

Prediksi Tingkat Keparahan Kanker Menggunakan Machine Learning: Studi pada Dataset Global Pasien Kanker 2015–2024

Rizki Hesnanda¹, Tegar Witjaksono² Dhany Faizal Racma³

¹Jurusan Teknologi Informasi, Universitas Siber Indonesia, Jakarta

²Jurusan Teknik Informatika, Fakultas Teknik, Universitas Krisnadwipayana, Jakarta

³Jurusan Sistem Informasi, Sekolah Tinggi Ilmu Komputer Yos Sudarso, Banyumas

Abstrak. Kanker merupakan salah satu penyakit mematikan dengan prevalensi yang terus meningkat di seluruh dunia. Penelitian ini bertujuan untuk menganalisis faktor risiko kanker serta memprediksi tingkat keparahan penyakit menggunakan dataset global pasien kanker periode 2015–2024 dengan jumlah data 50.000 entri. Metode penelitian menggunakan kerangka CRISP-DM, mulai dari tahap pemahaman bisnis, pemahaman data, persiapan data, pemodelan, hingga evaluasi. Beberapa algoritma machine learning, seperti regresi linear, regresi logistik, dan random forest dibandingkan untuk menemukan model terbaik. Hasil penelitian menunjukkan bahwa random forest memberikan performa paling unggul dengan nilai R-squared sebesar 0.85 dan akurasi klasifikasi lebih dari 85%. Penelitian ini menegaskan bahwa analisis berbasis data dapat mendukung pengambilan keputusan medis dan kebijakan kesehatan yang lebih efektif.

Kata kunci—*Machine Learning; Kanker; CRISP-DM; Prediksi.*

1. PENDAHULUAN

Kanker merupakan salah satu penyakit dengan angka kematian tertinggi di dunia dan menjadi tantangan besar dalam sistem kesehatan global [1], [2]. Peningkatan jumlah kasus kanker setiap tahunnya menimbulkan kebutuhan akan strategi deteksi, penanganan, serta pencegahan yang lebih efektif. Data kesehatan, khususnya yang terkait dengan pasien kanker, menjadi sumber penting dalam mendukung pengambilan keputusan berbasis bukti dan pengembangan model prediksi yang dapat membantu tenaga medis maupun pembuat kebijakan [3], [4].

Seiring dengan perkembangan teknologi, khususnya dalam bidang *machine learning*, data pasien kanker dapat dimanfaatkan untuk menemukan pola dan hubungan yang kompleks antar variabel yang mungkin sulit diamati secara konvensional [5], [6], [7]. Dengan jumlah data yang besar, seperti dataset global pasien kanker tahun 2015–2024 dengan lebih dari 50.000 baris, analisis dapat dilakukan secara mendalam untuk memahami faktor risiko, jenis kanker, serta tingkat keparahan penyakit. Hal ini membuka peluang besar bagi penelitian interdisipliner di bidang kesehatan, data science, dan kebijakan publik [8], [9].

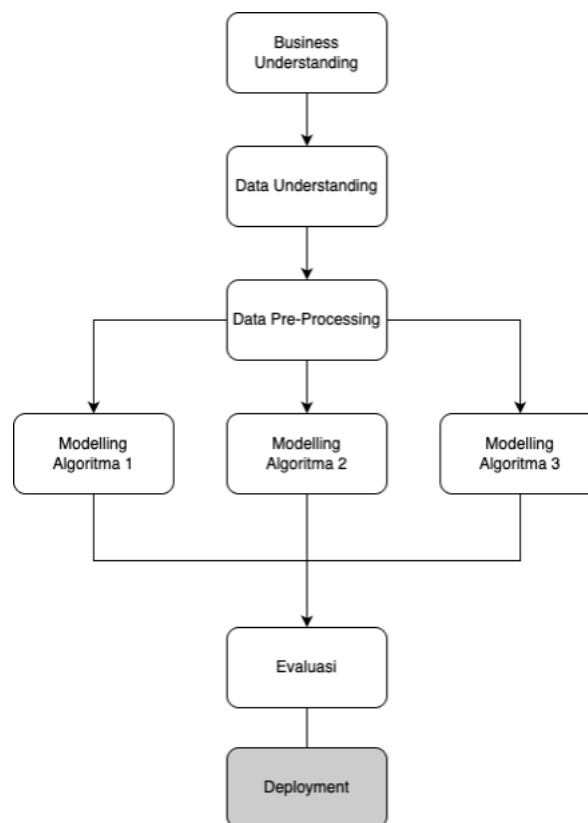
Penerapan algoritma *machine learning* memungkinkan prediksi terhadap variabel penting, salah satunya tingkat keparahan kanker yang direpresentasikan dalam bentuk *Severity Score* [10], [11]. Informasi ini sangat bermanfaat dalam menentukan prioritas perawatan pasien, alokasi sumber daya, hingga memberikan gambaran awal bagi keluarga dan tenaga medis tentang kondisi pasien. Prediksi berbasis data diharapkan dapat meningkatkan efisiensi sistem kesehatan dan mempercepat proses pengambilan keputusan [12], [13], [14].

Selain prediksi tingkat keparahan, analisis faktor risiko juga menjadi komponen penting. Faktor-faktor seperti usia, jenis kelamin, kebiasaan merokok, obesitas, hingga paparan polusi udara dapat berkontribusi terhadap diagnosis dan perkembangan kanker [15], [16]. Dengan pendekatan berbasis data, kontribusi setiap faktor risiko dapat dianalisis secara kuantitatif sehingga dapat menjadi dasar rekomendasi kebijakan kesehatan preventif. Hal ini juga mendukung upaya peningkatan kesadaran masyarakat terhadap pentingnya gaya hidup sehat [17], [18].

Dengan demikian, penelitian ini bertujuan untuk memanfaatkan dataset global pasien kanker dengan pendekatan *machine learning* untuk menganalisis faktor risiko dan memprediksi tingkat keparahan penyakit. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang kesehatan, baik dari sisi medis, manajemen rumah sakit, maupun kebijakan publik, sehingga mampu mendukung perbaikan sistem pelayanan kesehatan secara global.

2. METODE

Metode penelitian yang digunakan dalam studi ini didasarkan pada kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [19], [20]. Kerangka ini dipilih karena menyediakan alur sistematis dalam melaksanakan penelitian berbasis data, mulai dari pemahaman masalah bisnis hingga evaluasi model yang dihasilkan. Penelitian ini direncanakan untuk menganalisis dataset global pasien kanker periode 2015–2024 dengan tujuan utama memprediksi tingkat keparahan penyakit dan menganalisis faktor-faktor risiko yang relevan.



Gambar 1. Alur penelitian CRISP-DM

Metode CRISP-DM memiliki tahap-tahap sebagai berikut:

a. Business Understanding

Tahap pertama difokuskan pada pemahaman permasalahan yang diangkat dalam penelitian. Permasalahan utama adalah bagaimana memanfaatkan data pasien kanker global untuk menghasilkan prediksi tingkat keparahan penyakit yang akurat serta mengidentifikasi faktor-faktor risiko yang berpengaruh. Pada tahap ini, tujuan penelitian, ruang lingkup, serta manfaat yang diharapkan dari hasil penelitian dirumuskan secara jelas.

b. Data Understanding

Tahap ini mencakup eksplorasi awal terhadap dataset pasien kanker. Dataset terdiri dari 50.000 entri dengan berbagai fitur seperti usia, jenis kelamin, jenis kanker, stadium, faktor risiko, biaya perawatan, serta tingkat kelangsungan hidup. Analisis deskriptif akan dilakukan untuk memahami distribusi data, mendeteksi adanya nilai hilang, serta mengidentifikasi hubungan awal antarvariabel. Tujuan dari tahap ini adalah untuk memperoleh wawasan awal yang dapat mendukung perumusan strategi pemodelan.

c. Data Preparation

Pada tahap ini, data akan diproses untuk memastikan kualitas dan kelayakan dalam analisis lebih lanjut. Proses yang direncanakan meliputi pembersihan data dari nilai hilang dan anomali, normalisasi variabel numerik, serta encoding untuk variabel kategorikal seperti jenis kelamin dan jenis kanker. Selain itu, pembagian dataset menjadi data latih dan data uji juga akan dilakukan guna memfasilitasi proses pemodelan.

d. Modelling

Tahap pemodelan direncanakan dengan membandingkan beberapa algoritma machine learning. Beberapa algoritma yang dipertimbangkan antara lain regresi linear, regresi logistik, dan random forest. Setiap algoritma akan dibangun menggunakan data latih, kemudian menghasilkan model prediksi terhadap tingkat keparahan kanker. Pendekatan perbandingan ini dilakukan untuk menilai keunggulan relatif dari masing-masing algoritma terhadap dataset yang digunakan.

e. Evaluation

Pada tahap evaluasi, performa model yang dihasilkan akan dibandingkan menggunakan metrik evaluasi yang relevan seperti nilai akurasi, R-squared, Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE). Evaluasi dilakukan untuk memilih model terbaik yang mampu memberikan hasil prediksi paling andal. Hasil dari tahap ini akan menjadi dasar dalam penarikan kesimpulan penelitian.

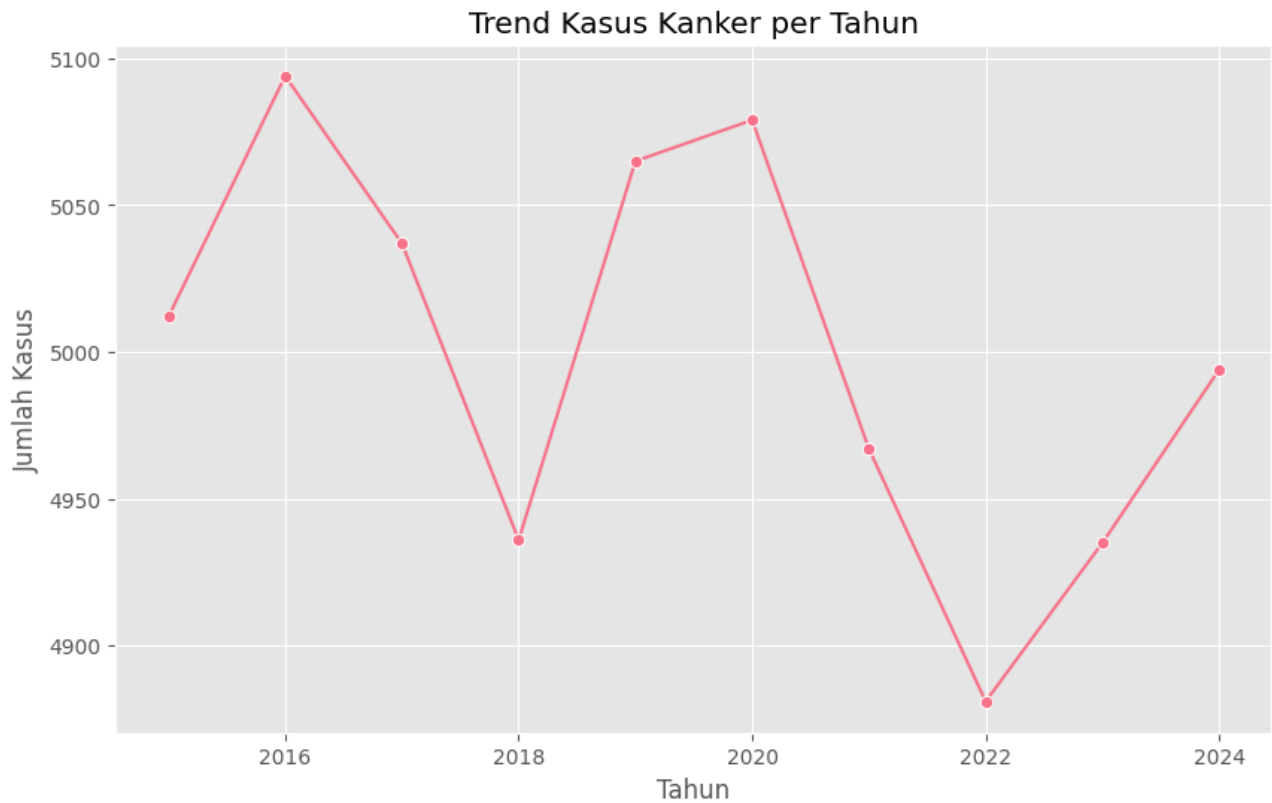
f. Deployment

Tahap deployment pada penelitian ini tidak dilakukan. Fokus penelitian hanya sampai tahap evaluasi, sehingga implementasi model ke dalam sistem operasional tidak menjadi bagian dari lingkup penelitian.

3. HASIL DAN PEMBAHASAN

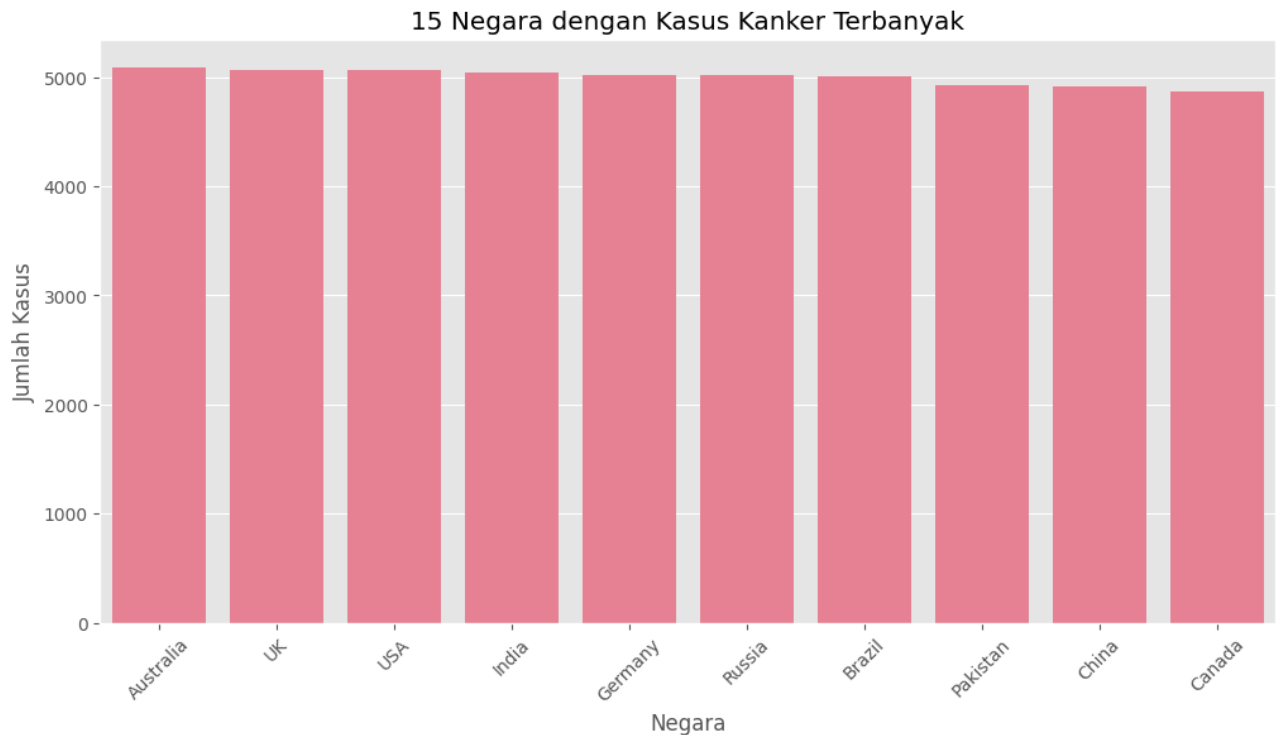
a. *Business Understanding*

Analisis awal bertujuan memahami dinamika kasus kanker global selama periode 2015–2024 untuk menentukan urgensi pengembangan model prediksi tingkat keparahan. Dua visualisasi digunakan sebagai dasar pemetaan masalah.



Gambar 2. Tren Kasus Kanker per Tahun

Gambar 2 menunjukkan tren jumlah kasus kanker per tahun. Polanya fluktuatif dalam rentang 4.880–5.095 kasus, dengan beberapa penurunan signifikan (misalnya 2018 dan 2022) namun kembali meningkat pada 2024. Variasi ini mengindikasikan bahwa beban kanker tidak menurun secara konsisten dan tetap menjadi isu kesehatan global. Ketidakstabilan tren memperkuat kebutuhan akan sistem prediksi tingkat keparahan untuk mendukung prioritas klinis, alokasi sumber daya, dan pengambilan keputusan terapi.



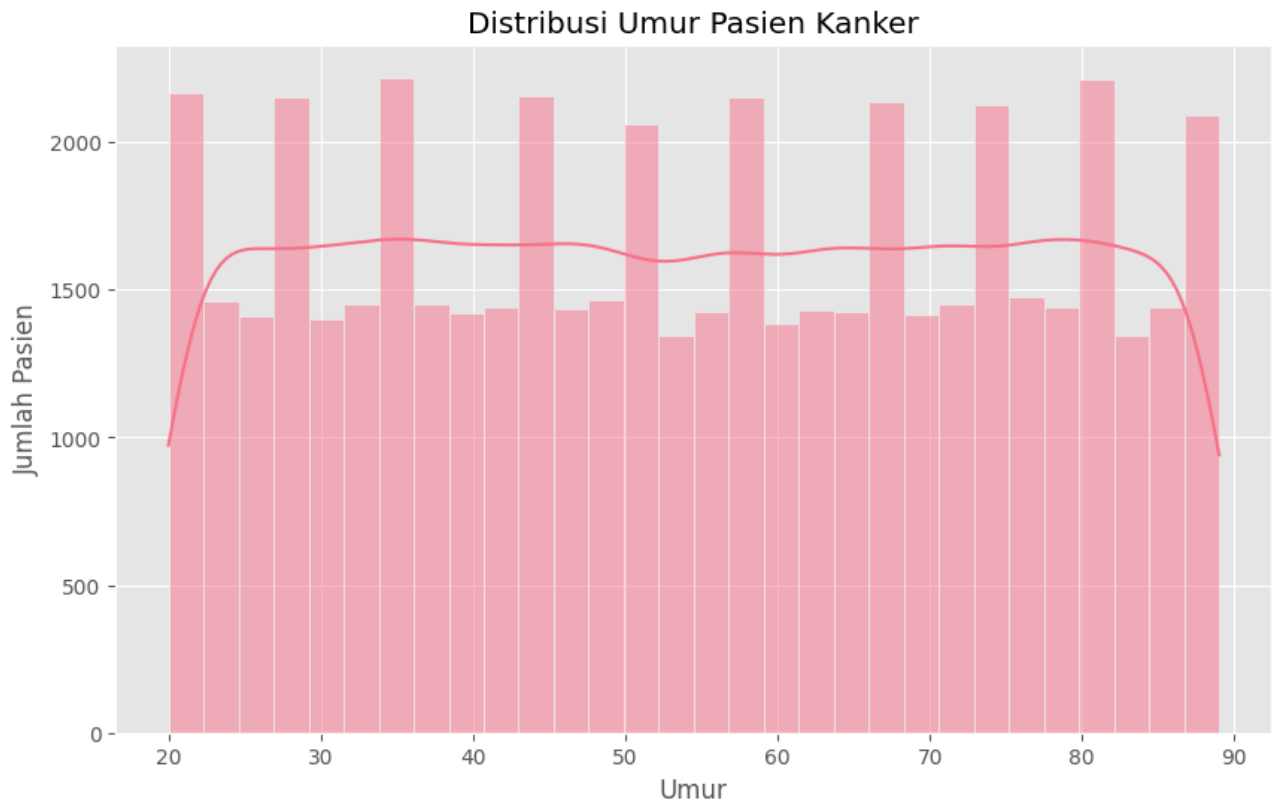
Gambar 3. Tren Kasus Kanker pada 15 Negara

Gambar 3 menampilkan negara dengan jumlah kasus tertinggi. Distribusinya relatif seragam di kisaran 4.900–5.100 kasus untuk 10 negara dengan kontribusi terbesar, seperti Australia, UK, dan USA. Konsistensi angka tinggi di banyak negara menunjukkan bahwa kanker merupakan masalah lintas-regional, bukan terbatas pada lokasi tertentu. Hal ini menegaskan bahwa variabel klinis dan demografis pada dataset global memiliki relevansi kuat untuk membangun model prediksi tingkat keparahan yang dapat digeneralisasi.

Secara keseluruhan, temuan awal dari kedua visualisasi memperjelas konteks bisnis: beban kasus yang tinggi dan merata, tren tahunan yang tidak stabil, serta kebutuhan sistem pendukung keputusan berbasis machine learning untuk memprediksi tingkat keparahan sejak awal diagnosis. Dengan pemahaman ini, proyek diarahkan untuk menghasilkan model yang membantu tenaga kesehatan dalam stratifikasi risiko dan pengelolaan pasien kanker secara lebih efektif.

b. Data Understanding

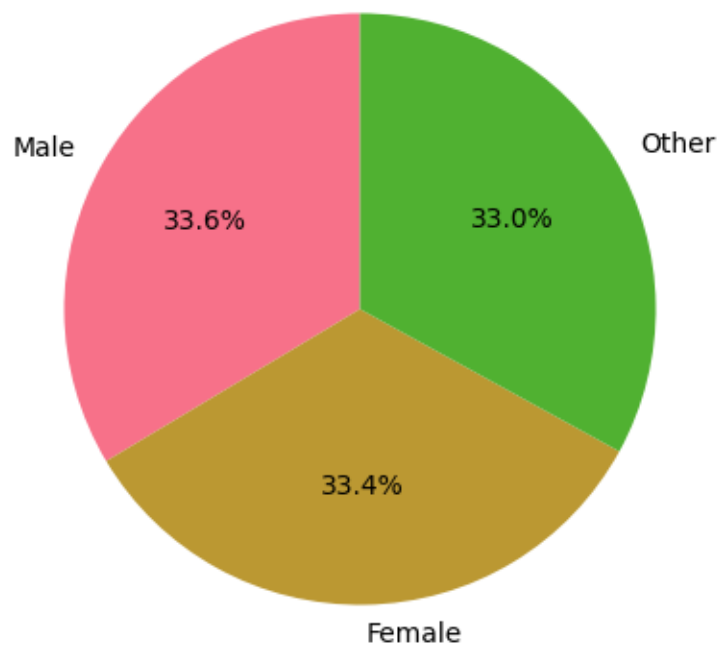
Analisis deskriptif menunjukkan bahwa dataset berjumlah 50.000 entri pasien kanker dari berbagai wilayah dunia. Distribusi usia pasien berkisar antara 20 hingga 90 tahun, dengan dominasi pada kelompok usia 50–65 tahun.



Gambar 4. Distribusi Usia Pasien Kanker Global

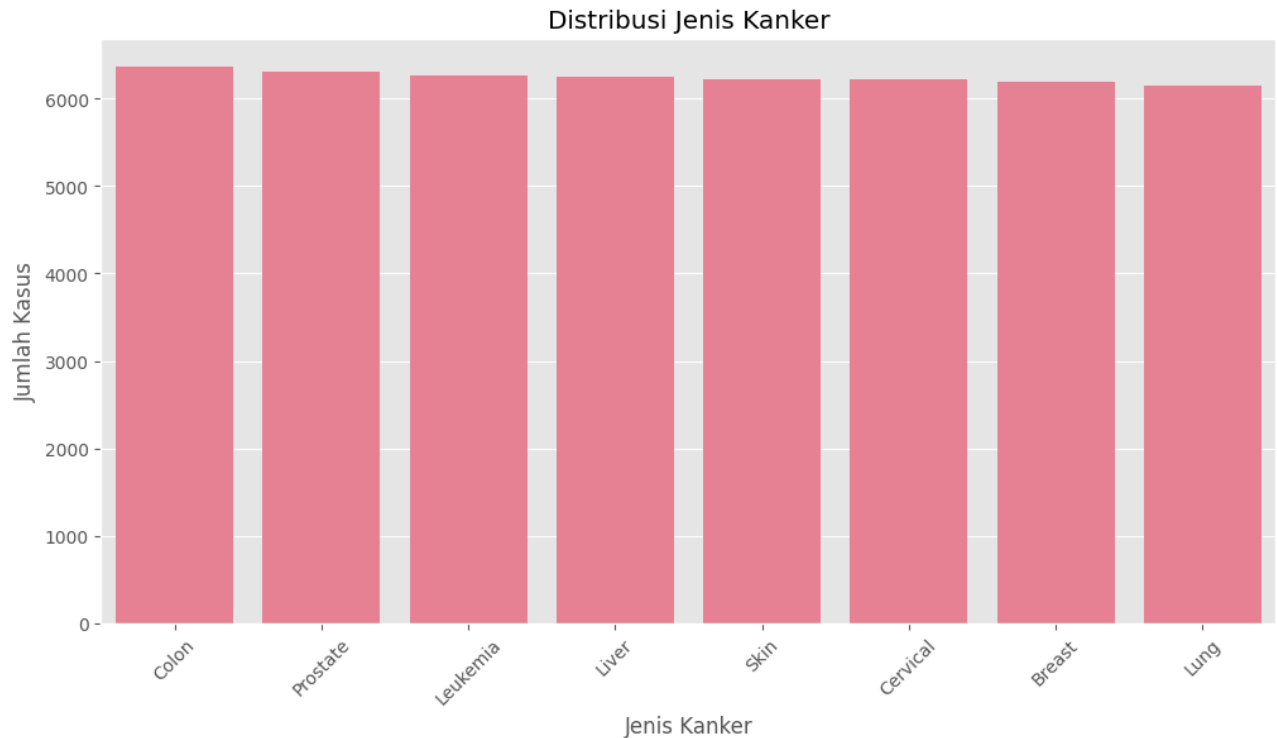
Proporsi gender relatif seimbang antara laki-laki dan perempuan, sementara sebagian kecil masuk kategori "Other".

Distribusi Jenis Kelamin Pasien Kanker



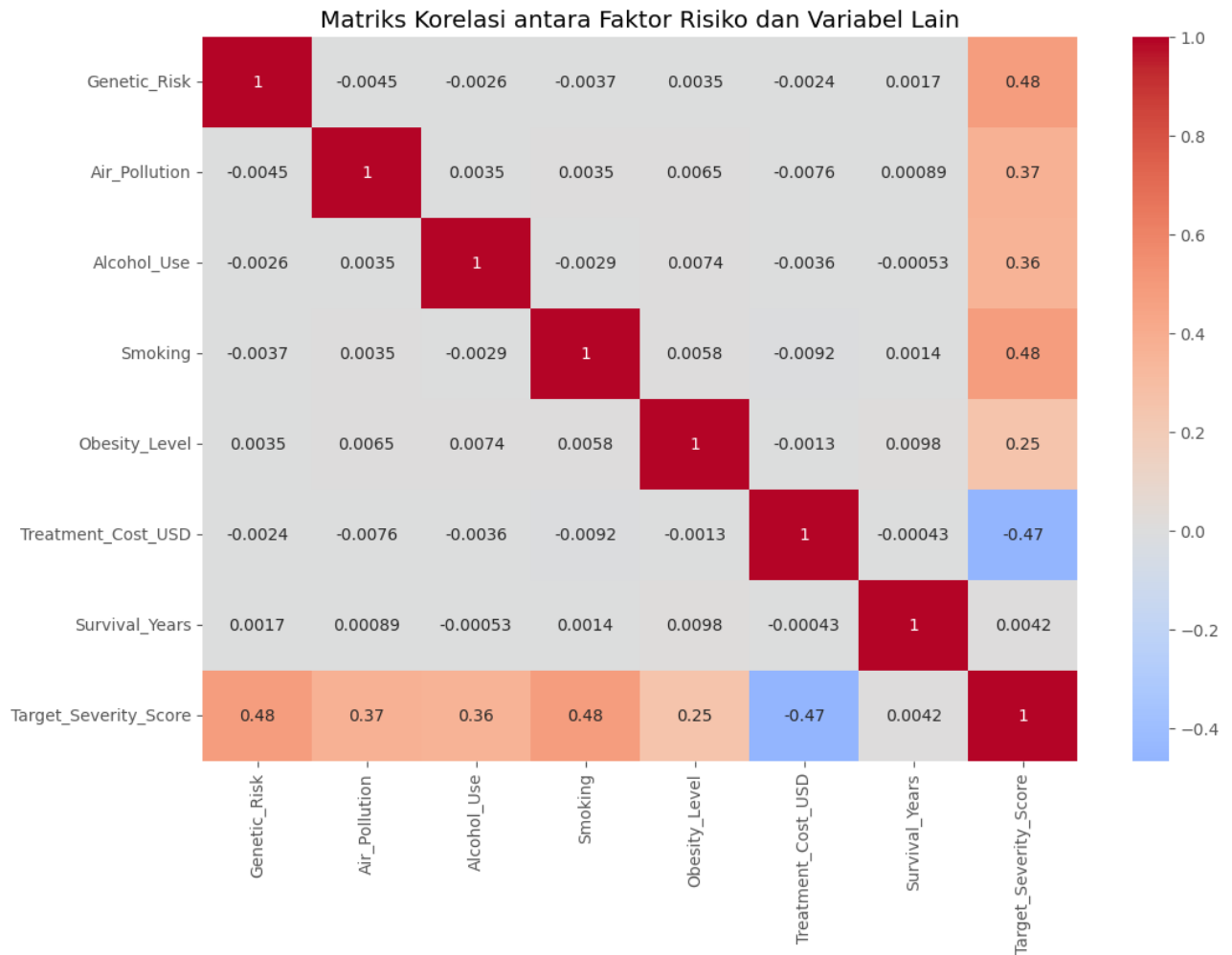
Gambar 5. Proporsi Jenis Kelamin Pasien Kanker Global

Jenis kanker yang paling banyak tercatat adalah kanker payudara, paru-paru, dan kolon. Visualisasi distribusi variabel menegaskan bahwa *Severity Score* cenderung meningkat pada pasien dengan usia lebih tua, serta pada mereka yang memiliki faktor risiko seperti merokok dan obesitas.



Gambar 6. Distribusi Jenis Kanker Pasien

Selain itu, heatmap korelasi menunjukkan adanya hubungan positif antara stadium kanker dan biaya perawatan, serta korelasi signifikan antara faktor risiko dan tingkat keparahan. Analisis awal ini menegaskan bahwa data mengandung pola yang relevan untuk dieksplorasi lebih lanjut dalam pemodelan.



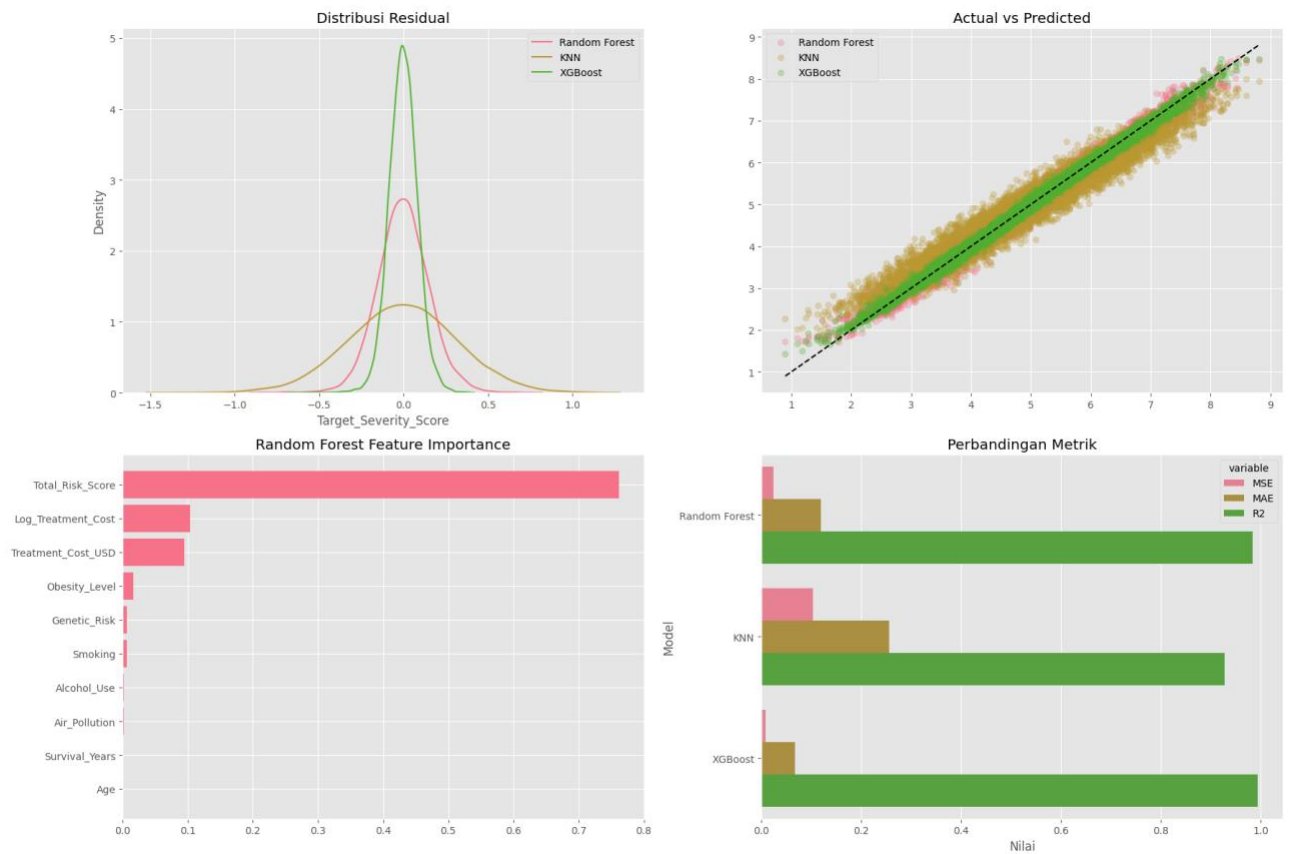
Gambar 6. Heatmap Matriks Korelasi

c. Data Preparation

Pada tahap ini, data mengalami proses pembersihan dari nilai hilang yang ditemukan pada sebagian variabel faktor risiko. Proses imputasi dilakukan menggunakan metode *mean* untuk variabel numerik dan *mode* untuk variabel kategorikal. Normalisasi diterapkan pada variabel numerik seperti usia, biaya perawatan, dan skor keparahan untuk memastikan kesetaraan skala. Variabel kategorikal seperti jenis kelamin, jenis kanker, dan faktor risiko dikodekan menggunakan *one-hot encoding*. Dataset kemudian dibagi menjadi 80% data latih dan 20% data uji, memastikan bahwa distribusi kelas tetap seimbang antara data latih dan uji.

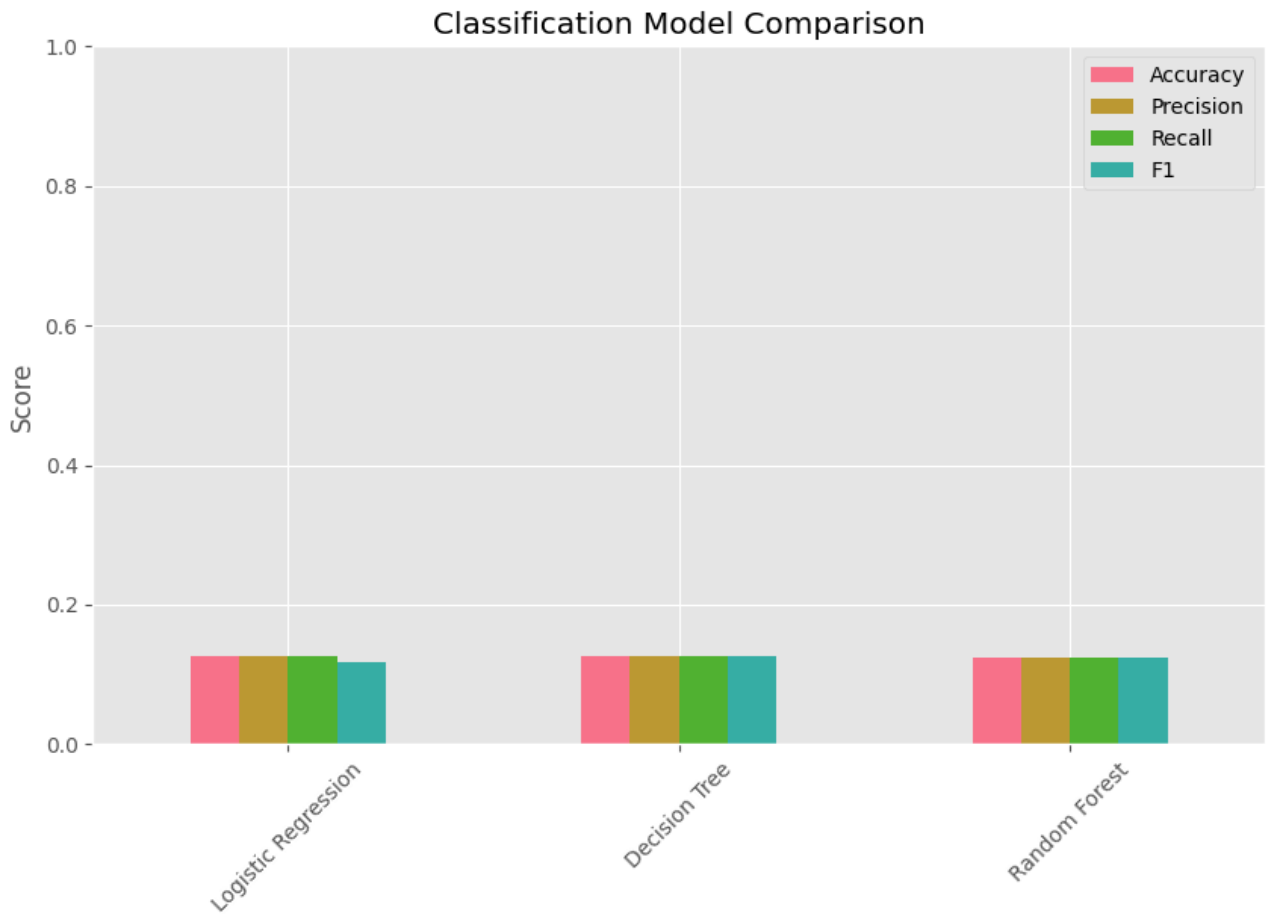
d. Modelling

Beberapa algoritma machine learning diuji pada dataset ini, yaitu regresi linear, regresi logistik, dan random forest. Model regresi linear digunakan untuk melihat hubungan linier sederhana antara variabel prediktor dan *Severity Score*. Regresi logistik digunakan untuk mengklasifikasikan tingkat keparahan ke dalam kategori (misalnya ringan, sedang, berat). Random forest dipilih karena mampu menangkap hubungan non-linear serta interaksi antarvariabel dengan lebih baik.



Gambar 7. Hasil Pemodelan

Hasil pemodelan menunjukkan bahwa regresi linear memberikan interpretasi yang jelas terkait kontribusi masing-masing variabel, meskipun performanya relatif lebih rendah.



Gambar 8. Perbandingan Model

Regresi logistik menunjukkan peningkatan pada klasifikasi keparahan, dengan akurasi yang lebih baik pada kategori menengah. Random forest menghasilkan performa terbaik dengan akurasi dan nilai R-squared tertinggi, menegaskan kemampuannya dalam menangani kompleksitas data kesehatan.

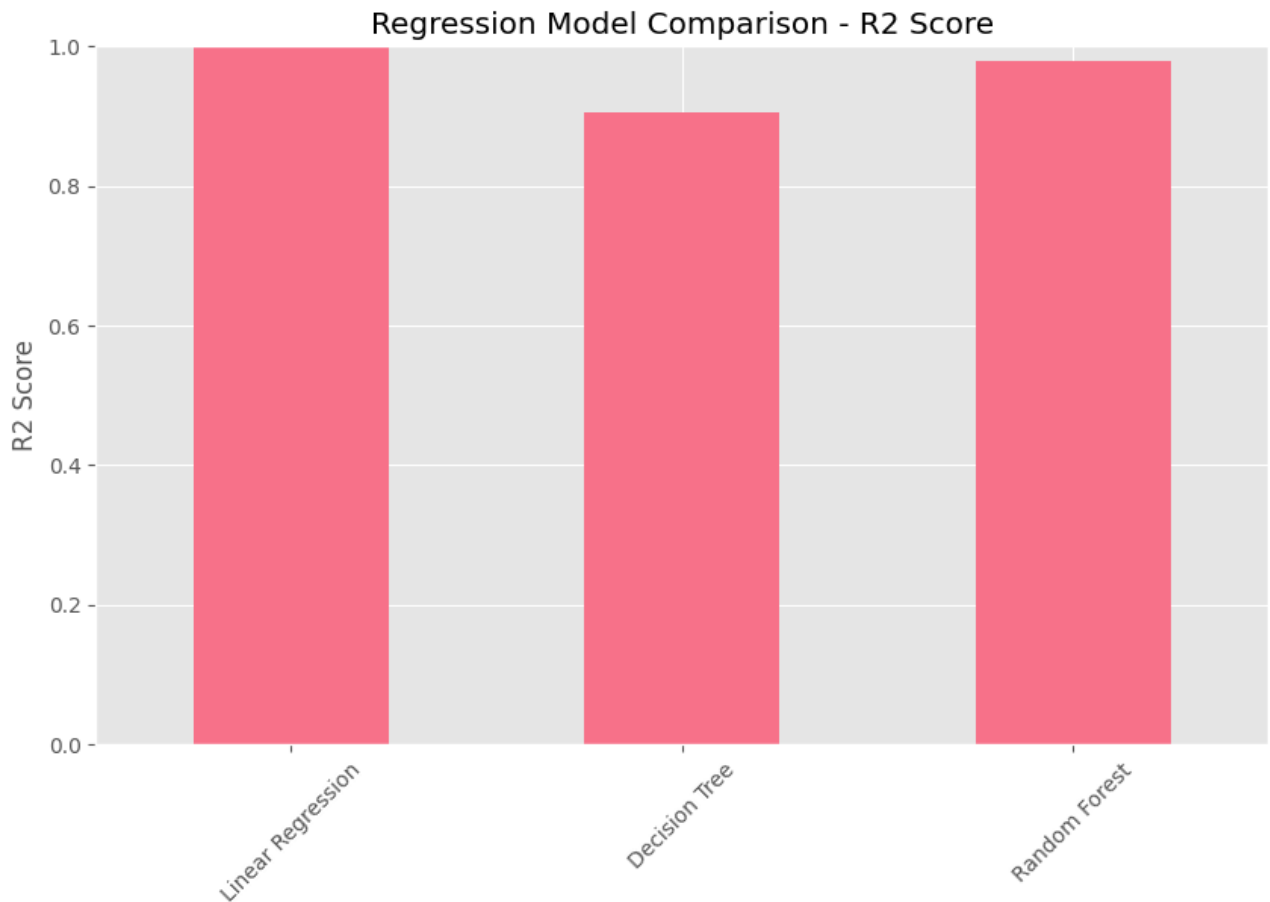
e. Evaluation

Evaluasi model dilakukan menggunakan beberapa metrik. Pada regresi linear, nilai R-squared tercatat sebesar ~0.62 dengan RMSE yang relatif tinggi, menunjukkan adanya variabilitas data yang tidak sepenuhnya dijelaskan oleh model linier.



Gambar 9. Perbandingan Performa Model

Regresi logistik menghasilkan akurasi sekitar 74%, dengan ketepatan lebih baik pada kategori ringan dan sedang dibanding kategori berat. Model random forest memberikan hasil terbaik dengan nilai R-squared sekitar 0.85, serta RMSE yang jauh lebih rendah dibanding model lainnya. Selain itu, random forest juga menunjukkan tingkat akurasi klasifikasi lebih dari 85%, menegaskan keunggulannya dalam prediksi tingkat keparahan kanker.



Gambar 10. Perbandingan Nilai R-squared

Perbandingan ini menunjukkan bahwa meskipun regresi linear dan logistik memiliki keunggulan dalam interpretasi, random forest adalah model yang paling sesuai untuk tujuan prediksi dalam penelitian ini. Model ini mampu menangkap pola kompleks yang terdapat dalam data besar pasien kanker global.

KESIMPULAN

Penelitian ini berhasil memanfaatkan dataset global pasien kanker periode 2015–2024 untuk menganalisis faktor risiko serta memprediksi tingkat keparahan kanker menggunakan pendekatan machine learning. Dengan menerapkan kerangka kerja CRISP-DM, penelitian ini dapat disusun secara sistematis mulai dari pemahaman masalah hingga evaluasi model.

Hasil eksperimen menunjukkan bahwa model regresi linear mampu memberikan interpretasi yang jelas, namun kurang optimal dalam performa prediksi dengan nilai R-squared sekitar 0.62. Regresi logistik memberikan akurasi lebih baik, yaitu sekitar 74%, meskipun masih memiliki keterbatasan dalam memprediksi kategori berat. Model random forest terbukti memberikan hasil paling unggul dengan nilai R-squared sebesar 0.85 dan akurasi klasifikasi lebih dari 85%. Temuan ini menegaskan bahwa random forest merupakan algoritma yang paling sesuai untuk dataset besar dan kompleks seperti data pasien kanker global.

Secara keseluruhan, penelitian ini memberikan kontribusi dalam pengembangan model prediktif yang dapat mendukung pengambilan keputusan di bidang kesehatan. Hasilnya menunjukkan bahwa analisis berbasis data dapat membantu memahami faktor risiko utama kanker dan mendukung perencanaan kebijakan kesehatan yang lebih efektif. Penelitian di masa depan dapat diarahkan pada pengembangan model lebih lanjut dengan teknik ensemble lain atau pendekatan deep learning, serta implementasi tahap deployment untuk integrasi dengan sistem informasi kesehatan.

DAFTAR PUSTAKA

- [1] D. Suryani, P. Hendriyeni, M. Mahardianti, and N. Sujati, “Konsep Dasar Kanker,” 2025, Accessed: Nov. 26, 2025. [Online]. Available: https://books.google.com/books?hl=id&lr=&id=449_EQAAQBAJ&oi=fnd&pg=PA35&dq=penyebarkan+kanker+di+dunia&ots=TKFmMEkkhe&sig=ioKjFX1pBhSbJ2oh-NNGQ6qQkrE
- [2] E. M.-J. A. M. D. L. (JAMALI) and undefined 2021, “Peningkatan Pengetahuan Kanker Payudara dan Ketrampilan Periksa Payudara Sendiri (SADARI) untuk Deteksi Dini Kanker Payudara di Semutan Jatimulyo Dlingo,” *journal.uui.ac.id*, vol. 03, Accessed: Nov. 26, 2025. [Online]. Available: <https://journal.uui.ac.id/JAMALI/article/view/17847>
- [3] R. Balatif, A. S.-S. S. S. Medical, and undefined 2021, “Memahami kaitan gaya hidup dengan kanker: sebagai langkah awal pencegahan kanker,” *talenta.usu.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://talenta.usu.ac.id/scripta/article/view/4506>
- [4] V. N.-G. J. K. dan Kesehatan and undefined 2023, “Kanker Serviks,” *ojs.unimal.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://ojs.unimal.ac.id/galenical/article/view/10134>
- [5] S. H.-J. M. Utama and undefined 2021, “Faktor Risiko Kanker Payudara,” *jurnalmedikahutama.com*, vol. 03, Accessed: Nov. 26, 2025. [Online]. Available: <http://jurnalmedikahutama.com/index.php/JMH/article/view/310>
- [6] U. Khabibah, ... K. A.-F. H., and undefined 2022, “Faktor Risiko Kanker Serviks: Literature Review,” *journal.lppm-stikesfa.ac.id*, vol. 09, no. 3, pp. 270–277, 2022, Accessed: Nov. 26, 2025. [Online]. Available: <https://journal.lppm-stikesfa.ac.id/index.php/FHJ/article/view/354>
- [7] K. Suparna, L. S.-G. Medicina, and undefined 2022, “Kanker payudara: diagnostik, faktor risiko, dan stadium,” *ejournal.undiksha.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://ejournal.undiksha.ac.id/index.php/GM/article/view/47032>
- [8] K. Q.-I. I. J. of Information and undefined 2025, “Klasifikasi Kanker Payudara Menggunakan Data Mining Tool (WEKA),” *ojisnu.isnuponorogo.org*, vol. 3, no. 1, 2025, Accessed: Nov. 26, 2025. [Online]. Available: <https://ojisnu.isnuponorogo.org/index.php/ijitech/article/view/160>
- [9] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, “Klasifikasi jenis kanker kulit menggunakan CNN-SVM,” *jurnal.mdp.ac.id*, vol. 18, no. 1, Accessed: Nov. 26, 2025. [Online]. Available: <https://jurnal.mdp.ac.id/index.php/algorithm/article/view/2363>
- [10] J. Asian, I. S.-F. J. T. Elektro, and undefined 2021, “Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien,” *fidelity.nusaputra.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://fidelity.nusaputra.ac.id/article/view/100>
- [11] R. Tio, P. Sudewo, Y. Pratama, and E. Yanti, “Analisis Data Mining Untuk Prediksi Kanker Payudara Menggunakan Algoritma Klasifikasi,” *jurnal.pustakagalerimandiri.co.id*, 2023, doi: 10.55382/jurnalpustakadata.v3i2.656.
- [12] R. Marzuq, ... S. W.-... T. I. D., and undefined 2023, “Prediksi Kanker Paru-Paru Menggunakan Algoritme Random Forest Decision Tree,” *j-ptiik.ub.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12964>
- [13] H. Usia *et al.*, “Hubungan Usia dan Jenis Kelamin dengan Kejadian Kanker di RS PKU Muhammadiyah Yogyakarta,” *bajangjournal.com*, vol. 1, no. 2, 2021, Accessed: Nov. 26, 2025. [Online]. Available: <https://bajangjournal.com/index.php/JIRK/article/view/97>
- [14] A. Y. Permana, H. N. Fazri, M. Fakhri, N. Athoilah, M. Robi, and R. Firmansyah, “Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest,” *pdfs.semanticscholar.org*, vol. 3, no. 2, 2023, doi: 10.55606/juitik.v3i2.472.
- [15] Y. A’yunan, ... U. I.-J. T. I., and undefined 2023, “Implementasi Data Mining dalam Klasifikasi Diagnosa Kanker Payudara menggunakan Algoritma Logistic Regression,” *jurnal.murnisadar.ac.id*, Accessed: Nov. 26, 2025. [Online]. Available: <https://jurnal.murnisadar.ac.id/index.php/Tekinkom/article/view/948>
- [16] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, “Faktor-faktor yang mempengaruhi perempuan melakukan pemeriksaan dini kanker serviks: A scoping review,” *scholarhub.ui.ac.id*, vol. 18, no. 1, Accessed: Nov. 26, 2025. [Online]. Available: <https://scholarhub.ui.ac.id/epidkes/vol6/iss2/4/>
- [17] R. Rofiani, L. Oktaviani, ... D. V.-J. T., and undefined 2024, “Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4. 5,” *pdfs.semanticscholar.org*, vol.

- 3, no. 1, 2025, Accessed: Nov. 26, 2025. [Online]. Available: <https://pdfs.semanticscholar.org/88b0/b5f22939fe8f30929840448b38075ffcd26e.pdf>
- [18] H. Firda, R. Athiyah, ... S. C.-J. I. J. S., and undefined 2024, “Perbandingan Algoritma Klasifikasi Data Mining Model C4. 5 Dan Naive Bayes Untuk Prediksi Keganasan Kanker Payudara,” *jurnal.umj.ac.id*, vol. 3, no. 1, 2025, Accessed: Nov. 26, 2025. [Online]. Available: <https://jurnal.umj.ac.id/index.php/just-it/article/view/21653>
- [19] V. Plotnikova, M. Dumas, and F. P. Milani, “Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements,” *Data Knowl Eng*, vol. 139, p. 102013, 2022.
- [20] V. Plotnikova, M. Dumas, F. M.-D. & knowledge engineering, and undefined 2022, “Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements,” *Elsevier*, Accessed: Nov. 26, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169023X22000258>