adalah preprocessing data, yang meliputi pembersihan data dari nilai yang hilang atau tidak valid, transformasi data ke format yang sesuai, dan normalisasi data. Tahap kedua adalah analisis data, dimana berbagai algoritma dan teknik diterapkan untuk menemukan pola dan hubungan dalam data. Tahap terakhir adalah interpretasi hasil, dimana pola dan hubungan yang ditemukan dianalisis untuk menghasilkan wawasan yang bermakna(Alrasyid et al., 2024).



Gambar 2. Tahapan Data Mining

Data mining memiliki berbagai teknik yang dapat digunakan sesuai dengan tujuan analisis. Teknik-teknik ini meliputi klasifikasi (pengelompokan data ke dalam kategori tertentu), clustering (pengelompokan data berdasarkan kemiripan karakteristik), asosiasi (menemukan hubungan antar variabel), dan prediksi (memperkirakan nilai di masa depan berdasarkan pola historis). Dalam konteks medis, khususnya diagnosis kanker payudara, teknik klasifikasi menjadi sangat relevan karena dapat membantu mengkategorikan tumor sebagai jinak atau ganas berdasarkan karakteristik terukur(Jalil et al., 2024). Penerapan data mining dalam bidang kesehatan telah mengalami perkembangan pesat dalam beberapa tahun terakhir. Teknologi ini telah terbukti efektif dalam berbagai aplikasi medis, seperti diagnosis penyakit, prediksi risiko kesehatan, analisis citra medis, dan pengoptimalan protokol pengobatan. Khusus dalam kasus kanker payudara, data mining memungkinkan analisis yang lebih akurat dan objektif terhadap data-data hasil pemeriksaan medis, seperti hasil mamografi dan biopsi, yang dapat membantu dokter dalam membuat keputusan diagnosis yang lebih tepat.

Cross Validation

Cross Validation adalah teknik validasi model yang digunakan untuk menilai bagaimana hasil analisis statistik akan digeneralisasi terhadap dataset independen. Teknik ini terutama digunakan dalam konteks di mana tujuannya adalah prediksi, dan ingin memperkirakan seberapa akurat model prediktif akan bekerja dalam praktik. Metode yang paling umum digunakan adalah k-fold cross validation, di mana data dibagi menjadi k subset yang sama besar. Dalam setiap iterasi, satu subset digunakan sebagai data uji (testing data) sementara k-1 subset lainnya digunakan sebagai data latih (training data). Proses ini diulang sebanyak k kali sehingga setiap subset mendapat kesempatan menjadi data uji(Hadistio et al., 2022).

Naïve Bayes

Naive Bayes adalah algoritma klasifikasi yang didasarkan pada teorema Bayes dengan asumsi independensi yang kuat (naif) antara fitur. Meskipun asumsi ini sering tidak realistis, algoritma ini tetap memberikan hasil yang baik dalam banyak kasus nyata(Munazilin & Nasta'in, 2023).

Teorema Bayes yang menjadi dasar algoritma ini dapat diformulasikan sebagai berikut:

$P(C|X) = P(X|C) \times P(C) / P(X)$

Dimana:

$P(C|X)$ adalah probabilitas posterior kelas C dengan fitur X

$P(X|C)$ adalah likelihood atau probabilitas fitur X pada kelas C

$P(C)$ adalah probabilitas prior kelas C

$P(X)$ adalah probabilitas fitur X

Dalam implementasinya untuk klasifikasi, Naive Bayes menghitung probabilitas setiap kelas untuk fitur yang diberikan dan memilih kelas dengan probabilitas tertinggi. Untuk fitur numerik, biasanya digunakan distribusi Gaussian dengan rumus(Rahayu & Qurrota, 2022):

$P(x|C) = 1/\sqrt{(2\pi\sigma^2)} \times e^{(-(x-\mu)^2/(2\sigma^2))}$

Dimana:

μ adalah mean dari fitur dalam kelas tersebut

σ^2 adalah varians dari fitur dalam kelas tersebut

Random Forest

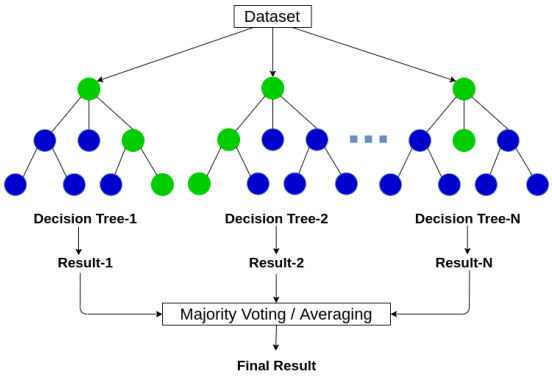
Random Forest adalah algoritma ensemble learning yang terdiri dari kumpulan decision tree. Setiap tree memberikan prediksi, dan hasil akhir ditentukan melalui voting mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi)(Rigatti, 2017).

Algoritma ini bekerja dengan beberapa prinsip utama:

Bootstrap Aggregating (Bagging): Setiap tree dilatih menggunakan sampel acak dengan pengembalian dari dataset asli.

Random Feature Selection: Pada setiap split node, hanya subset acak dari fitur yang dipertimbangkan. Untuk setiap tree dalam Random Forest, probabilitas kelas k dapat dihitung dengan rumus:

$P(k|x) = 1/N \times \Sigma(i=1 \text{ to } N) I(hi(x) = k)$



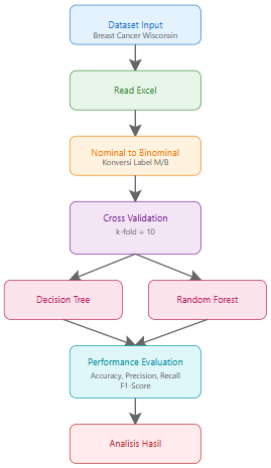
Gambar 3. Ilustrasi *Random Forest*

Dimana:
N adalah jumlah tree
hi(x) adalah prediksi tree ke-i
I adalah fungsi indikator yang bernilai 1 jika prediksi benar dan 0 jika salah
Keunggulan Random Forest terletak pada kemampuannya untuk(Devella et al., 2020):

a) Mengatasi overfitting melalui proses averaging atau voting
b) Memberikan estimasi pentingnya variabel (feature importance)
c) Menangani data yang tidak seimbang dan missing values

Metodologi Penelitian

Penelitian ini menggunakan pendekatan eksperimental dalam menganalisis perbandingan algoritma klasifikasi untuk diagnosis kanker payudara. Metodologi penelitian terdiri dari beberapa tahapan yang sistematis, dimulai dari pengumpulan data hingga analisis hasil.



Gambar 4. Alur Metodologi Penelitian

Tahap awal penelitian dimulai dengan pengumpulan data menggunakan dataset Breast Cancer Wisconsin yang terdiri dari 569 sampel dengan 32 atribut. Dataset ini dibaca menggunakan operator Read Excel di RapidMiner, yang memungkinkan data diproses lebih lanjut dalam format yang sesuai untuk analisis. Selanjutnya, dilakukan tahap preprocessing data dengan menggunakan operator Nominal to Binominal untuk mengkonversi atribut diagnosis dari format nominal (M untuk Malignant dan B untuk Benign) menjadi format binominal yang diperlukan untuk proses klasifikasi. Proses ini penting untuk memastikan data dapat diolah dengan baik oleh algoritma klasifikasi. Tahap berikutnya adalah implementasi Cross Validation dengan menggunakan k-fold = 10, yang membagi dataset menjadi 10 bagian yang sama besar. Proses ini memungkinkan evaluasi yang lebih robust terhadap performa model, dimana setiap bagian data akan mendapat kesempatan menjadi data testing, sementara bagian lainnya menjadi data training(Rifa & Kunci, 2023).

HASIL DAN PEMBAHASAN

Dataset Kanker Payudara

Dalam penelitian ini, dataset yang digunakan adalah Breast Cancer Wisconsin (Diagnostic) yang bersumber dari Database Coffee Quality Institute melalui platform GitHub. Dataset ini terdiri dari 569 sampel data dengan total 32 atribut yang merepresentasikan berbagai karakteristik sel tumor payudara. Dari keseluruhan sampel data tersebut, terdapat dua klasifikasi utama yaitu tumor ganas (Malignant) dan tumor jinak (Benign) yang ditandai dengan label 'M' untuk ganas dan 'B' untuk jinak.

Atribut-atribut dalam dataset ini mencakup pengukuran berbagai karakteristik sel tumor yang diperoleh dari analisis gambar digital hasil Fine Needle Aspirate (FNA). Setiap karakteristik sel diukur dalam tiga nilai statistik, yaitu mean (rata-rata), standard error (se), dan worst (rata-rata dari tiga nilai tertinggi), sehingga menghasilkan total 30 fitur numerik. Karakteristik yang diukur meliputi:

a) Radius (rata-rata jarak dari pusat ke titik-titik pada tepi sel)
b) Texture (standar deviasi dari nilai-nilai skala abu-abu)
c) Perimeter (ukuran keliling sel tumor)
d) Area (luas area sel tumor)
e) Smoothness (variasi lokal dalam panjang radius)
f) Compactness (perimeter² / area - 1.0)
g) Concavity (tingkat keparahan bagian cekung dari kontur sel)
h) Concave points (jumlah bagian cekung dari kontur sel)
i) Symmetry (tingkat simetri dari sel tumor)
j) Fractal dimension (pengukuran "ketidakteraturan" garis tepi sel)

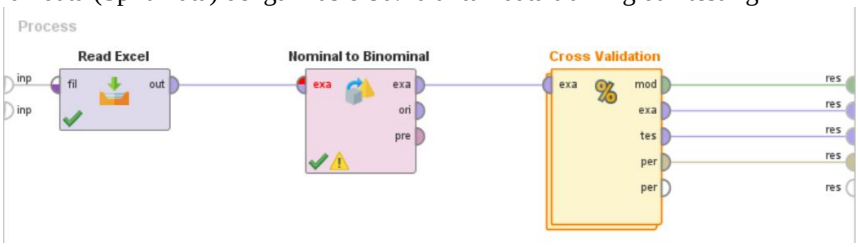
Distribusi kelas dalam dataset menunjukkan proporsi yang cukup seimbang, dengan 357 kasus tumor jinak (62.7%) dan 212 kasus tumor ganas (37.3%). Keseimbangan ini penting untuk menghindari bias dalam proses pembelajaran model.

Preprocessing Data

Pada tahap preprocessing data, penelitian ini melakukan beberapa langkah penting menggunakan tools RapidMiner untuk mempersiapkan dataset Breast Cancer Wisconsin sebelum proses klasifikasi. Dataset ini memiliki 32 atribut, dengan satu atribut sebagai label kelas (diagnosis).

Berdasarkan alur proses di RapidMiner yang ditunjukkan, tahapan preprocessing yang dilakukan adalah:

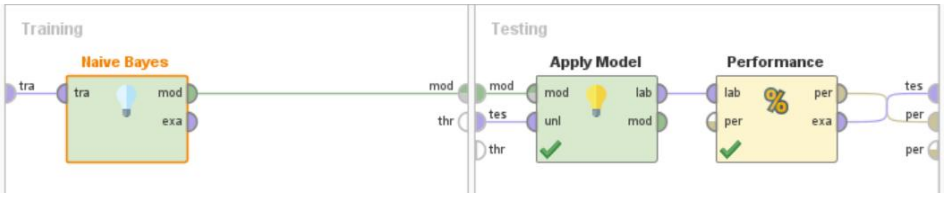
a) Pembacaan data melalui operator "Read Excel"
b) Konversi data dari nominal ke binominal untuk atribut diagnosis (M dan B)
c) Pemisahan data (Split Data) dengan rasio 80:20 untuk data training dan testing



Gambar 5. Tahapan Rapid Miner

Implementasi *Naïve Bayes*

Implementasi algoritma Naive Bayes pada penelitian ini dilakukan menggunakan tools RapidMiner dengan menggunakan operator Naive Bayes yang diterapkan pada dataset Breast Cancer Wisconsin. Proses implementasi dimulai dengan membagi data menjadi data training dan testing melalui proses cross validation dengan k-fold = 10. Operator Naive Bayes digunakan dalam proses training untuk membangun model probabilistik berdasarkan teorema Bayes.



Gambar 6. Implementasi Naïve Bayes

Berdasarkan hasil eksperimen yang telah dilakukan, algoritma Naive Bayes menunjukkan performa yang cukup baik dalam mengklasifikasikan kanker payudara. Hal ini dapat dilihat dari hasil evaluasi model yang menunjukkan nilai accuracy sebesar 93,51% dengan standar deviasi ±5,30%. Tingkat akurasi ini menunjukkan bahwa dari seluruh prediksi yang dilakukan, model Naive Bayes mampu memberikan klasifikasi yang tepat sebesar 93,51% dari seluruh kasus yang diuji.

Analisis lebih detail dapat dilihat dari confusion matrix yang dihasilkan, dimana dari total kasus malignant (M), model berhasil mengklasifikasikan dengan benar sebanyak 190 kasus dan salah mengklasifikasikan 15 kasus. Sementara untuk kasus benign (B), model berhasil mengklasifikasikan dengan benar sebanyak 342 kasus dan salah mengklasifikasikan 22 kasus. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali kedua kelas, meskipun sedikit lebih baik dalam mengenali kasus benign dibandingkan malignant. Performa model juga dapat dilihat dari nilai weighted mean recall yang mencapai 92,67% dengan standar deviasi ±6,58%. Nilai recall ini

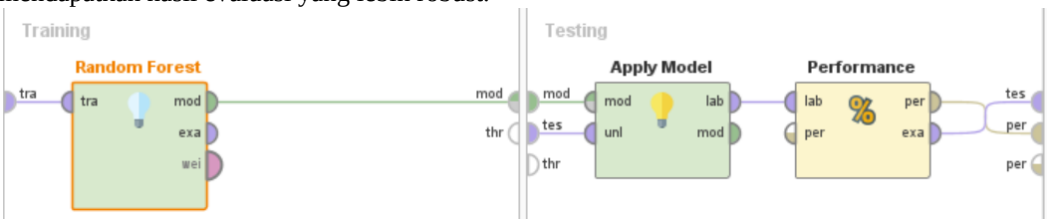
mengindikasikan kemampuan model dalam mengidentifikasi kasus positif dari seluruh kasus yang sebenarnya positif. Sementara itu, weighted mean precision mencapai 93,68% dengan standar deviasi ±4,74%, yang menunjukkan tingkat ketepatan model dalam memberikan prediksi positif.

```
PerformanceVector:
accuracy: 93.51% +/- 5.30% (micro average: 93.50%)
ConfusionMatrix:
True:   M       B
M:      190     15
B:      22      342
weighted_mean_recall: 92.67% +/- 6.58% (micro average: 92.71%), weights: 1, 1
ConfusionMatrix:
True:   M       B
M:      190     15
B:      22      342
weighted_mean_precision: 93.68% +/- 4.74% (micro average: 93.32%), weights: 1, 1
ConfusionMatrix:
True:   M       B
M:      190     15
B:      22      342
```

Gambar 7. Evaluasi Model Naive Bayes

Implementasi *Random Forest*

Implementasi algoritma Random Forest dalam penelitian ini dilakukan menggunakan tools RapidMiner dengan operator Random Forest yang diterapkan pada dataset Breast Cancer Wisconsin. Proses implementasi menggunakan pendekatan cross validation dengan k-fold = 10, dimana data dibagi menjadi data training dan testing secara berulang untuk mendapatkan hasil evaluasi yang lebih robust.



Gambar 8. Implementasi Random Forest

Berdasarkan hasil eksperimen yang telah dilakukan, algoritma Random Forest menunjukkan performa yang sangat baik dalam klasifikasi kanker payudara. Hasil evaluasi model menunjukkan tingkat accuracy yang mencapai 94,91% dengan standar deviasi ±5,06%. Nilai akurasi ini menggambarkan bahwa model Random Forest mampu memberikan prediksi yang tepat untuk 94,91% dari seluruh kasus yang diuji, yang merupakan peningkatan dibandingkan dengan performa Naive Bayes.

Analisis confusion matrix dari model Random Forest menunjukkan hasil yang lebih baik dalam hal presisi klasifikasi. Dari total kasus malignant (M), model berhasil mengklasifikasikan dengan benar sebanyak 191 kasus dan hanya salah mengklasifikasikan 8 kasus. Sementara untuk kasus benign (B), model berhasil mengklasifikasikan dengan benar sebanyak 349 kasus dan salah mengklasifikasikan 21 kasus. Hal ini menunjukkan peningkatan signifikan dalam hal mengurangi false positive dibandingkan dengan model Naive Bayes. Performa model Random Forest juga terlihat unggul dari nilai weighted mean recall yang mencapai 93,90% dengan standar deviasi ±6,53%. Peningkatan yang lebih signifikan terlihat pada weighted mean precision yang mencapai 95,33% dengan standar deviasi ±4,37%. Nilai precision yang tinggi ini menunjukkan bahwa Random Forest lebih akurat dalam memberikan prediksi positif, yang sangat penting dalam konteks diagnosis medis.

```
PerformanceVector:
accuracy: 94.91% +/- 5.06% (micro average: 94.90%)
ConfusionMatrix:
True:   M       B
M:      191     8
B:      21      349
weighted_mean_recall: 93.90% +/- 6.53% (micro average: 93.93%), weights: 1, 1
ConfusionMatrix:
True:   M       B
M:      191     8
B:      21      349
weighted_mean_precision: 95.33% +/- 4.37% (micro average: 95.15%), weights: 1, 1
ConfusionMatrix:
True:   M       B
M:      191     8
B:      21      349
```

Gambar 9. Evaluasi Model Random Forest

Evaluasi Model

Berdasarkan hasil eksperimen yang telah dilakukan, kedua algoritma yaitu Naive Bayes dan Random Forest menunjukkan performa yang berbeda dalam klasifikasi kanker payudara. Analisis komparasi dilakukan dengan membandingkan beberapa metrik evaluasi utama untuk menilai efektivitas masing-masing algoritma.

Dari segi accuracy, Random Forest menunjukkan performa yang lebih unggul dengan nilai 94,91% (±5,06%), sementara Naive Bayes mencapai 93,51% (±5,30%). Perbedaan accuracy sebesar 1,4% ini mungkin terlihat kecil, namun dalam konteks diagnosis medis, peningkatan ini dapat berarti signifikan dalam mengurangi kesalahan diagnosis.

Standar deviasi yang lebih kecil pada Random Forest juga mengindikasikan bahwa algoritma ini lebih stabil dalam memberikan prediksi.

Perbandingan Hasil Evaluasi Naive Bayes dan Random Forest

Metrik Evaluasi	Naive Bayes	Random Forest
Accuracy	93.51% ± 5.30%	94.91% ± 5.06%
Precision	93.68% ± 4.74%	95.33% ± 4.37%
Recall	92.67% ± 6.58%	93.90% ± 6.53%
False Positive	15 kasus	8 kasus
False Negative	22 kasus	21 kasus

Gambar 10. Hasil Komparasi Kedua Algoritma

Analisis confusion matrix kedua algoritma memberikan wawasan yang lebih mendalam tentang karakteristik klasifikasi masing-masing model:

a) Naive Bayes: Berhasil mengklasifikasikan 190 kasus malignant dengan benar (true positive) dan 342 kasus benign dengan benar (true negative), dengan 15 false positive dan 22 false negative.
b) Random Forest: Berhasil mengklasifikasikan 191 kasus malignant dengan benar (true positive) dan 349 kasus benign dengan benar (true negative), dengan 8 false positive dan 21 false negative.

Dari segi precision, Random Forest kembali menunjukkan keunggulan dengan weighted mean precision 95,33% (±4,37%) dibandingkan Naive Bayes yang mencapai 93,68% (±4,74%). Peningkatan precision ini terutama terlihat dari berkurangnya jumlah false positive pada Random Forest, yang penting dalam menghindari diagnosis positif palsu yang dapat menyebabkan kecemasan tidak perlu pada pasien. Dalam hal recall, kedua algoritma menunjukkan performa yang relatif mirip, dengan Random Forest mencapai 93,90% (±6,53%) dan Naive Bayes 92,67% (±6,58%). Nilai recall yang hampir setara ini menunjukkan bahwa kedua algoritma memiliki kemampuan yang sebanding dalam mengidentifikasi kasus positif, meskipun Random Forest tetap sedikit lebih unggul.

KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang telah dilakukan mengenai komparasi algoritma Naive Bayes dan Random Forest untuk klasifikasi kanker payudara menggunakan dataset Breast Cancer Wisconsin, dapat ditarik beberapa kesimpulan penting. Implementasi kedua algoritma menunjukkan performa yang baik dalam klasifikasi kanker payudara, dengan Random Forest menunjukkan hasil yang lebih unggul secara keseluruhan. Random Forest mencapai tingkat akurasi 94,91% dengan standar deviasi ±5,06%, sementara Naive Bayes mencapai akurasi 93,51% dengan standar deviasi ±5,30%. Perbedaan performa ini juga tercermin dalam metrik evaluasi lainnya, dimana Random Forest menunjukkan precision 95,33% dan recall 93,90%, sedangkan Naive Bayes mencapai precision 93,68% dan recall 92,67%.

Keunggulan Random Forest terutama terlihat dalam kemampuannya mengurangi false positive, yang sangat penting dalam konteks diagnosis medis. Algoritma ini hanya menghasilkan 8 kasus false positive dibandingkan dengan 15 kasus pada Naive Bayes. Hal ini menunjukkan bahwa Random Forest lebih efektif dalam menghindari diagnosis positif palsu yang dapat menyebabkan kecemasan tidak perlu pada pasien. Meskipun Random Forest menunjukkan performa yang lebih baik, perlu dicatat bahwa Naive Bayes juga memberikan hasil yang sangat kompetitif dengan akurasi di atas 93%. Kedua algoritma menunjukkan standar deviasi yang relatif kecil pada semua metrik evaluasi, yang mengindikasikan konsistensi dan keandalan dalam melakukan prediksi.

DAFTAR PUSTAKA

A'yunan, Y. A. D. K., Indahyanti, U., & Busono, S. (2023). Implementasi Data Mining dalam Klasifikasi Diagnosa Kanker Payudara menggunakan Algoritma Logistic Regression. *Jurnal TEKINKOM*, *6*(2), 400–407. https://doi.org/10.37600/tekinkom.v6i2.948

Alrasyid, H., Homaidi, A., Kom, M., Fatah, Z., & Kom, M. (2024). *Comparison Support Vector Machine and Random Forest Algorithms in Detect Diabetes*. *1*(1), 447–453.

Angkasa, V., & Pangaribuan, J. J. (2022). Information System Development Komparasi Tingkat Akurasi Random Forest Dan Knn Untuk Mendiagnosis Penyakit Kanker Payudara. *Journal Information System Development (ISD)*, *7*(1), 37–38. http://dx.doi.org/10.19166/xxxx

Devella, S., Yohannes, Y., & Rahmawati, F. N. (2020). Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*(2), 310–320. https://doi.org/10.35957/jatisi.v7i2.289

DOI: http://dx.doi.org/10.33846/sf12307 Faktor yang Mempengaruhi Keterlambatan Diagnosis Awal Pasien Kanker Payudara Tri Cita Pelima. (2021). *12*, 258–260.

Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam

Algoritma Klasifikasi. *Teknika*, *8*(1), 11–16. https://doi.org/10.34148/teknika.v8i1.95

Hadistio, R. R., Mawengkang, H., & Zarlis, M. (2022). *Perbandingan Algoritma Stochastic Gradient Descent dan Naïve Bayes Pada Klasifikasi Diabetic Retinopathy*. *6*, 271–277. https://doi.org/10.30865/mib.v6i1.3426

Ismai. (2017). *Data Mining: Pengolahan Data Menjadi Informasi dengan RapidMiner*.

Jalil, A., Homaidi, A., & Fatah, Z. (2024). Implementasi Algoritma Support Vector Machine Untuk Klasifikasi Status Stunting Pada Balita. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 2070–2079. https://doi.org/10.33379/gtech.v8i3.4811

Munazilin, A., & Nasta'in, M. (2023). Analisis Sentimen Pengguna Aplikasi Sistem Pembayaran UTAP Pondok Pesantren Salafiyah Syafi'iyah Situbondo. *Elektriese: Jurnal Sains Dan Teknologi Elektro*, *13*(01), 50–55. https://doi.org/10.47709/elektriese.v13i01.2581

Muntiari, N. R., & Hanif, K. H. (2022). Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Ilmu Komputer Dan Teknologi*, *3*(1), 1–6. https://doi.org/10.35960/ikomti.v3i1.766

Nugraheni, F., Anisah, F., & Susetyo, G. A. (2022). Analisis Efek Radiasi Sinar-X pada Tubuh Manusia. *Prosiding SNFA (Seminar Nasional Fisika Dan Aplikasinya)*, *1*(1), 19–25.

Rahayu, P. T., & Qurrota, A. (2022). *Jurnal Smart Teknologi Perbandingan Algoritma K-Nearest Neighbor Dan Gaussian Naïve Bayes Pada Klsifikai Penyakit Diabetes Melitus Comparison Of K-Nears Neighbor And Gaussian Naïve Bayes Algorithm On The Classification Of Diabetes Mellitus Jurnal Smart Te*. *3*(4), 366–373.

Rahmadini, A. F., D.S, R. K., & Agustiani, T. (2022). Edukasi Perilaku Pemeriksaan Payudara Sendiri (Sadari) Dalam Pencegahan Kanker Payudara Pada Remaja. *Jurnal Pemberdayaan Dan Pendidikan Kesehatan (JPPK)*, *1*(02), 105–113. https://doi.org/10.34305/jppk.v1i02.433

Rifa, Y., & Kunci, K. (2023). *Analisis Metodologi Penelitian Kulitatif dalam Pengumpulan Data di Penelitian Ilmiah pada Penyusunan Mini Riset*. *1*(1), 31–37.

Rigatti, S. J. (2017). *Random Forest*. 31–39.

Risiko, F., & Payudara, K. (2013). Faktor Risiko Kanker Payudara Wanita. *KESMAS - Jurnal Kesehatan Masyarakat*, *8*(2), 121–126. https://doi.org/10.15294/kemas.v8i2.2635

Shidqi, Z. N., Saraswati, L. D., Kusariana, N., Sutiningsih, D., & Udiyono, A. (2022). Faktor-Faktor Keterlambatan Diagnosis Kanker Pada Pasien Kanker Payudara : Systematic Review. *Jurnal Epidemiologi Kesehatan Komunitas*, *7*(2), 471–481. https://doi.org/10.14710/jekk.v7i2.14911

Sigit, A., & Yuita, A. S. (2018). Implementasi Data Mining Menggunakan Weka. In *Universitas Brawijaya Press*.