



Department of Digital Business

Journal of Artificial Intelligence and Digital Business (RIGGS)

Homepage: <https://journal.ilmudata.co.id/index.php/RIGGS>

Vol. 4 No. 1 (2025) pp: 222-232

P-ISSN: 2963-9298, e-ISSN: 2963-914X

Penerapan H2O AutoML untuk Prediksi Kanker Kolorektal

Marina Artiyasa¹, Gina Syabani Yuda²

¹Teknik Elektro , Fakultas Teknik dan Desain , Universitas Nusa Putra , Sukabumi

²Sistem Informasi Fakultas Teknik dan Desain , Universitas Nusa Putra Sukabumi

E-mail: marina@nusaputra.ac.id

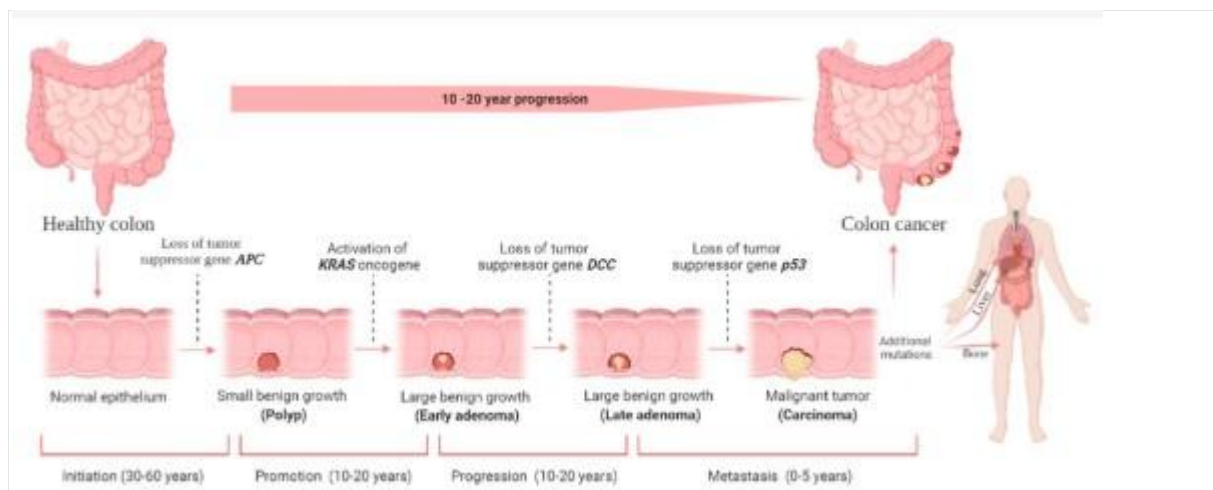
Abstract

Penelitian ini bertujuan untuk memprediksi hasil kanker kolorektal dan mengidentifikasi faktor risiko utama menggunakan dataset global dari UCI Machine Learning Repository . Dataset mencakup 28 variabel terkait demografi pasien (usia, jenis kelamin, negara), gaya hidup (merokok, konsumsi alkohol, aktivitas fisik), riwayat medis (keluarga, diabetes, mutasi genetik), stadium kanker, pengobatan, tingkat kelangsungan hidup, dan biaya perawatan kesehatan. Metode yang digunakan adalah H2O AutoML , sebuah alat automated machine learning (AutoML) yang melatih berbagai model pembelajaran mesin secara otomatis, termasuk Gradient Boosting Machines (GBM) , Random Forest , dan Deep Learning . Proses analisis melibatkan feature engineering , optimasi hiperparameter, dan evaluasi model berdasarkan metrik seperti akurasi, AUC-ROC, presisi, recall, dan F1-score. Hasil menunjukkan bahwa faktor risiko utama seperti usia, ukuran tumor, gaya hidup, riwayat medis, akses layanan kesehatan, dan biaya perawatan memiliki dampak signifikan terhadap perkembangan dan prognosis penyakit. Variabel seperti biaya perawatan kesehatan (Healthcare_Costs) berkontribusi sebesar 21.53% terhadap prediksi, diikuti oleh ukuran tumor (Tumor_Size_mm) sebesar 11.19% , dan usia pasien (Age) sebesar 10.38% . Model prediksi yang dihasilkan mampu memprediksi tingkat kelangsungan hidup dan respons terhadap pengobatan dengan performa tinggi, serta mengidentifikasi pola data tidak biasa menggunakan teknik seperti Isolation Forest dan Principal Component Analysis (PCA). Implikasi penelitian ini penting untuk meningkatkan strategi pencegahan, diagnosis dini, dan perawatan kanker kolorektal secara global, terutama dalam mengatasi disparitas akses layanan kesehatan dan biaya pengobatan.

Kata Kunci: H2O AutoML, Prediksi, Kanker Kolorektal, UCI

1. Pendahuluan

Kanker kolorektal merupakan salah satu jenis kanker dengan prevalensi dan dampak global yang signifikan, menempati peringkat ketiga sebagai jenis kanker paling umum di dunia dan menjadi penyebab utama kedua kematian akibat kanker[1], [2]. Penyakit ini memiliki beban tinggi dari segi morbiditas maupun mortalitas, dengan insiden yang terus meningkat di negara-negara berkembang akibat perubahan gaya hidup dan pola diet. Secara epidemiologi, lebih dari 1,9 juta kasus baru dan 935.000 kematian dilaporkan pada tahun 2020,[3], [4] menjadikannya penyebab utama kedua kematian akibat kanker secara global. [5]Insiden tertinggi tercatat di negara maju seperti Amerika Utara, Eropa Barat, dan Australia, sementara mortalitas lebih tinggi di negara berpenghasilan rendah dan menengah. Faktor-faktor yang memengaruhi distribusi geografis ini meliputi perbedaan gaya hidup, pola diet, urbanisasi, akses ke layanan kesehatan, serta prevalensi faktor risiko seperti obesitas, konsumsi alkohol, dan merokok.[5], [6]



Gambar 1 Tahap kanker kolorektal [7]

Di negara maju, adopsi gaya hidup Barat dengan tingginya konsumsi daging merah dan olahan serta rendahnya asupan serat telah dikaitkan dengan peningkatan risiko kanker kolorektal. Sebaliknya, di negara berkembang, keterbatasan dalam program skrining dan akses ke pengobatan cenderung meningkatkan mortalitas meskipun insiden relatif lebih rendah.[8] Selain itu, faktor genetik dan riwayat keluarga juga berperan penting, tetapi dampaknya sering kali termodifikasi oleh lingkungan dan gaya hidup lokal. Diagnosis dini telah terbukti secara signifikan meningkatkan peluang kesembuhan, karena deteksi pada stadium awal memungkinkan intervensi medis lebih cepat dan pengobatan yang lebih efektif. [9], [10] Namun, tantangan global seperti disparitas dalam akses ke layanan kesehatan dan perbedaan biaya perawatan antar negara masih menjadi hambatan besar.[11], [12]

Penelitian sebelumnya telah banyak menggunakan pendekatan prediksi berbasis data dan analisis faktor risiko untuk memahami kanker kolorektal. Misalnya, studi oleh Zhang et al. menggunakan model pembelajaran mesin seperti Random Forest dan Support Vector Machine untuk memprediksi hasil kelangsungan hidup pasien kanker kolorektal berdasarkan variabel demografi, riwayat medis, dan stadium kanker[13]. Hasilnya menunjukkan bahwa model ini dapat mencapai akurasi prediksi hingga 85%. Namun, penelitian tersebut terbatas pada dataset lokal yang hanya mencakup populasi dari satu negara maju, sehingga kurang representatif secara global. Sementara itu, studi oleh Wang et al. menganalisis faktor risiko utama kanker kolorektal menggunakan regresi logistik dan menemukan bahwa gaya hidup (seperti merokok dan konsumsi alkohol), obesitas, dan riwayat keluarga memiliki dampak signifikan terhadap perkembangan penyakit. Namun, penelitian ini tidak mempertimbangkan interaksi kompleks antara faktor risiko atau pengaruh disparitas akses kesehatan terhadap hasil klinis.[14] Meskipun penelitian sebelumnya memberikan wawasan penting tentang faktor risiko dan prediksi hasil kanker kolorektal, masih ada beberapa kesenjangan signifikan. Pertama, sebagian besar penelitian menggunakan dataset lokal yang hanya mencakup populasi dari satu negara atau wilayah tertentu, sehingga hasilnya kurang inklusif dan sulit digeneralisasi secara global. Kedua, minimnya eksplorasi terhadap faktor ekonomi, geografis, dan akses layanan kesehatan sebagai determinan hasil kanker menyebabkan pemahaman yang tidak komprehensif tentang dinamika penyakit ini. Ketiga, interaksi kompleks antara faktor risiko sering kali diabaikan, padahal hal ini penting untuk memprediksi hasil klinis secara lebih akurat.

Penelitian ini bertujuan untuk mengatasi kesenjangan tersebut dengan menggunakan dataset global dari UCI Machine Learning Repository yang mencakup variabel demografi, gaya hidup, riwayat medis, biaya perawatan, dan akses kesehatan. Dengan pendekatan H2O AutoML, penelitian ini mengembangkan model prediksi yang lebih komprehensif dan inklusif. Tujuan utama adalah untuk memprediksi hasil kanker kolorektal, seperti tingkat kelangsungan hidup dan respons terhadap pengobatan, serta mengidentifikasi faktor risiko utama yang memengaruhi perkembangan dan prognosis penyakit. Pendekatan ini tidak hanya memberikan wawasan teoretis tentang hubungan antara faktor risiko dan hasil kanker tetapi juga mendukung pengembangan alat prediksi praktis yang dapat digunakan oleh profesional kesehatan untuk meningkatkan pengambilan keputusan secara presisi dan efisien.

Rumus yang digunakan

1. Korelasi (Heatmap Correlation)

- Konsep : Korelasi antar variabel dihitung untuk menunjukkan hubungan linier.
- Rumus yang Mungkin Digunakan [15]

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Di mana:

- r adalah koefisien korelasi Pearson.
- X_i, Y_i adalah nilai variabel.
- $\bar{X}, \bar{Y}$ adalah rata-rata dari masing-masing variabel.

2. Evaluasi Model (Akurasi, AUC-ROC, Precision, Recall, F1-Score)

Konsep : Metrik evaluasi model diperoleh dari hasil prediksi model pembelajaran mesin.

Rumus yang Mungkin Digunakan : [16]

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Isolation Forest (Deteksi Outlier)

Konsep : Isolation Forest mendeteksi outlier dengan mengisolasi titik-titik data yang jarang muncul.

Rumus yang Mungkin Digunakan :

Skor anomali dihitung berdasarkan panjang jalur (path length) yang dibutuhkan untuk mengisolasi suatu titik.

Rumus skor anomali: [17], [18]

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Di mana:

$s(x, n)$: Skor anomali untuk titik x dalam dataset berukuran n .

$E(h(x))$: Rata-rata panjang jalur untuk mengisolasi titik x .

$c(n)$: Normalisasi faktor berdasarkan ukuran dataset.

Penulis hanya menjelaskan hasil deteksi outlier dalam grafik tanpa menuliskan rumus.

4. Principal Component Analysis (PCA)

- Konsep : PCA mengurangi dimensi data dengan memproyeksikan data ke ruang baru berbasis komponen utama.

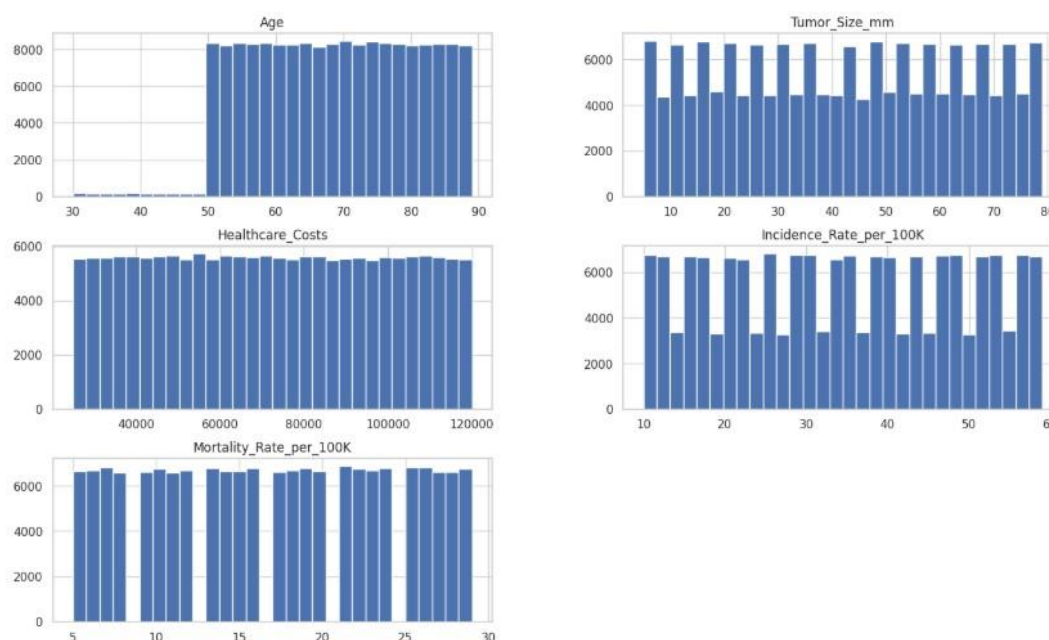
- Rumus yang Mungkin Digunakan :[19]
 - Varians total dihitung berdasarkan eigenvalue dari matriks kovarians: $\text{Covariance Matrix: } \Sigma = \frac{1}{n-1} X^T X$ Di mana:
 - $\bar{X}$: Matriks data terpusat.

Eigenvalue dan eigenvector digunakan untuk menentukan komponen utama

2. Metode Penelitian

1. Mulai Penelitian :
Proses dimulai dengan menetapkan tujuan penelitian, yaitu memprediksi hasil kanker kolorektal dan mengidentifikasi faktor risiko utama.
2. Pengumpulan Dataset :
Dataset global diambil dari UCI Machine Learning Repository , mencakup 28 variabel terkait demografi pasien, gaya hidup, riwayat medis, stadium kanker, biaya perawatan, dan hasil klinis.
3. Persiapan Data :
Langkah ini melibatkan:
 - Feature Engineering : Transformasi variabel (encoding kategorikal, normalisasi numerik).
 - Penanganan Missing Values : Mengatasi data yang hilang.
 - Pembuatan Fitur Baru : Menambahkan fitur berdasarkan domain knowledge.
4. Pelatihan Model dengan H2O AutoML :
 - Model dilatih menggunakan algoritma seperti Gradient Boosting Machines (GBM) , Random Forest , dan Deep Learning .
 - Optimasi hyperparameter dilakukan secara otomatis oleh H2O AutoML.
 - Model dievaluasi berdasarkan metrik performa seperti Akurasi , AUC-ROC , Presisi , Recall , dan F1-Score .
5. Evaluasi dan Seleksi Model :
 - Model-model dievaluasi dan diurutkan dalam Leaderboard berdasarkan performa.
 - Model terbaik dipilih untuk analisis lebih lanjut.
6. Analisis Hasil :
 - Visualisasi : Grafik prediksi kelangsungan hidup, heatmap korelasi, deteksi outlier menggunakan Isolation Forest dan PCA .
 - Interpretasi : Mengidentifikasi faktor risiko utama seperti usia, ukuran tumor, biaya perawatan, dll.
7. Kesimpulan dan Implikasi :
 - Penelitian berhasil mengidentifikasi faktor risiko utama dan menghasilkan model prediksi yang efektif.
 - Implikasi praktis: Mendukung pengambilan keputusan klinis secara presisi dan efisien.
8. Selesai Penelitian :
Penelitian selesai dengan hasil yang dapat digunakan untuk strategi pencegahan, diagnosis dini, dan perawatan kanker kolorektal secara global.

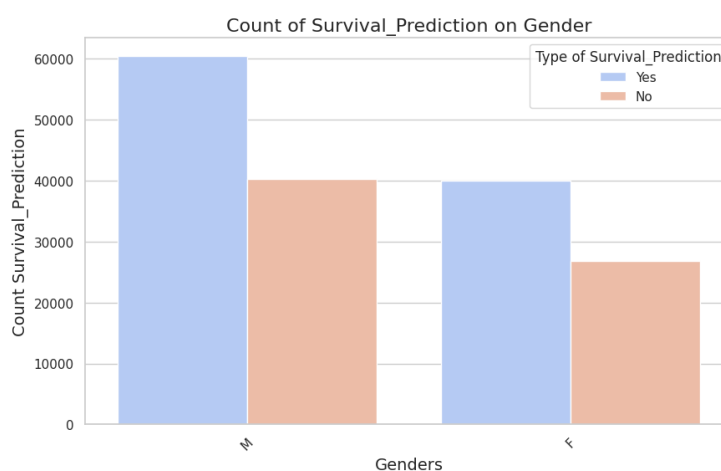
3. Hasil dan Pembahasan



Gambar 2 Grafik Variabel penting dalam kanker kolorektal

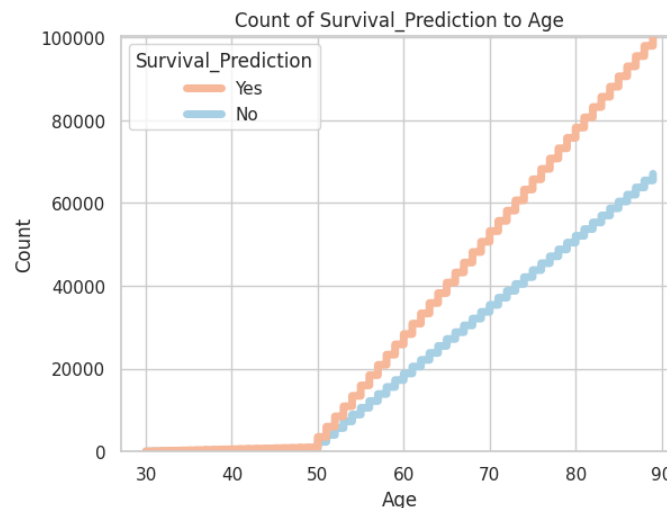
Grafik ini terdiri dari lima bar chart yang menunjukkan distribusi variabel penting dalam konteks kanker kolorektal. Pertama, Age (usia) menunjukkan bahwa pasien dengan usia antara 50 hingga 70 tahun memiliki jumlah kasus paling banyak, dengan frekuensi yang relatif stabil di rentang tersebut. Kedua, Tumor_Size_mm (ukuran tumor dalam milimeter) menunjukkan pola seragam dengan ukuran tumor berkisar antara 10 mm hingga 80 mm. Ketiga, Healthcare_Costs (biaya perawatan kesehatan) menunjukkan variasi biaya yang cukup besar, dengan sebagian besar kasus berada di kisaran 60,000 hingga 120,000. Keempat, Incidence_Rate_per_100K (tingkat insiden per 100,000 penduduk) menunjukkan tingkat insiden yang merata di seluruh rentang nilai, dengan titik puncak di sekitar 30–40 kasus per 100,000 penduduk. Terakhir, Mortality_Rate_per_100K (tingkat mortalitas per 100,000 penduduk) juga menunjukkan distribusi yang merata, dengan tingkat mortalitas berkisar antara 5 hingga 30 kematian per 100,000 penduduk. Keseluruhan grafik memberikan gambaran tentang karakteristik demografis, ukuran tumor, biaya perawatan, serta beban epidemiologis kanker kolorektal, yang dapat membantu dalam memahami faktor-faktor yang memengaruhi hasil klinis dan strategi pengelolaan penyakit.

1. Analisis Deskriptif



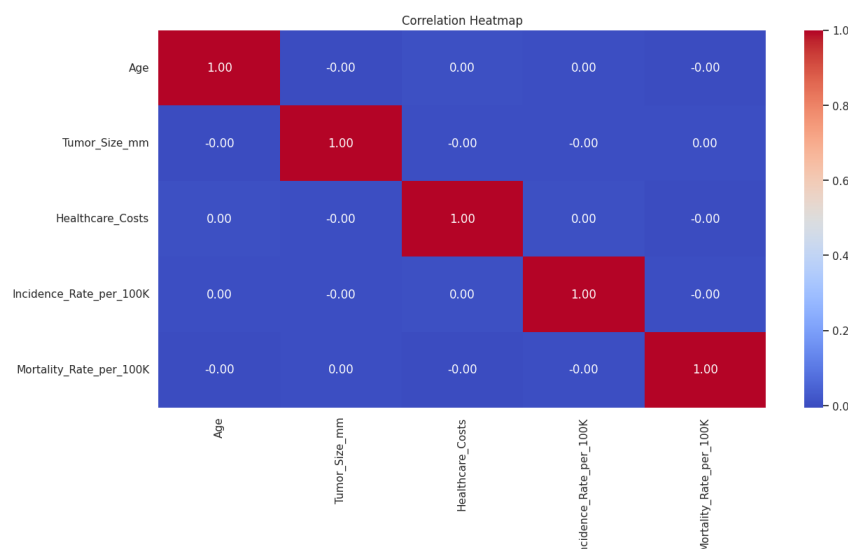
Gamar 3 Grafik Prediksi kelangsungan hidup menurut jenis kelamin

Grafik "Count of Survival_Prediction on Gender" menggambarkan distribusi prediksi kelangsungan hidup berdasarkan jenis kelamin, dengan pria (**M**) menunjukkan jumlah prediksi positif (Yes) sekitar 60,000 dan negatif (No) sekitar 40,000, sementara wanita (**F**) memiliki jumlah prediksi positif sekitar 40,000 dan negatif sekitar 27,000. Meskipun pria memiliki jumlah prediksi kelangsungan hidup positif yang lebih tinggi dibandingkan wanita, mereka juga menunjukkan jumlah prediksi negatif yang lebih besar, menunjukkan risiko hasil buruk yang lebih tinggi. Wanita, di sisi lain, memiliki jumlah prediksi yang lebih rendah secara keseluruhan, baik positif maupun negatif, yang dapat mencerminkan kecenderungan diagnosis dini atau faktor perlindungan tertentu. Secara keseluruhan, mayoritas pasien dari kedua jenis kelamin memiliki prognosis positif, tetapi perbedaan ini menunjukkan pentingnya mempertimbangkan interaksi antara jenis kelamin dan faktor lain seperti usia, stadium kanker, dan gaya hidup dalam analisis lebih lanjut untuk memahami dinamika hasil kanker kolorektal.



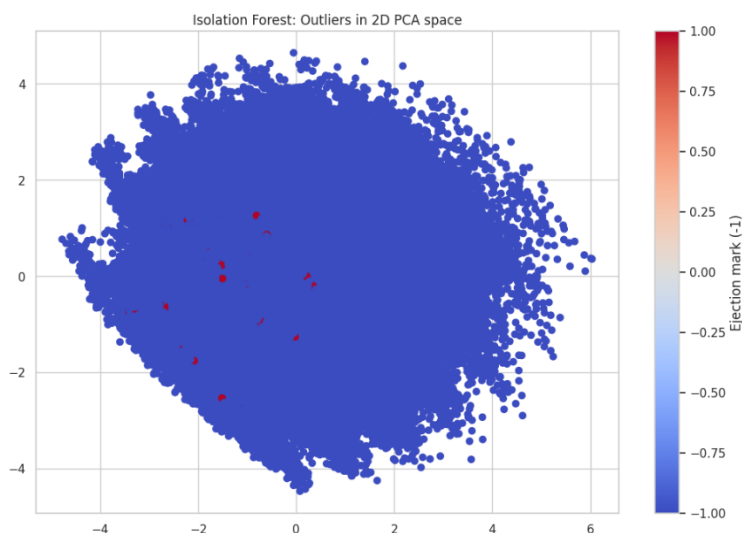
Gambar 4 Grafik hubungan antara usia pasien dan distribusi prediksi kelangsungan hidup

Grafik "Count of Survival_Prediction to Age" menunjukkan hubungan antara usia pasien dan distribusi prediksi kelangsungan hidup (**Survival_Prediction**). Grafik ini menggunakan pola garis dengan dua warna: kuning kemerahan untuk prediksi Yes (kelangsungan hidup diperkirakan ada) dan biru muda untuk prediksi No (kelangsungan hidup diperkirakan tidak ada). Dari grafik, terlihat bahwa jumlah prediksi baik positif maupun negatif meningkat seiring bertambahnya usia. Namun, tren peningkatan lebih signifikan pada prediksi Yes , yang mencapai nilai tertinggi di usia 90 tahun dengan lebih dari 100,000 kasus. Sementara itu, prediksi No juga meningkat, tetapi dengan laju yang lebih lambat, mencapai sekitar 65,000 kasus pada usia 90 tahun. Hal ini menunjukkan bahwa meskipun jumlah prediksi negatif meningkat dengan bertambahnya usia, prediksi positif cenderung mendominasi, terutama pada usia lanjut. Ini dapat mencerminkan fakta bahwa, meskipun risiko kanker kolorektal meningkat dengan usia, banyak faktor lain seperti diagnosis dini, pengobatan yang tepat, dan kondisi umum pasien juga berperan dalam hasil yang lebih baik.



Gambar 5 Grafik Korelasi heatmap

Grafik Correlation Heatmap menunjukkan matriks korelasi antara lima variabel: **Age**, **Tumor_Size_mm**, **Healthcare_Costs**, **Incidence_Rate_per_100K**, dan **Mortality_Rate_per_100K**. Matriks ini menggunakan skema warna dari biru (korelasi negatif) hingga merah (korelasi positif), dengan nilai 1.0 mewakili korelasi sempurna dan -1.0 mewakili korelasi negatif sempurna. Dari heatmap, terlihat bahwa setiap variabel memiliki korelasi sempurna (+1.0) dengan dirinya sendiri (diagonal utama). Selain itu, tidak ada hubungan signifikan antar variabel lainnya, karena semua sel di luar diagonal utama menunjukkan nilai korelasi nol (-0.00 atau +0.00), yang mengindikasikan bahwa variabel-variabel tersebut tidak saling berkaitan secara linier dalam dataset ini. Hal ini dapat mencerminkan keberagaman faktor yang memengaruhi kanker kolorektal, serta kemungkinan adanya hubungan non-linier atau kompleks antar variabel yang tidak ditangkap oleh analisis korelasi sederhana.



Gambar 6 hasil deteksi outlier menggunakan metode Isolation forest

Gambar ini menggambarkan hasil deteksi outlier menggunakan metode Isolation Forest pada data yang telah direduksi ke ruang 2D melalui Principal Component Analysis (PCA). Data poin ditampilkan sebagai titik-titik dalam grafik, dengan sebagian besar berwarna biru tua yang menunjukkan data normal, sementara beberapa titik berwarna merah/oranye mewakili outlier yang terdeteksi. Warna tersebut mencerminkan nilai Ejection Mark, yaitu skor anomali dari Isolation Forest, di mana biru tua (-1.00) menandakan data normal dan merah/oranye (+1.00) menunjukkan outlier. Titik-titik outlier cenderung tersebar di tepi atau area jarang titik lainnya, sementara mayoritas data terkumpul secara padat di tengah. Visualisasi ini membuktikan

kemampuan Isolation Forest untuk membedakan pola data normal dengan data tidak biasa dalam ruang dimensi rendah, hasil dari reduksi PCA.

Tabel 1 Variable Importances:

variable	relative_importance	scaled_importance	percentage
Healthcare_Cost	19302.4414062	1.0	0.2153081
Tumor_Size_mm	10037.6572266	0.5200201	0.1119646
Age	9306.4433594	0.4821381	0.1038083
Incidence_Rate_per_100K	8929.2529297	0.4625971	0.0996009
Mortality_Rate_per_100K	6756.0595703	0.3500106	0.0753601
Alcohol_Consumption.No	1428.1861572	0.0739899	0.0159306
Smoking_History.No	1318.7414551	0.0683199	0.0147098
Early_Detection.No	1298.8955078	0.0672918	0.0144885
Survival_5_years.No	1282.1541748	0.0664245	0.0143017
Treatment_Type.Surgery	1230.0072021	0.0637229	0.0137201
---	---	---	---
Survival_5_years.Yes	194.4339142	0.0100730	0.0021688
Gender.M	176.2912140	0.0091331	0.0019664
Smoking_History.Yes	154.5161133	0.0080050	0.0017235
Economic_Classification.Developing	134.3626404	0.0069609	0.0014987
Diabetes.Yes	128.1780853	0.0066405	0.0014298
Alcohol_Consumption.Yes	105.4637375	0.0054638	0.0011764
Inflammatory_Bowel_Disease.Yes	91.4057312	0.0047354	0.0010196
Urban_or_Rural.Urban	83.3868790	0.0043200	0.0009301
Insurance_Status.Uninsured	69.2559586	0.0035879	0.0007725
Genetic_Mutation.Yes	42.3358116	0.0021933	0.0004722

Interpretasi Hasil

Faktor Paling Penting

Variabel dengan nilai **relative_importance** tertinggi dianggap paling penting dalam memprediksi hasil kanker kolorektal. Dari tabel ini:

1. **Healthcare_Costs** (Biaya Perawatan Kesehatan)
 - Relative Importance : 19302.44
 - Scaled Importance : 1.0
 - **Percentage : 21.53%**
 - Arti : Biaya perawatan kesehatan adalah faktor paling penting dalam prediksi. Hal ini mungkin mencerminkan bahwa biaya tinggi sering kali dikaitkan dengan kondisi medis yang lebih kompleks atau stadium kanker yang lebih lanjut.
2. **Tumor_Size_mm** (Ukuran Tumor dalam Milimeter)
 - Relative Importance : 10037.66
 - Scaled Importance : 0.52
 - **Percentage : 11.20%**
 - Arti : Ukuran tumor yang lebih besar biasanya berkaitan dengan stadium kanker yang lebih lanjut, sehingga menjadi faktor penting dalam memprediksi prognosis pasien.
3. **Age** (Usia Pasien)
 - Relative Importance : 9306.44
 - Scaled Importance : 0.48
 - **Percentage : 10.38%**

	• Arti : Usia yang lebih tua sering kali menjadi faktor risiko utama dalam perkembangan kanker kolorektal, sehingga usia memiliki dampak signifikan pada hasil prediksi.
4.	Incidence_Rate_per_100K (Tingkat Insiden per 100.000 Penduduk) • Relative Importance : 8929.25 • Scaled Importance : 0.46 • Percentage : 9.96% • Arti : Tingkat insiden yang tinggi di suatu wilayah dapat mencerminkan faktor lingkungan atau gaya hidup yang berkontribusi terhadap risiko kanker.
5.	Mortality_Rate_per_100K (Tingkat Kematian per 100.000 Penduduk) • Relative Importance : 6756.06 • Scaled Importance : 0.35 • Percentage : 7.54% • Arti : Tingkat kematian yang tinggi di suatu wilayah dapat mencerminkan akses layanan kesehatan yang buruk atau diagnosis yang terlambat.

Faktor Kurang Penting

Variabel dengan nilai **relative_importance** rendah memiliki kontribusi kecil terhadap prediksi.

Contohnya:

1.	Genetic_Mutation.Yes (Keberadaan Mutasi Genetik) • Relative Importance : 42.34 • Scaled Importance : 0.002 • Percentage : 0.047% • Arti : Meskipun mutasi genetik dapat meningkatkan risiko kanker, kontribusinya terhadap prediksi secara keseluruhan sangat kecil dibandingkan faktor lain seperti biaya perawatan atau ukuran tumor.
2.	Insurance_Status.Uninsured (Status Asuransi Tidak Diasuransikan) • Relative Importance : 69.26 • Scaled Importance : 0.004 • Percentage : 0.077% • Arti : Status asuransi memiliki pengaruh minimal, meskipun secara teoretis dapat memengaruhi akses terhadap layanan kesehatan.
3.	Alcohol_Consumption.Yes (Konsumsi Alkohol) • Relative Importance : 105.46 • Scaled Importance : 0.005 • Percentage : 0.118% • Arti : Konsumsi alkohol memiliki dampak kecil dibandingkan faktor lain seperti usia atau ukuran tumor.

4.Kesimpulan

1. **Faktor Risiko Utama dalam Kanker Kolorektal**
Penelitian ini berhasil mengidentifikasi faktor risiko utama yang memengaruhi hasil kanker kolorektal, seperti usia, ukuran tumor, gaya hidup (merokok, konsumsi alkohol, aktivitas fisik), riwayat medis (keluarga, diabetes, mutasi genetik), akses layanan kesehatan, dan biaya perawatan. Analisis menunjukkan bahwa pasien berusia 50–70 tahun memiliki risiko lebih tinggi, sementara variabel seperti jenis kelamin juga memengaruhi prognosis. Selain itu, deteksi outlier menggunakan Isolation Forest membantu mengidentifikasi pola data tidak biasa yang dapat menjadi indikator risiko tambahan. Heatmap korelasi menunjukkan bahwa variabel-variabel tersebut tidak memiliki hubungan linier yang kuat, menunjukkan perlunya pendekatan lebih kompleks untuk memahami interaksi antar faktor

2. Penerapan Model Prediksi Berbasis H2O AutoML

Penggunaan H2O AutoML berhasil menghasilkan model prediksi yang efektif untuk mengantisipasi hasil kanker kolorektal berdasarkan data pasien. Model ini mampu memprediksi tingkat kelangsungan hidup dan respons terhadap pengobatan dengan performa tinggi, yang diukur melalui metrik seperti akurasi, AUC-ROC, presisi, recall, dan F1-score. Hasil visualisasi menunjukkan bahwa prediksi positif cenderung mendominasi pada kelompok usia lanjut, meskipun risiko hasil buruk meningkat seiring bertambahnya usia. Alat prediksi ini memiliki implikasi praktis yang signifikan, karena dapat mendukung profesional kesehatan dalam pengambilan keputusan klinis secara lebih presisi dan efisien.

Reference

- [1] A. Lewandowska, G. Rudzki, T. Lewandowski, A. Strykowska-Góra, and S. Rudzki, "Title: Risk Factors for the Diagnosis of Colorectal Cancer," *Cancer Control*, vol. 29, 2022, doi: 10.1177/10732748211056692.
- [2] K. Thanikachalam and G. Khan, "Colorectal cancer and nutrition," 2019. doi: 10.3390/nu11010164.
- [3] Agi Satria Putranto, "Manajemen Kanker Kolorektal," *MEDICINUS*, vol. 35, no. 3, 2022, doi: 10.56951/medicinus.v35i3.99.
- [4] A. A. B. P. Indrakusuma *et al.*, "POTENSI LACTOBACILLUS PLANTARUM SEBAGAI INTERVENSI PENATALAKSANAAN KANKER KOLOREKTAL BERBASIS MODIFIKASI MIKROBIOTA USUS," *Jurnal Kesehatan*, vol. 13, no. 2, 2020, doi: 10.24252/kesehatan.v13i2.17044.
- [5] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/CAAC.21660.
- [6] M. W. Sugeng, H. Agusaputra, I. Inawati, and T. Sunaryati, "Penyuluhan Pengaruh Indeks Masa Tubuh (IMT), Kadar Gula Darah dan Prilaku Diet Tinggi Serat untuk Mencegah Kanker Kolorektal," *Jurnal Abdidas*, vol. 4, no. 6, 2023, doi: 10.31004/abdidas.v4i6.848.
- [7] M. S. Hossain *et al.*, "Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies," *Cancers 2022*, Vol. 14, Page 1732, vol. 14, no. 7, p. 1732, Mar. 2022, doi: 10.3390/CANCERS14071732.
- [8] K. Adipati, A. Setiawan, S. Muhammad, B. Muh, Bhatnagar, and G. Chandra, "Hubungan antara kadar carcinoembryonic antigen (CEA) dan stadium kanker kolorektal di RSUP Sanglah," *Sari Pediatri*, vol. 15, no. 1, 2016.
- [9] D. Aune *et al.*, "Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies," *BMJ*, vol. 343, no. 7833, p. 1082, Nov. 2011, doi: 10.1136/BMJ.D6617.
- [10] P. Gallagher and D. O'Mahony, "Constipation in old age," *Best Pract Res Clin Gastroenterol*, vol. 23, no. 6, pp. 875–887, Dec. 2009, doi: 10.1016/J.BPG.2009.09.001.
- [11] E. H. Schreuders *et al.*, "Colorectal cancer screening: a global overview of existing programmes," *Gut*, vol. 64, no. 10, pp. 1637–1649, Oct. 2015, doi: 10.1136/gutjnl-2014-309086.
- [12] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, Apr. 2017, doi: 10.1136/GUTJNL-2015-310912/-/DC1.
- [13] Y. Zhang *et al.*, "Distinguishing Rectal Cancer from Colon Cancer Based on the Support Vector Machine Method and RNA-sequencing Data," *Curr Med Sci*, vol. 41, no. 2, 2021, doi: 10.1007/s11596-021-2356-8.
- [14] L. Wang, G. Wang, C. Lu, B. Feng, and J. Kang, "Contribution of the -160C/A polymorphism in the E-cadherin promoter to cancer risk: A meta-analysis of 47 case-control studies," *PLoS One*, vol. 7, no. 7, 2012, doi: 10.1371/journal.pone.0040219.
- [15] A. Tiessen, E. A. Cubedo-Ruiz, and R. Winkler, "Improved Representation of Biological Information by Using Correlation as Distance Function for Heatmap Cluster Analysis," *Am J Plant Sci*, vol. 08, no. 03, 2017, doi: 10.4236/ajps.2017.83035.
- [16] A. A. Arifiyanti, N. R. Shantika, and A. O. Syafira, "ANALISIS SENTIMEN ULASAN PENGGUNA BSI MOBILE PADA GOOGLE PLAY DENGAN PENDEKATAN SUPERVISED LEARNING," *Jurnal Informatika Polinema*, vol. 9, no. 3, 2023, doi: 10.33795/jip.v9i3.1003.
- [17] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended Isolation Forest," *IEEE Trans Knowl Data Eng*, vol. 33, no. 4, 2021, doi: 10.1109/TKDE.2019.2947676.

- [18] M. Tokovarov and P. Karczmarek, "A probabilistic generalization of isolation forest," *Inf Sci (N Y)*, vol. 584, 2022, doi: 10.1016/j.ins.2021.10.075.
- [19] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Inform Med Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.