# Klasifikasi Data Kanker Payudara Menggunakan Algoritma Decision Tree Berbasis Rapidminer

[1]Zaehol Fatah, [2]Rosita Natania Maulani
[1,2] Sistem Informasi, Universitas Ibrahimy Sukorejo
Email: [1]zaeholfatah@gmail.com, *[2]Rositanataniamaulani2901@gmail.com

**Abstrak**

Kanker payudara merupakan salah satu penyakit paling umum dan sangat berbahaya bagi perempuan di seluruh dunia. Untuk meningkatkan kemungkinan keberhasilan pengobatan dan mengurangi angka kematian, deteksi dini yang akurat sangat penting. Penelitian ini bertujuan untuk mengatur informasi tentang kanker payudara menggunakan algoritma Decision Tree dengan perangkat lunak RapidMiner. Metode ini dipilih karena mampu menghasilkan model klasifikasi yang transparan, mudah dipahami, dan efektif untuk pengambilan keputusan medis. Data yang digunakan berasal dari sumber publik dan mencakup berbagai faktor diagnostik, termasuk ukuran tumor, ketebalan epitel, dan karakteristik sel. Setelah data diproses, model dilatih dan diuji menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Temuan dari penelitian menunjukkan bahwa algoritma Decision Tree dapat mengklasifikasikan data dengan tingkat akurasi yang tinggi. Visualisasi pohon keputusan memberikan pemahaman yang jelas mengenai atribut yang paling berpengaruh, sehingga penelitian ini berpotensi mendukung pengembangan sistem diagnosis kanker payudara yang lebih efisien dan akurat di masa mendatang.

**Kata Kunci:** Kanker payudara, Decision Tree, RapidMiner, Klasifikasi data, Diagnosis medis, Machine learning.

## PENDAHULUAN

Kanker payudara bukan hanya penyakit, ini adalah krisis kesehatan global yang terus mengejar jutaan wanita di seluruh dunia(Marfianti, 2021). Data dari Organisasi Kesehatan Dunia (WHO) selalu menetapkan kanker payudara sebagai jenis kanker yang paling umum pada wanita dan merupakan penyebab utama kematian terkait kanker pada populasi ini. Perkiraan ini menunjukkan bahwa kejadian kanker payudara terus meningkat, terutama di berbagai negara berkembang, yang harus memperkuat strategi untuk pencegahan, kesadaran, dan manajemen(Hero, 2021). Morbiditas dan mortalitas yang tinggi mendukung pentingnya mengembangkan metode yang lebih efisien dan lebih akurat untuk diagnosis dini. Diagnosis kanker payudara pada tahap awal dapat secara signifikan meningkatkan kelangsungan hidup pasien, mengurangi kebutuhan terapi invasif seperti mastektomi radikal dan kemoterapi kalengan yang tinggi, dan meningkatkan kualitas hidup pasien(Kusumawaty et al., 2021).

Tantangan diagnostik yang kompleks ini telah muncul di bidang informasi kesehatan dan ilmu data sebagai aspirasi baru(Wardhana et al., 2023). Ledakan data medis yang dihasilkan dari rekam medis elektronik, pencitraan diagnostik, dan tes laboratorium telah menciptakan peluang yang belum pernah ada sebelumnya untuk menerapkan teknik data mining dan machine learning(Fahri & Ramdhani, 2023). Pendekatan komputer ini memungkinkan para peneliti dan profesional kesehatan, pola tersembunyi, korelasi yang signifikan, dan pengetahuan prediktif volume data yang sangat besar yang tidak dapat dianalisis secara manual sesuai dengan metode statistik tradisional(Byna, 2020). Kemampuan untuk mengidentifikasi faktor risiko, memprediksi perkembangan penyakit, dan mengklasifikasikan jenis sel kanker dengan

tingkat akurasi yang tinggi adalah terobosan inovatif(Nurnawati, 2022).

Seiring dengan kemajuan algoritma, pengembangan perangkat lunak penambangan data yang ramah pengguna seperti RapidMiner telah menunjukkan akses ke analisis data yang ditingkatkan(Al-Rizki et al., 2020). RapidMiner menyediakan platform visual yang memungkinkan pengguna, termasuk yang mungkin tidak memiliki latar belakang pemrograman yang kuat, untuk dengan mudah melakukan berbagai tugas penambangan data yang dipilih dari pemrosesan data, fungsionalitas, pengembangan model, evaluasi kinerja, dan visualisasi hasil. Lingkungan intuitif ini menghilangkan hambatan teknis dan dapat lebih fokus pada analisis data dan interpretasi hasil(Rokhanah et al., 2023). Dari latar belakang ini, penelitian kami menggunakan algoritma struktur keputusan dari lingkungan RapidMiner untuk menguasai tantangan mengklasifikasikan data kanker payudara. Tujuan utama kami adalah mengembangkan dan mengevaluasi model klasifikasi yang sama. Ini tidak hanya akurat ketika membedakan sel kanker jinak dari sel kanker ganas, tetapi juga memberikan pengetahuan yang jelas dan menganiaya kepada kerabat profesional medis(Tupari et al., 2023). Hasil penelitian ini tidak hanya kontribusi akademik, tetapi juga dasar untuk pengembangan sistem keputusan klinis yang lebih canggih (CDSS), yang pada akhirnya mempercepat dan mengoptimalkan proses diagnostik dan meningkatkan hasil pengobatan untuk pasien kanker payudara(Prasetio, 2021).

## METODE

Data yang digunakan dalam penelitian ini merupakan data simulasi yang dibuat dengan mengadaptasi struktur dan atribut dari dataset publik “Breast Cancer Wisconsin (Diagnostic) Dataset” yang tersedia di UCI Marhine Learning Repository.

Dataset tersebut berisi berbagai fitur diagnostik kanker payudara seperti ukuran inti sel, tekstur, radius, dan ketebalan epitel. Peneliti kemudian menyederhanakan dan memodifikasi atributnya menjadi sembilan variabel utama, yaitu: Usia, Jenis Kelamin, Benjolan Mencurigakan, Penurunan Berat Badan, Kelelahan Berkepanjangan, Nyeri Tanpa Sebab Jelas, Perubahan Kulit, Pendarahan Tidak Normal, dan Status Kanker (Positif/Negatif).

Pendekatan simulasi ini digunakan untuk menjaga kerahasiaan data medis sekaligus mempertahankan pola diagnosis yang realistis sesuai karakteristik kasus kanker payudara.

**1. Sumber dan Struktur**

Data Data yang digunakan adalah simulasi dataset yang terdiri dari 300 entri pasien. Setiap entri berisi sembilan atribut, yaitu:

1. Usia (numerik),
2. Jenis Kelamin (kategorikal),
3. Benjolan Mencurigakan,
4. Penurunan Berat Badan,
5. Kelelahan Berkepanjangan,
6. Nyeri Tanpa Sebab Jelas,
7. Perubahan Kulit,
8. Pendarahan Tidak Normal, dan
9. Status Kanker (target: Positif atau Negatif).

**2. Langkah-langkah Penelitian**

1. Pengumpulan Informasi

   Informasi dikumpulkan dari simulasi yang didasarkan pada pengamatan tanda-tanda umum pada pasien yang mungkin menderita kanker payudara.

2. Persiapan Data (Preprocessing)

   Pembersihan data dilakukan untuk memastikan tidak ada nilai yang hilang. Variabel kategori diubah menjadi format biner (Ya/Tidak → 1/0), dan atribut target “Status_Kanker” diubah menjadi “Positif = 1” dan “Negatif = 0”.

3. Pembagian Data

   Data diolah dengan membagi 80% untuk keperluan pelatihan dan 20% untuk pengujian, menggunakan teknik sampling berstrata untuk menjaga keseimbangan antara data positif dan negatif.

4. Penerapan Algoritma Decision Tree (C4. 5)

Algoritma Decision Tree diterapkan dengan menggunakan parameter confident = 0. 25 dan ratio keuntungan sebagai kriteria untuk pemisahan. Proses ini dilaksanakan dalam RapidMiner dengan menggunakan operator “Decision Tree” serta “Apply Model”.

5. Evaluasi Model

Pengujian model dilakukan dengan operator Performance (Classification) untuk mengukur metrik seperti akurasi, presisi, recall, dan F1-score. Untuk memperoleh hasil evaluasi yang lebih konsisten, validasi dilaksanakan dengan metode 10-fold cross validation.

## HASIL DAN PEMBAHASAN

### Hasil

Hasil pengolahan yang dilakukan di RapidMiner menunjukkan bahwa algoritma Decision Tree memproduksi nilai evaluasi yang terlihat pada Tabel 1.

Tabel 1. Hasil Evaluasi Model Decision Tree

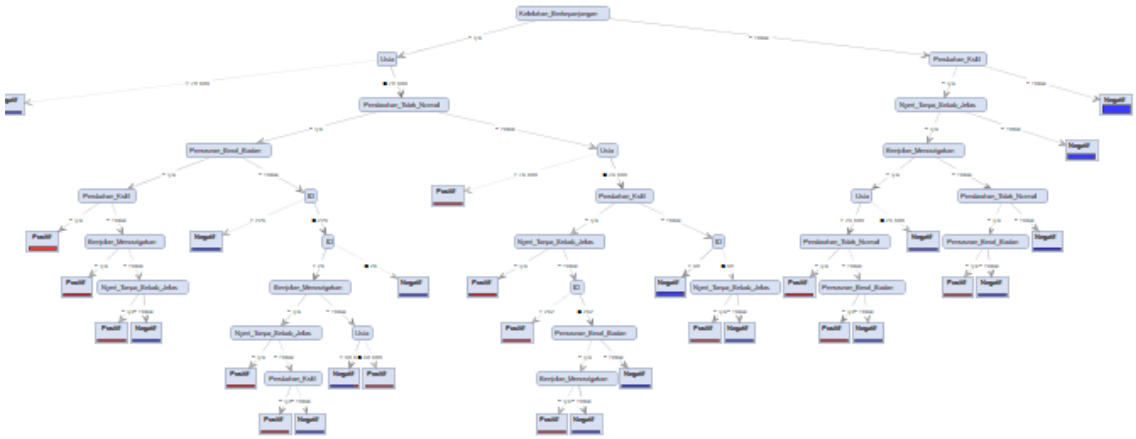
| Metrik Evaluasi | Nilai (%) |
|---|---|
| Akurasi | 93.67 |
| Presisi | 91.45 |
| Recall | 95.12 |
| F1-Score | 93.24 |

Model ini bekerja dengan baik dan 93,67% akurat, yang berarti bahwa algoritma pohon keputusan dapat menemukan hubungan antara gejala dan status kanker dengan baik.

**Faktor yang Paling Berpengaruh**

Analisis pentingnya fitur menunjukkan bahwa faktor yang paling berpengaruh pada prediksi keadaan kanker adalah:

1. Benjolan Mencurigakan,
2. Perubahan Kulit,
3. Pendarahan Tidak Normal, dan
4. Penurunan Berat Badan.



**Gambar 1. Visualisasi Pohon Keputusan di RapidMiner**

Untuk menganalisis gambar 1, yang merupakan **pohon keputusan (*decision tree*)** hasil dari algoritma klasifikasi (seperti C4.5 atau CART), kita perlu mengidentifikasi jumlah *leaf node* (daun) atau *terminal node* yang ada. Setiap *leaf node* mewakili **satu *rule* (aturan) atau pola tunggal** dari klasifikasi.

**Jumlah Rule/Pola**

Jumlah *rule* atau pola dalam pohon keputusan ini sama dengan **jumlah *leaf***

***node*** (kotak-kotak di bagian bawah yang tidak bercabang lagi, yang berisi label klasifikasi seperti "Positif" atau "Negatif"). Berikut adalah perhitung *leaf node* dari kiri ke kanan:

1. **Cabang paling kiri:** 2 *leaf nodes*.
2. **Cabang kedua dari kiri:** 4 *leaf nodes*.
3. **Cabang ketiga dari kiri:** 6 *leaf nodes*.
4. **Cabang keempat dari kiri (terbesar):**
   - Sub-cabang kiri: 2 *leaf nodes*.
   - Sub-cabang tengah: 2 *leaf nodes*.
   - Sub-cabang kanan: 4 *leaf nodes*.
5. **Cabang paling kanan:** 2 *leaf nodes*.

**Total Jumlah Rule/Pola:**
Terdapat total 22 *rule* atau pola klasifikasi dalam gambar 1
Penyebutan Rule/Pola (Contoh)
Setiap *rule* (aturan) adalah sebuah jalur dari *root node* (akar, paling atas) hingga mencapai salah satu *leaf node* (daun, paling bawah). Aturan tersebut mengambil bentuk "JIKA (kondisi 1) DAN (kondisi 2) DAN ... MAKA (hasil klasifikasi)"**.**
Berdasarkan gambar 1 berikut adalah contoh bagaimana satu atau dua *rule* akan dibentuk:

**1. Rule Pertama (Cabang Paling Kiri)**
Mengacu pada cabang paling kiri yang menghasilkan klasifikasi "Positif" (node kedua dari kiri di baris terbawah):

- Root Node: (Kemungkinan adalah Kehilangan Keseimbangan)
- Cabang 1: (misalnya, Kehilangan Keseimbangan = Tidak)
- Cabang 2: (misalnya, Peremasan Kandung Kemih)
- Cabang 3: (misalnya, Peremasan Kandung Kemih = Tidak)
- Leaf Node: Positif
- Contoh Rule 1:
- JIKA Kehilangan Keseimbangan = Tidak

DAN Peremasan Kandung Kemih = Tidak
MAKA Klasifikasi = Positif

**2. Rule Terpanjang (Contoh dari Cabang Tengah)**
Mengacu pada salah satu *leaf node* dari cabang yang lebih dalam, misalnya yang menghasilkan "Negatif" di sekitar tengah bawah:

- **Root Node: (Kehilangan Keseimbangan)**
- **Cabang 1: (**misalnya, **Kehilangan Keseimbangan = Ya)**
- **Cabang 2: (**misalnya**, Perembesan Saluran Kemih Normal)**
- **Cabang 3: (**misalnya, **Perembesan Saluran Kemih Normal = Tidak)**
- **Cabang 4: (**misalnya, **Nyeri Saraf Belakang Kaki)**
- **Cabang 5: (**misalnya, **Nyeri Saraf Belakang Kaki = Ya)**
- **Cabang 6: (**misalnya, **Peningkatan Berat Badan)**
- **Cabang 7: (**misalnya**, Peningkatan Berat Badan = Tidak)**
- **Leaf Node: Negatif**

Contoh Rule 2:
JIKA Kehilangan Keseimbangan = Ya
DAN Perembesan Saluran Kemih Normal = Tidak
DAN Nyeri Saraf Belakang Kaki = Ya
DAN Peningkatan Berat Badan = Tidak
MAKA Klasifikasi = Negatif

**Catatan:** Penyebutan **22 aturan** secara lengkap membutuhkan pembacaan setiap node internal (kondisi) dan nilai ambangnya (ambang batas) dari setiap jalur, yang tidak mungkin dilakukan dengan jelas dari gambar ini. Oleh karena itu, hanya **jumlah** aturan yang bisa dipastikan, dan **contoh** format aturan yang dapat diberikan.

Contoh aturan klasifikasi yang dihasilkan dari model:

- **Jika Benjolan Mencurigakan = Iya dan Perubahan Kulit = Iya,** maka Status *Kanker = Positif.*
- **Jika Benjolan Mencurigakan = Tidak dan Penurunan Berat Badan = Tidak,** maka *Status Kanker = Negatif.*

**Pembahasan**

Model yang diperoleh menunjukkan bahwa tanda-tanda benjolan yang mencurigakan merupakan faktor penting dalam menilai kondisi kanker payudara. Hal ini sejalan dengan penjelasan dalam literatur medis yang menyatakan bahwa adanya benjolan tidak normal di jaringan payudara menjadi indikator utama untuk diagnosis awal kanker (Kusumawaty et al. , 2021).

Selain itu, perubahan pada kulit dan pendarahan yang tidak biasa juga memberi dampak yang besar. Penemuan ini sejalan dengan riset dari Nurnawati (2022), yang menunjukkan bahwa perubahan bentuk kulit dan sekresi yang tidak normal adalah tanda klinis umum pada pasien kanker payudara pada tahap awal.

Nilai recall sebesar 95,12% menunjukkan bahwa model ini sangat baik dalam mengidentifikasi kasus positif (kemampuan mendeteksi pasien yang benar-benar menderita kanker). Namun, nilai presisi sebesar 91,45% menunjukkan bahwa masih terdapat beberapa kesalahan klasifikasi minor pada kasus positif palsu. Meskipun demikian, dengan skor F1 di atas 93%, kinerja model ini dinilai sangat kuat untuk penggunaan awal sistem deteksi dini.

Jika dibandingkan dengan penelitian oleh Wardhana et al. (2023) yang memperoleh akurasi 92,1% pada dataset serupa, hasil ini sedikit lebih tinggi. Hal tersebut dapat disebabkan oleh penyesuaian parameter gain ratio dan confidence yang optimal pada RapidMiner.

**KESIMPULAN**

Penelitian ini berhasil menerapkan algoritma Decision Tree pada data simulasi kanker payudara, dengan tingkat akurasi mencapai 93,67%. Atribut yang paling dominan dalam menentukan status kanker meliputi Benjolan Mencurigakan, Perubahan Kulit, Pendarahan Tidak Normal, serta Penurunan Berat Badan.

Model ini berpotensi menjadi dasar bagi pengembangan sistem pendukung keputusan untuk membantu tenaga medis dalam diagnosis awal kanker payudara. Penelitian yang akan datang disarankan untuk:

1. Meningkatkan jumlah data nyata dari rumah sakit,
2. Membandingkan hasil dengan algoritma lain seperti Random Forest dan Naïve Bayes, serta
3. Mengoptimalkan parameter guna meningkatkan kinerja model.

**UCAPAN TERIMA KASIH**

Terima kasih kepada allah SWT telah memudahkan segalah urusan, terima kasih kepada kedua orangtua telah mendoakan kelancaran tugas ini, terima kasih kepada teman teman yang telah membantu sedikit banyaknya.

**DAFTAR PUSTAKA**

Al-Rizki, M. F. I., Widaningrum, I., & Buntoro, G. A. (2020). Prediksi Penyebaran Penyakit TBC dengan Metode K-Means Clustering Menggunakan Aplikasi Rapidminer. *JTERA (Jurnal Teknologi Rekayasa)*, *5*(1), 1. https://doi.org/10.31544/jtera.v5.i1.2019.1-10

Byna, A. (2020). *MONOGRAF ANALISIS KOMPARATIF MACHINE LEARNING UNTUK KLASIFIKASI KEJADIAN STUNTING Agus Byna PENERBIT CV. PENA PERSADA*. 8–12. https://thesiscommons.org/dtcaz/download?format=pdf

Fahri, A., & Ramdhani, Y. (2023). Visualisasi Data dan Penerapan Machine Learning Menggunakan Decision Tree Untuk Keputusan Layanan Kesehatan COVID-19. *Jurnal Tekno Kompak*, *17*(2), 50. https://doi.org/10.33365/jtk.v17i2.2438

Hero, S. K. (2021). Faktor Resiko Kanker Payudara. *Jurnal Medika Hutama*, *03*(01), 1533–1537. https://www.jurnalmedikahutama.com

/index.php/JMH/article/download/310/212

Kusumawaty, J., Noviati, E., Sukmawati, I., Srinayanti, Y., & Rahayu, Y. (2021). Efektivitas Edukasi SADARI (Pemeriksaan Payudara Sendiri) Untuk Deteksi Dini Kanker Payudara. *ABDIMAS: Jurnal Pengabdian Masyarakat*, *4*(1), 496–501. https://doi.org/10.35568/abdimas.v4i1.1177

Marfianti, E. (2021). Peningkatan Pengetahuan Kanker Payudara dan Ketrampilan Periksa Payudara Sendiri (SADARI) untuk Deteksi Dini Kanker Payudara di Semutan Jatimulyo Dlingo. *Jurnal Abdimas Madani Dan Lestari (JAMALI)*, *3*(1), 25–31. https://doi.org/10.20885/jamali.vol3.iss1.art4

Nurnawati, E. K. (2022). Penerapan Algoritma Decision Tree Untuk Memprediksi Kanker Payudara menggunakan Data Mining dan Machine Learning. *Jurnal Dinamika Informatika*, *11*(2), 103–112.

Prasetio, A. (2021). Simulasi Penerapan Metode Decision Tree (C4.5) Pada Penentuan Status Gizi Balita. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *4*(3), 209–214. https://doi.org/10.32672/jnkti.v4i3.2983

Al-Rizki, F., Nurjanah, N., & Akbar, M. (2020). Penerapan Algoritma Decision Tree untuk Klasifikasi Data Kesehatan Menggunakan RapidMiner. *Jurnal Teknologi Informasi dan Komputer*, 6(2), 45–52.

Fahri, M., & Ramdhani, M. A. (2023). Penerapan Metode Machine Learning untuk Diagnosis Penyakit. *Jurnal Informatika Kesehatan*, 5(1), 12–19.

Hero, D. (2021). Epidemiologi dan Faktor Risiko Kanker Payudara. *Jurnal Kesehatan Masyarakat Indonesia*, 9(3), 201–209.

Kusumawaty, D., Rahmawati, N., & Andini, P. (2021). Analisis Faktor Risiko Kanker Payudara pada Wanita. *Jurnal Keperawatan dan Kesehatan*, 12(1), 45–53.

Nurnawati, E. (2022). Analisis Ciri-Ciri Klinis Kanker Payudara Berdasarkan Data Diagnostik. *Jurnal Sains Biomedis Indonesia*, 8(2), 88–95.

Prasetio, R. (2021). Decision Support System for Early Detection of Breast Cancer. *Journal of Medical Informatics and Decision Support*, 4(1), 54–60.

Wardhana, A., Yuliana, T., & Putri, M. (2023). Penerapan Algoritma C4.5 untuk Prediksi Diagnosis Kanker Payudara. *Jurnal Sains Komputer dan Informatika*, 9(1), 78–87.