Analisis Prediktif dalam Klasifikasi Keganasan Kanker Payudara

Yogyakarta, 3 Juli 2024

Disusun Oleh:

1. Cornelia Happy Rahmawati (22/496781/PA/21362)
2. Zumrotul Inayah (22/498498/PA/21524)
3. Farhan Ubaidillah (22/503097/PA/21604)

Dosen Pengampu:

1. Drs. Danardono, MPH., Ph.D.
2. Mohamad Fahruli Wahyujati, S.Si., M.Si.

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS GADJAH MADA

2024

Abstrak

Penelitian ini bertujuan untuk mengevaluasi kinerja beberapa algoritma machine learning dalam mengklasifikasikan keganasan kanker payudara menggunakan dataset dari Kaggle yang terdiri dari 569 pasien dengan 32 variabel. Algoritma yang digunakan meliputi Logistic Regression, Decision Tree, Random Forest, dan K-Nearest Neighbors (KNN). Data preprocessing dilakukan dengan menghapus variabel yang tidak relevan dan menangani missing values. Feature selection menggunakan metode Random Forest menghasilkan empat fitur terpenting: radius_mean, texture_mean, perimeter_mean, dan area_mean.

Hasil penelitian menunjukkan bahwa algoritma Decision Tree memiliki akurasi tertinggi sebesar 99,415% pada data uji, diikuti oleh Logistic Regression dan KNN dengan akurasi masing-masing sebesar 93,567%. Random Forest menunjukkan akurasi sebesar 92,982%. Confusion matrix digunakan untuk mengevaluasi kinerja model, dengan Decision Tree menunjukkan nilai True Positive (TP) dan True Negative (TN) yang lebih tinggi dibandingkan algoritma lainnya.

Kesimpulan dari penelitian ini adalah bahwa Decision Tree merupakan algoritma terbaik untuk klasifikasi keganasan kanker payudara dalam penelitian ini. Untuk meningkatkan akurasi dan efektivitas model, diperlukan studi lanjutan dengan dataset yang lebih besar dan beragam serta personalisasi model untuk setiap pasien. Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam diagnosis kanker payudara melalui pemanfaatan teknologi machine learning.

A. Pendahuluan

1. Latar belakang

Kanker payudara merupakan salah satu jenis kanker yang paling umum dan mematikan di dunia, terutama di kalangan perempuan. Menurut Organisasi Kesehatan Dunia (WHO), jutaan kasus baru didiagnosis setiap tahun, dan kanker payudara menjadi penyebab utama kematian terkait kanker di banyak negara (WHO, 2021). Deteksi dini dan diagnosis yang akurat sangat penting untuk meningkatkan peluang kesembuhan dan mengurangi angka kematian. Dalam konteks ini, penerapan teknologi machine learning menawarkan potensi besar untuk meningkatkan akurasi diagnosis kanker payudara melalui analisis data medis secara otomatis dan efisien.

Teknologi machine learning telah digunakan dalam berbagai bidang medis untuk memprediksi dan mendiagnosis penyakit dengan akurasi tinggi (Topol, 2019). Dengan memanfaatkan algoritma machine learning, kita dapat mengolah data pasien yang kompleks untuk mengenali pola-pola yang mungkin terlewatkan oleh metode konvensional. Studi ini bertujuan untuk mengeksplorasi penggunaan berbagai algoritma machine learning dalam mengklasifikasikan keganasan kanker payudara berdasarkan data medis pasien.

2. Tujuan penelitian

Penelitian ini memiliki beberapa tujuan utama:

a. Mengevaluasi kinerja beberapa algoritma machine learning dalam mengklasifikasikan keganasan kanker payudara.
b. Mengidentifikasi fitur-fitur penting yang berkontribusi signifikan dalam prediksi kanker payudara.
c. Membandingkan efektivitas model-model machine learning yang digunakan untuk menemukan model terbaik dalam klasifikasi kanker payudara.

3. Manfaat penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, antara lain:

a. Memberikan wawasan tentang kinerja berbagai algoritma machine learning dalam diagnosis kanker payudara.
b. Menyediakan model prediktif yang akurat dan dapat diandalkan untuk membantu tenaga medis dalam membuat keputusan klinis.
c. Meningkatkan efisiensi dan efektivitas dalam proses diagnosis kanker payudara, yang pada akhirnya dapat meningkatkan kualitas perawatan pasien dan menurunkan angka kematian.

B. Tinjauan Pustaka

Pada penelitian ini akan memprediksi hasil klasifikasi dari empat algoritma yang kemudian akan dibandingkan hasilnya untuk menentukan algoritma yang paling baik untuk pengujian dataset.

1. Regresi Logistik

   Logistic Regression adalah algoritma statistik yang digunakan untuk prediksi klasifikasi. Algoritma ini bekerja dengan mengukur hubungan antara variabel independen dan variabel dependen yang bersifat biner, seperti diagnosis kanker payudara (jinak atau ganas). Logistic Regression mengestimasi probabilitas bahwa suatu kejadian akan terjadi dengan menggunakan fungsi logit (Dritsas & Trigka, 2022).

2. Decision Tree

   Decision Tree adalah algoritma yang menggunakan model pohon keputusan untuk membuat prediksi berdasarkan serangkaian aturan keputusan yang diturunkan dari data. Algoritma ini sangat efektif dalam menangani data yang kompleks dan non-linear. Pada setiap node, data dibagi berdasarkan fitur yang memberikan peningkatan terbesar dalam pemisahan antara kelas yang berbeda (Hidayah, Arifitama, & Permana, 2024).

3. Random Forest

   Random Forest adalah metode ensemble learning yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi overfitting. Setiap pohon dalam Random Forest dilatih pada subset acak dari data, dan hasil prediksi akhir diambil berdasarkan mayoritas voting dari semua pohon (Akbar, 2020).

4. K-Nearest Neighbors (KNN)

   K-Nearest Neighbors adalah algoritma non-parametrik yang digunakan untuk klasifikasi dan regresi. Algoritma ini mengklasifikasikan sampel baru berdasarkan mayoritas kelas dari K tetangga terdekat di ruang fitur. KNN sangat sederhana namun efektif untuk berbagai jenis masalah klasifikasi, termasuk deteksi kanker payudara (Meyana, Sitorus, & Soleh, 2023).

5. Naive Bayes

   Naive Bayes adalah algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes. Algoritma ini mengasumsikan bahwa setiap fitur bersifat independen terhadap fitur lainnya, yang menjadikannya sangat cepat dan efisien untuk masalah klasifikasi teks dan data medis (Supriyadi et al., 2020).

C. Metode Penelitian

1. Data

   Penelitian ini menggunakan dataset pasien kanker payudara yang didapatkan dari *website* Kaggle. Data terdiri dari 569 pasien dengan 32 variabel. Data tersebut memiliki atribut ID, beberapa tipe uji lab, dan target sebagai indikasi apakah pasien menderita kanker payudara yang ganas atau tidak.

2. Preprocessing Data

   Pada dataset terdapat variabel "ID" yang tidak akan digunakan dalam analisis sehingga variabel tersebut dihapus. Ditemukan *missing value* pada variabel "Unnamed : 32" sebanyak 569 baris atau seluruh baris tidak memiliki nilai sehingga variabel tersebut juga dihapus.

   Dari eksplorasi data, didapatkan data terdiri dari 357 pasien dengan label "benign" atau jinak dan 212 pasien dengan tipe kanker "malignant" atau ganas. Sebelum dilakukan analisis, dikarenakan nilai dari fitur sangat beragam maka dilakukan normalisasi data dan pemilihan fitur utama.

3. Pemilihan Fitur

   Pemilihan fitur (*feature selection*) merupakan langkah yang penting dalam analisis data karena dapat meningkatkan kinerja model machine learning, meningkatkan akurasi model, mengurangi overfitting, mengatasi *curse of dimensionality* yang biasa terjadi pada data dengan dimensi tinggi, dan mengurangi redudansi karena beberapa fitur mungkin memiliki informasi yang sama atau mirip.

   Data yang digunakan dalam penelitian ini memiliki banyak variabel sehingga untuk mengurangi redudansi dan meningkatkan akurasi model, dilakukan pemilihan fitur terpenting. Berdasarkan PCA, didapatkan bahwa tiga hingga empat komponen utama dapat dipilih. *Feature selection* dilakukan dengan metode Random Forest dan didapatkan 4 fitur terpenting yang akan digunakan dalam analisis.

Variabel	Keterangan
radius_mean	Rata-rata jarak dari pusat kanker ke titik-titik di sekeliling
texture_mean	Standar deviasi dari nilai gray-scale
perimeter_mean	Ukuran rata-rata dari tumor inti

area_mean	Rata-rata area tumor

4. Prediksi Data

   Prediksi data pasien kanker payudara pada penelitian ini akan digunakan metode machine learning dengan bantuan *python.* Machine learning adalah salah satu cabang Artificial Intelligence (AI) yang berfokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk belajar dari data dan membuat keputusan tanpa memerlukan pemrograman eksplisit untuk setiap tugas. Machine learning terdiri dari dua tipe, yaitu supervised learning dan unsupervised learning.

   Supervised learning adalah pendekatan machine learning yang menggunakan data-data yang sudah diberi label yang berarti data input sudah dipasangkan dengan output yang diinginkan. Komputer kemudian belajar untuk memprediksi output untuk data baru. Supervised learning sering digunakan untuk klasifikasi dan regresi.

5. Metode Evaluasi

   Untuk mengevaluasi kinerja model machine learning yang digunakan dalam penelitian ini, beberapa metrik evaluasi yang penting digunakan. Berikut adalah penjelasan mengenai metode evaluasi yang diterapkan:

   1. Akurasi

      Akurasi adalah metrik yang mengukur proporsi prediksi yang benar dari total prediksi yang dibuat oleh model. Akurasi dihitung dengan rumus:

      $$Akurasi = \frac{Jumlah\ prediksi\ benar}{Total\ prediksi}$$

   2. Confusion Matrix

      Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi. Tabel ini menunjukkan jumlah prediksi benar dan salah yang dibagi menjadi empat kategori: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

      Dari confusion matrix, dapat dihitung metrik-metrik tambahan seperti Precision, Recall, dan F1-Score.

   3. Precision dan Recall

      Precision mengukur ketepatan dari prediksi positif yang dibuat oleh model, dihitung dengan rumus:

      $$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

Recall (juga dikenal sebagai Sensitivity) mengukur kemampuan model dalam mendeteksi semua kasus positif, dihitung dengan rumus:

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

4. F1-Score

   F1-Score adalah metrik yang menggabungkan Precision dan Recall menjadi satu nilai yang merepresentasikan keseimbangan antara keduanya. F1-Score dihitung dengan rumus:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. ROC Curve dan AUC (Area Under Curve)

   ROC Curve adalah grafik yang menunjukkan trade-off antara True Positive Rate (Recall) dan False Positive Rate (FPR) pada berbagai threshold. AUC adalah metrik yang mengukur seluruh area di bawah ROC Curve. Nilai AUC berkisar antara 0 dan 1, dengan nilai yang lebih tinggi menunjukkan kinerja model yang lebih baik.

Dalam penelitian ini, model-model machine learning dievaluasi menggunakan metrik-metrik di atas untuk menentukan algoritma terbaik dalam mengklasifikasikan keganasan kanker payudara. Berdasarkan hasil evaluasi, algoritma Decision Tree menunjukkan kinerja terbaik dengan akurasi tertinggi dan metrik-metrik evaluasi lainnya yang mendukung.

D. Hasil dan Diskusi

Pengklasifikasian kanker payudara dilakukan dengan 4 metode, yaitu Logistic Regression, Decision Tree, Random Forest, dan KNN untuk memprediksi keganasan kanker. Analisis dilakukan menggunakan 569 data pasien dengan 4 fitur terpenting, yaitu radius_mean, texture_mean, perimeter_mean, dan area_mean. Data tersebut dibagi menjadi 70% data latih dan 30% data uji.

1. Hasil Penelitian

Pada penelitian ini dilakukan dengan empat algoritma yang berbeda untuk melakukan klasifikasi terhadap keganasan kanker payudara, berikut hasil dari masing-masing algoritma.

1. Logistic Regression

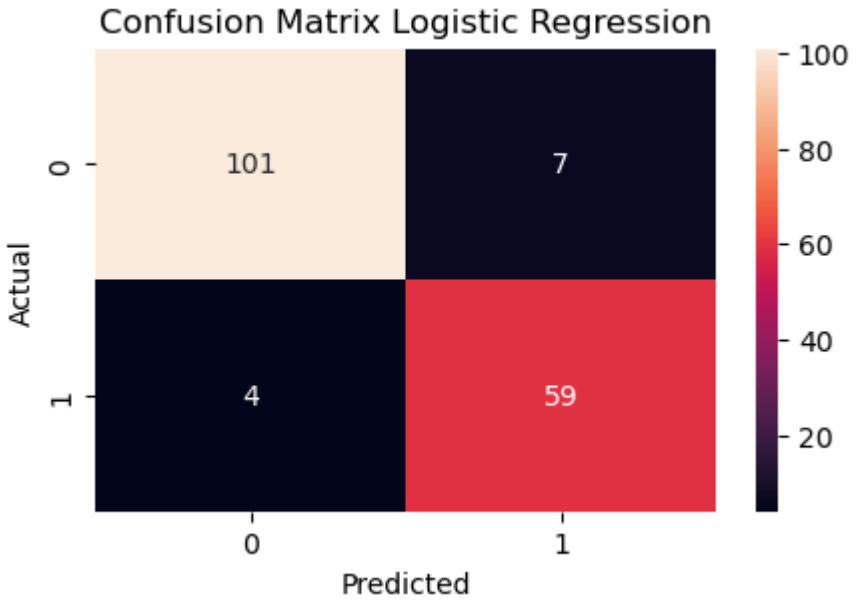
   Pada pengujian yang dilakukan dengan algoritma Logistic Regression mendapatkan hasil seperti pada Tabel 1.

Tabel 1. Akurasi pada algoritma Logistic Regression

Data	Akurasi
Train	0,90201
Test	0,93567

Setelah melakukan pengujian klasifikasi, akan dibuat *confusion matrix* pada data test untuk mengevaluasi proses klasifikasi. Berdasarkan Gambar 1 didapatkan nilai *confusion matrix* dari Logistic Regression,

- True Positif (TP) = 59
- True Negatif (TN) = 101
- False Positif (FP) = 7
- False Negatif (FN) = 4



Gambar 1. *Confusion Matrix* Logistic Regression

2. Decision Tree

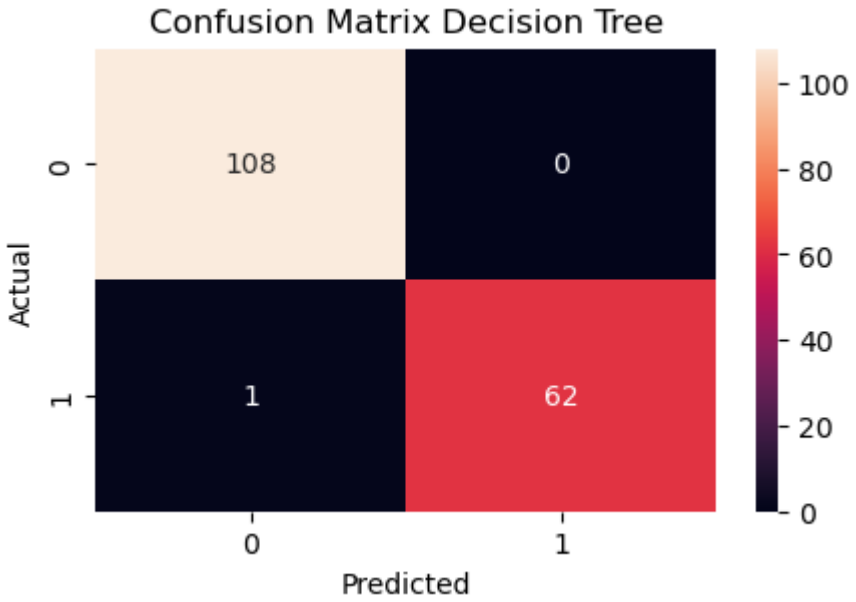
Pada pengujian yang dilakukan dengan algoritma Decision Tree mendapatkan hasil seperti pada tabel 2.

Tabel 2. Akurasi pada algoritma Decision Tree

Data	Akurasi
Train	0,94221
Test	0,99415

Setelah melakukan pengujian klasifikasi, akan dibuat *confusion matrix* pada data test untuk mengevaluasi proses klasifikasi. Berdasarkan Gambar 2 didapatkan nilai *confusion matrix* dari Decision Tree,

- True Positif (TP) = 62
- True Negatif (TN) = 108
- False Positif (FP) = 0
- False Negatif (FN) = 1



Gambar 2. *Confusion Matrix* Decision Tree

3. Random Forest

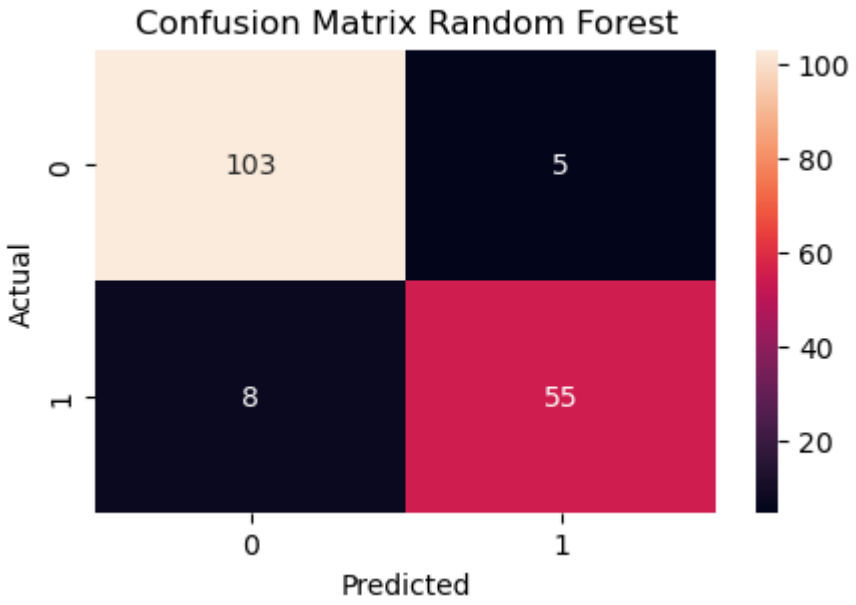
Pada pengujian yang dilakukan dengan algoritma Random Forest mendapatkan hasil seperti pada tabel 3.

Tabel 3. Akurasi pada algoritma Random Forest

Data	Akurasi
Train	0,91960
Test	0,92982

Setelah melakukan pengujian klasifikasi, akan dibuat *confusion matrix* pada data test untuk mengevaluasi proses klasifikasi. Berdasarkan Gambar 3 didapatkan nilai *confusion matrix* dari Random Forest,

- True Positif (TP) = 55
- True Negatif (TN) = 103
- False Positif (FP) = 5
- False Negatif (FN) = 8



Gambar 3. *Confusion Matrix* Random Forest

4. K-Nearest Neighbors (KNN)

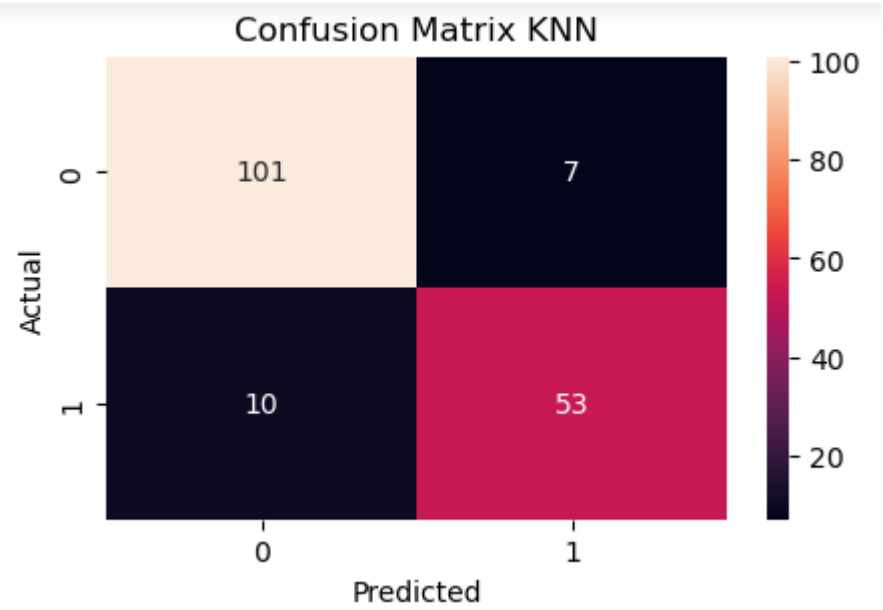
   Pada pengujian yang dilakukan dengan algoritma K-Nearest Neigbors mendapatkan hasil seperti pada tabel 4.

Tabel 4. Akurasi pada algoritma KNN

Data	Akurasi
Train	0,90452
Test	0,93567

Setelah melakukan pengujian klasifikasi, akan dibuat *confusion matrix* pada data test untuk mengevaluasi proses klasifikasi. Berdasarkan Gambar 4 didapatkan nilai *confusion matrix* dari KNN,

- True Positif (TP) = 53
- True Negatif (TN) = 101
- False Positif (FP) = 7
- False Negatif (FN) = 10



Gambar 4. *Confusion Matrix* KNN

2. Hasil Perbandingan

Setelah dilakukan pengujian dari setiap algoritma, hasil yang didapatkan dibandingkan akurasi nya untuk menentukan algoritma yang cocok untuk mengklasifikasi data keganasan kanker payudara. Berdasarkan hasil penelitian yang telah dilakukan, berikut hasil perbandingan akurasi pada data test dari keempat metode.

Tabel 5. Perbandingan Akurasi

Algoritma	Accuracy
Logistic Regression	0,93567
Decision Tree	0,99415
Random Forest	0,92982
KNN	0,93567

Berdasarkan Tabel 5, dapat dilihat dari keempat algoritma *machine learning* yang digunakan, yaitu Logistic Regression, Decision Tree, Random Forest, dan KNN yang menunjukkan hasil terbaik adalah pada algoritma Decision Tree dengan nilai akurasi 99,415%. Nilai akurasi yang sangat tinggi menunjukkan performa model dan seleksi fitur yang baik.

E. Kesimpulan

Klasifikasi keganasan kanker payudara merupakan salah satu cara untuk meningkatkan efektivitas penanganan kanker melalui pemanfaatan teknologi machine learning. Penelitian ini menggunakan empat algoritma yaitu Logistic Regression, Decision Tree, Random Forest, dan K-Nearest Neighbors (KNN) pada dataset pasien dari Kaggle dengan 569 data pasien dan empat fitur utama yang telah dipilih. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memiliki tingkat akurasi terbaik sebesar 99,415%, mengungguli algoritma lainnya. Untuk meningkatkan akurasi prediksi dan efektivitas model machine learning dalam klasifikasi keganasan kanker payudara, diperlukan upaya personalisasi untuk setiap pasien dan studi lanjutan dengan dataset yang lebih besar dan lebih beragam. Hal ini akan membantu dalam meningkatkan pelayanan kesehatan, mengoptimalkan alokasi sumber daya, dan mengurangi angka kematian akibat kanker payudara.

Saran

Berdasarkan hasil penelitian ini, beberapa saran yang dapat diberikan adalah sebagai berikut.

- Pemerintah perlu meningkatkan investasi dalam teknologi kesehatan, khususnya dalam pengembangan dan penerapan machine learning untuk diagnosis dan penanganan kanker. Hal ini termasuk pengadaan perangkat keras dan lunak yang diperlukan serta pelatihan tenaga medis untuk mengoperasikannya.
- Mengembangkan program untuk meningkatkan akses masyarakat terhadap teknologi diagnostik terbaru dan meningkatkan kesadaran tentang pentingnya deteksi dini kanker payudara melalui kampanye edukasi dan pemeriksaan rutin.

Daftar Pustaka

Akbar, M. N. (2020). *KLASIFIKASI KANKER MENGGUNAKAN ALGORITMA NNGE, RANDOM FOREST, DAN RANDOM COMMITEE*.

Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learningtechniques. *Sensors*, *22*(13), 4670. https://doi.org/10.3390/s22134670

Hidayah, K. T., Arifitama, B., & Permana, S. D. (2024). Klasifikasi Penyakit Kanker Serviks Berdasarkan Kebiasaan Dan Rekam medis Dengan Metode c4.5. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *10*(1), 36–44. https://doi.org/10.25077/teknosi.v10i1.2024.36-44

Meyana, N. A., Sitorus, Y. T. S., & Soleh, A. Z. (2023). *Klasifikasi Kanker Payudara Menggunakan Metode K-Nearest Neighbor*.

Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma random forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 67–75. https://doi.org/10.51903/e-bisnis.v13i2.247

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

WHO. (2021). Breast cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Lampiran

Dataset : Data Project PDS Kelompok 10

Syntax : Syntax Project PDS Kelompok 10