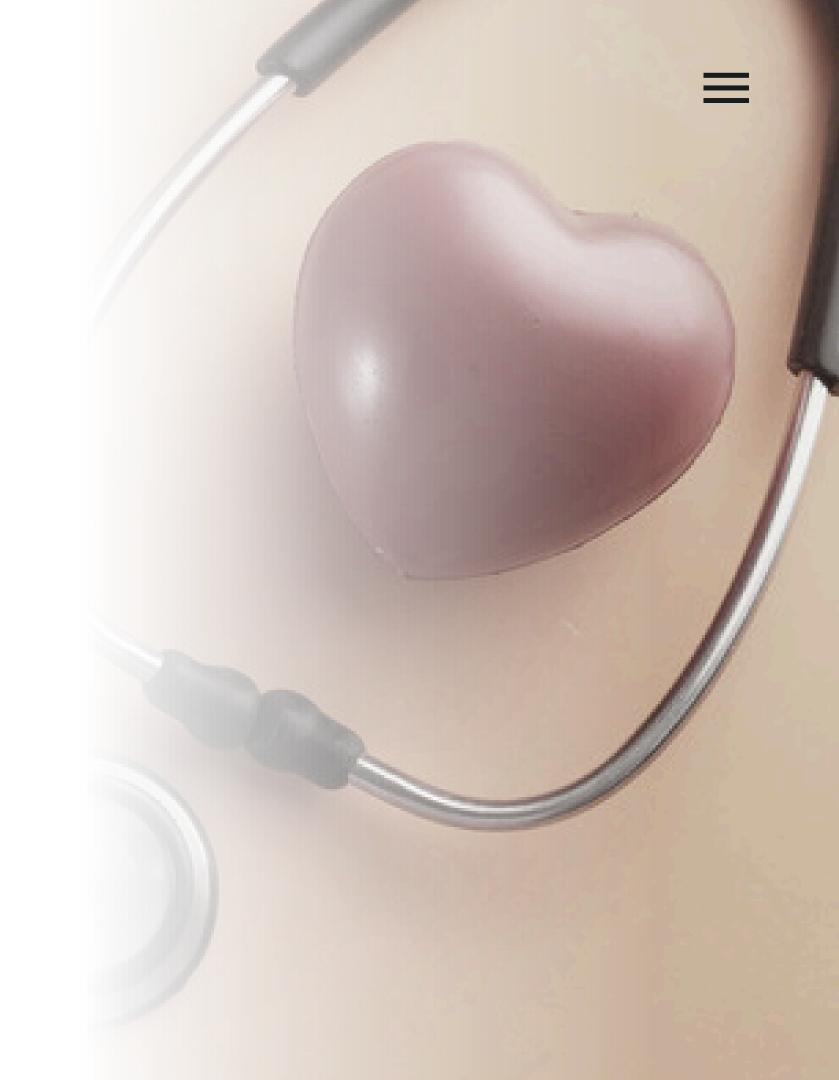
SISTEM PERINGATAN DINI UNTUK SKRINING RISIKO PENYAKIT JANTUNG



M

2702243016 - Ricky Rudiansyah

2201848545 - Hafid Nur Shiddiq

2702268555 - Muhammad Rafly

2702362320 - Shandy Shulton Shihab

2702359175 - Dio Obriel Saragih



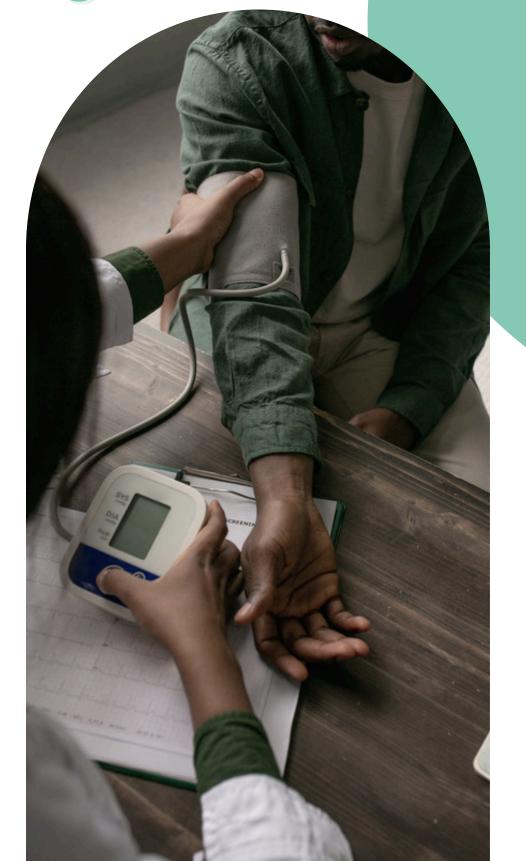


Latar Belakang dan Tujuan Proyek

 Penyakit jantung adalah salah satu penyebab utama kematian global dengan biaya perawatan yang sangat tinggi.

• Deteksi dini seringkali terhambat oleh biaya dan aksesibilitas tes medis yang komprehensif.

• Tujuan Utama: Membangun model klasifikasi Machine Learning sebagai alat skrining berbiaya rendah dan non-invasif.

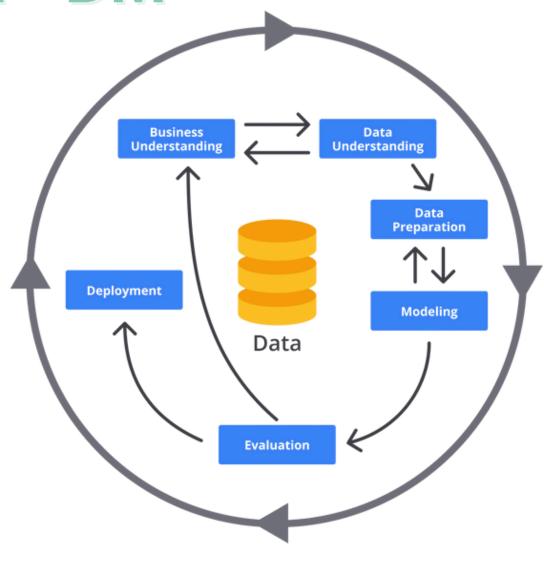




//// Fauget.



Metodologi Proyek: CRISP-DM



Data Understanding

- Sumber Data: Dataset CDC BRFSS 2022 (Behavioral Risk Factor Surveillance System).
- Lingkup: Subset yang telah dikurasi, berisi lebih dari 300.000 responden di Amerika Serikat.
- Konteks: Merupakan survei kesehatan telepon berkelanjutan terbesar di dunia, memberikan data yang sangat kredibel mengenai status kesehatan dan faktor risiko.







Data Understanding

Karakteristik Data: Gabungan 40 variabel yang mencakup:

- Demografis: Age_Category (Kategori Usia), Sex (Jenis Kelamin), Race.
- Kondisi Kesehatan Klinis: Diabetic (Status Diabetes), PhysicalHealth (Jumlah hari sakit), SkinCancer, dll.
- Faktor Gaya Hidup & Perilaku: Smoking (Status Merokok), AlcoholDrinking, PhysicalActivity (Aktivitas Fisik).
- Indikator Umum: BMI (Indeks Massa Tubuh), GeneralHealth (Penilaian Kesehatan Umum).





								leatm	an Ko	rolas	i Anta	r Fitu	-			44.76	-	10	
General Health -	1.00	-0.09	0.28	-0.23	-0.05	-0.15			_					-0.25	-0.17	0.12	0.10	0.12	-0.03
_																			
Checkup -	-0.09	1.00	-0.03	0.09	0.08	0.09	0.04	0.14	0.16	-0.11	0.24	-0.10	0.01	0.06	-0.01	-0.05	0.05	0.04	-0.06
Exercise -	0.28	-0.03	1.00	-0.10	-0.00	-0.05	-0.08	-0.14	-0.12	0.06	-0.12	0.09	-0.09	-0.16	-0.09	0.09	0.14	0.13	-0.04
Heart_Disease -	-0.23	0.09	-0.10	1.00	0.09	0.09	0.03	0.18	0.15	0.07	0.23	0.02	0.05	0.04	0.11	-0.04	-0.02	-0.02	-0.01
Skin_Cancer	-0.05	0.08	-0.00	0.09	1.00	0.15	-0.01	0.04	0.14	0.01	0.27	0.01	-0.03	-0.04	0.03	0.04	0.02	0.01	-0.04
Other_Cancer -	-0.15	0.09	-0.05	0.09	0.15	1.00	0.02	0.07	0.13	-0.04	0.23	-0.04	-0.02	0.00	0.05	-0.01	0.01	-0.00	-0.03
Depression -	-0.21	0.04	-0.08	0.03	-0.01	0.02	1.00	0.05	0.12	-0.14	-0.10	-0.09	0.05	0.11	0.10	-0.03	-0.04	-0.05	0.02
Diabetes -	-0.28	0.14	-0.14	0.18	0.04	0.07	0.05	1.00	0.14	0.00	0.21	-0.03	0.17	0.21	0.06	-0.12	-0.02	-0.03	-0.00
Arthritis -																			
Sex -	0.02	-0.11	0.06	0.07	0.01	-0.04	-0.14	0.00	-0.10	1.00	-0.06	0.70	0.35	0.01	0.07	0.13	-0.09	-0.07	0.13
Age_Category -	-0.17	0.24	-0.12	0.23	0.27	0.23	-0.10	0.21	0.37	-0.06	1.00	-0.12	-0.06	-0.01	0.13	0.01	0.04	0.04	-0.14
Height_(cm) -	0.07	-0.10	0.09	0.02	0.01	-0.04	-0.09	-0.03	-0.10	0.70	-0.12	1.00	0.47	-0.03	0.05	0.13	-0.05	-0.03	0.11
Weight_(kg) -	-0.18	0.01	-0.09	0.05	-0.03	-0.02	0.05	0.17	0.07	0.35	-0.06	0.47	1.00	0.86	0.05	-0.03	-0.09	-0.08	0.10
BMI -	-0.25	0.06	-0.16	0.04	-0.04	0.00	0.11	0.21	0.14	0.01	-0.01	-0.03	0.86	1.00	0.02	-0.11	-0.08	-0.07	0.05
Smoking History -	-0.17	-0.01	-0.09	0.11	0.03	0.05	0.10	0.06	0.12	0.07	0.13	0.05	0.05	0.02	1.00	0.10	-0.09	-0.03	0.04
Alcohol_Consumption -	0.12	-0.05	0.09	-0.04	0.04	-0.01	-0.03	-0.12	-0.02	0.13	0.01	0.13	-0.03	-0.11	0.10	1.00	-0.01	0.06	0.03
Fruit_Consumption -	0.10	0.05	0.14	-0.02	0.02	0.01	-0.04	-0.02	-0.00	-0.09	0.04	-0.05	-0.09	-0.08	-0.09	-0.01	1.00	0.27	-0.06
Green_Vegetables_Consumption -	0.12	0.04	0.13	-0.02	0.01	-0.00	-0.05	-0.03	-0.02	-0.07	0.04	-0.03	-0.08	-0.07	-0.03	0.06	0.27	1.00	0.00
FriedPotato_Consumption -	-0.03	-0.06	-0.04	-0.01	-0.04	-0.03	0.02	-0.00	-0.05	0.13	-0.14	0.11	0.10	0.05	0.04	0.02	-0.06	0.00	1.00
	General_Health -	Checkup -	Exercise -	Heart_Disease -	Skin_Cancer -	Other_Cancer -	Depression -	Diabetes -	Arthritis -	- xex	Age_Category -	Height_(cm) -	Weight_(kg) -	BMI -	Smoking_History -	cohol_Consumption -	Fruit_Consumption -	ables_Consumption -	otato_Consumption -



- 0.6

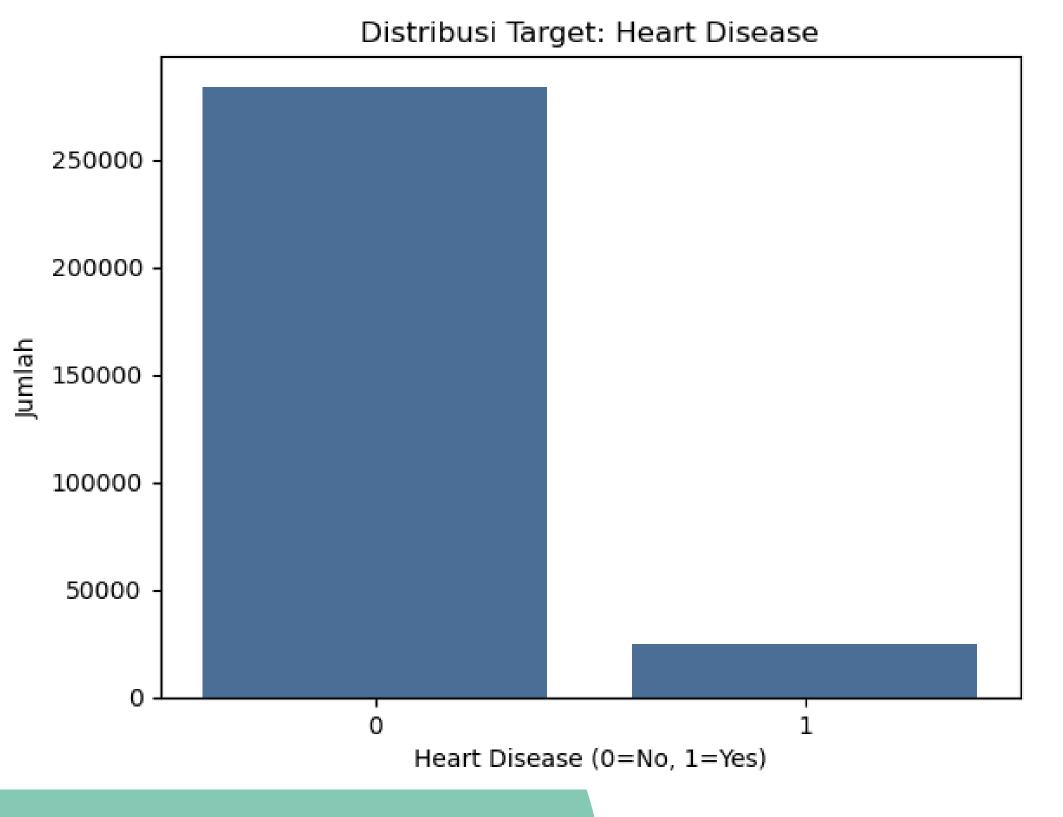
- 0.4

- 0.2

- 0.0

- -0.2

ויוון

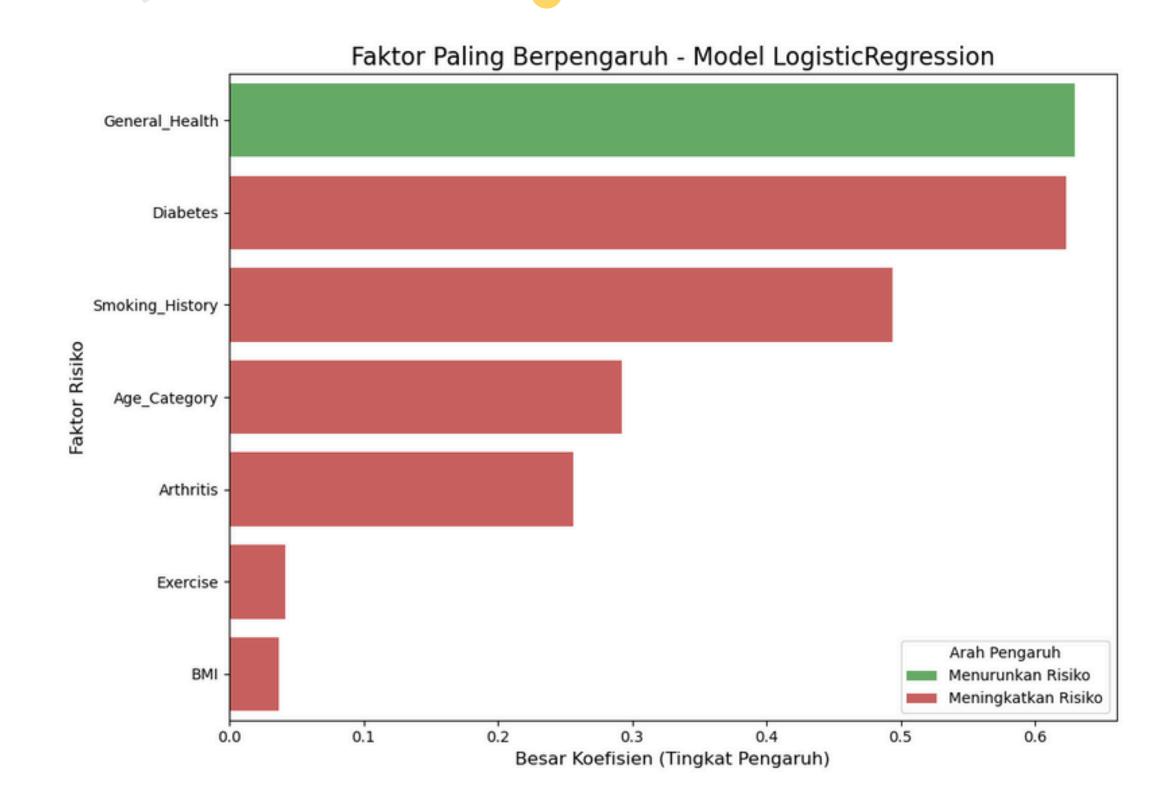


Data Understanding

- Distribusi data target sangat tidak seimbang.
 Jumlah responden tanpa penyakit jantung (Kelas 0) jauh lebih banyak daripada yang memiliki penyakit jantung (Kelas 1).
- Risiko: Jika tidak ditangani, model akan cenderung "malas" dan sangat bias dalam memprediksi kelas mayoritas (Kelas 0), sehingga tidak berguna untuk mendeteksi penyakit.

Data Preparation

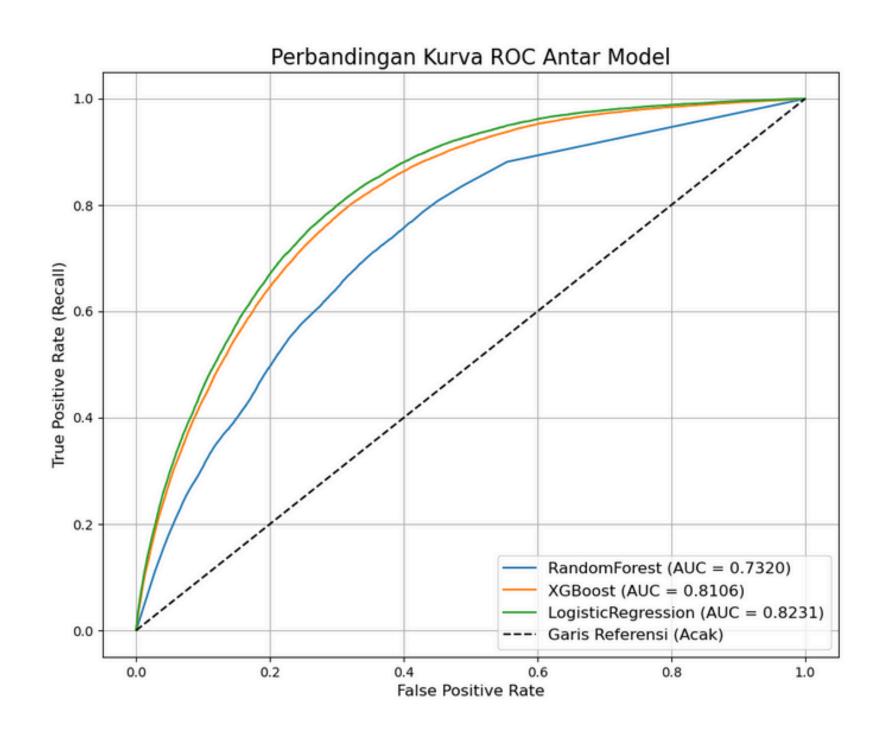
- Encoding Kategorikal: Mengubah data teks menjadi numerik.
- Ordinal Encoding: Untuk fitur berurutan seperti General_Health.
- Nuanced Encoding: Menerapkan logika klinis pada Diabetes untuk merefleksikan tingkat risiko (Tidak=0, Pra-diabetes=0.5, Gestasional=0.75, Ya=1).
- Feature Selection: Berdasarkan korelasi dan domain knowledge, kami memfokuskan model pada 7 fitur paling berpengaruh untuk menciptakan model yang ramping dan efisien.



Modeling

- Penanganan Imbalance: Menggunakan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menyeimbangkan data training secara sintetis.
- Pipeline Terpadu: Menggabungkan semua langkah (preprocessing, SMOTE, dan classifier) ke dalam satu Scikit-learn Pipeline untuk memastikan proses yang konsisten dan bebas dari kebocoran data.
- Validasi Robust: Menggunakan Stratified K-Fold Cross-Validation untuk mendapatkan evaluasi performa yang stabil dan dapat diandalkan.
 - Model yang Diuji
 - Logistic Regression (sebagai baseline)
 - Random Forest
 - XGBoost (dengan dan tanpa Hyperparameter Tuning)





Evaluasi: Perbandingan Kinerja Model

Fokus pada Recall (kemampuan menemukan kasus positif) dan ROC AUC (kemampuan membedakan kelas).

- Hasil Perbandingan
- RandomForest menunjukkan performa di bawah standar (Recall rendah).
- XGBoost (baik sebelum maupun sesudah tuning) menunjukkan performa yang baik, namun tidak secara signifikan melampaui model baseline.
- Logistic Regression menunjukkan Recall tertinggi (79%) dan ROC AUC yang sangat kompetitif (82%).

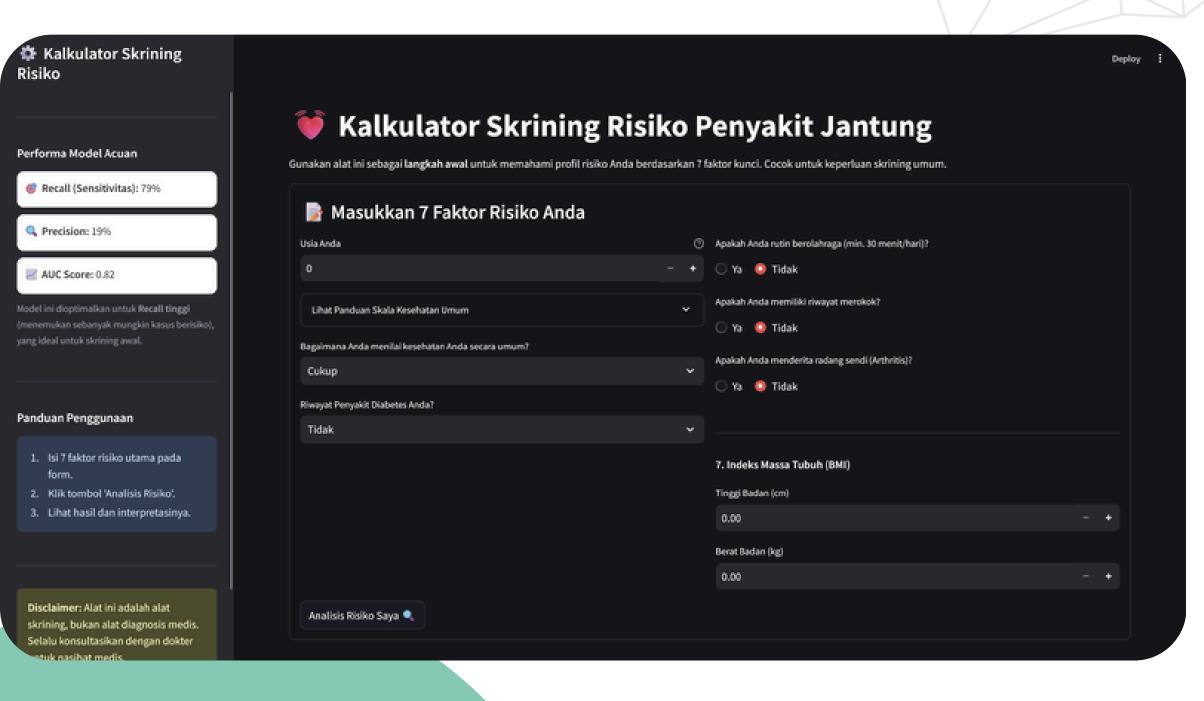
Confusion Matrix (Jumlah Absolut) Normalized Confusion Matrix (Recall) LogisticRegression_HeartDisease LogisticRegression_HeartDisease No Disease 202066 81817 71.18% 28.82% Aktual 5329 19642 21.34% 78.66% No Disease Disease No Disease Disease Prediksi Prediksi

Evaluasi: Analisis Model Terpilih

- True Positives (TP): Berhasil mengidentifikasi [jumlah] pasien sakit.
- True Negatives (TN): Berhasil mengidentifikasi [jumlah] pasien sehat.
- False Positives (FP): [Jumlah] pasien sehat yang keliru diidentifikasi berisiko (sumber presisi rendah).
- False Negatives (FN): Hanya [jumlah] pasien sakit yang terlewatkan (bukti recall tinggi).

Kesimpulan Evaluasi: Dengan Recall tertinggi dan kompleksitas paling rendah, Logistic Regression adalah model terpilih untuk implementasi.

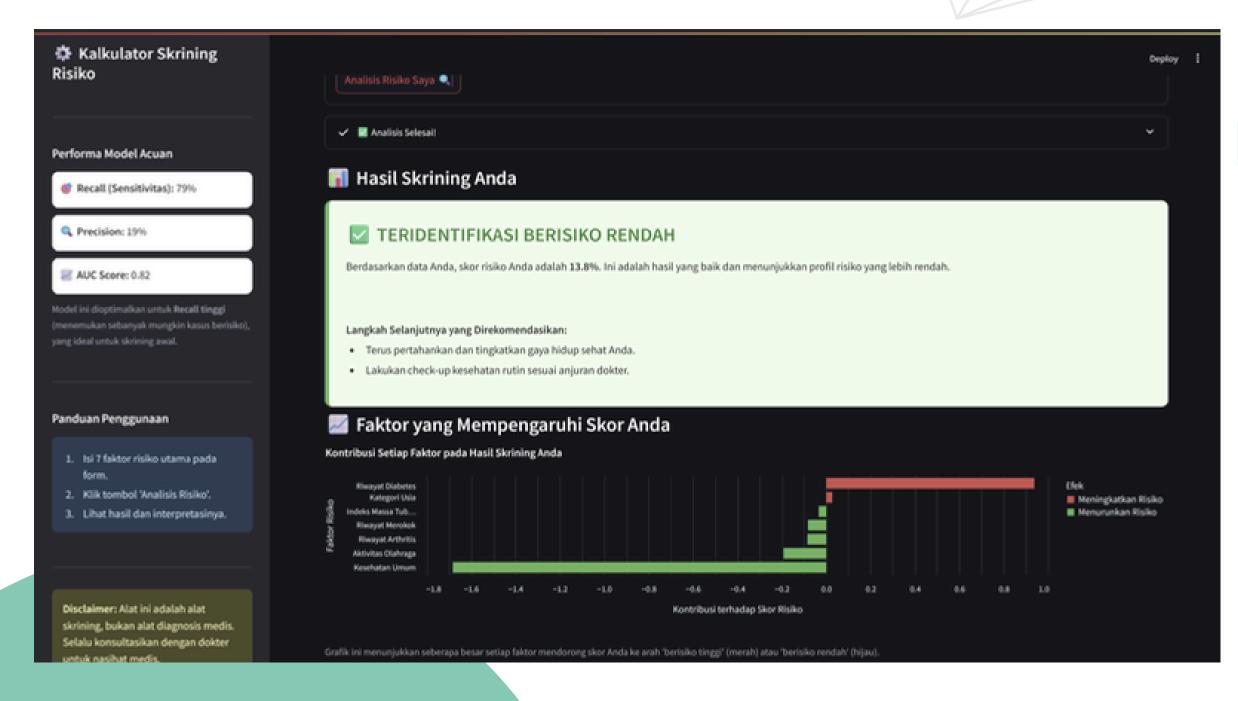




Deployment

- Produk: Sebuah dashboard interaktif yang dibangun menggunakan Streamlit.
- Persona Alat: Didesain sebagai "Kalkulator Skrining Risiko", bukan alat diagnostik.
 - Fitur Utama
- Form input yang sederhana (hanya 7 faktor).
- Hasil prediksi real-time.
- Analisis faktor risiko personal (SHAP).





Implementasi: Interpretasi Hasil & Faktor Risiko

- Model ini bukan "kotak hitam".
- Kita menggunakan analisis SHAP untuk menjelaskan faktor-faktor apa saja yang paling berkontribusi pada skor risiko setiap individu.
 - Manfaat
- Membangun kepercayaan pengguna.
- Memberikan wawasan yang dapat ditindaklanjuti (misal: "BMI Anda adalah faktor risiko terbesar, pertimbangkan untuk menurunkannya").

Kesimpulan Proyek

Pencapaian

- Berhasil membangun model skrining (Logistic Regression) dengan Recall 79% menggunakan hanya 7 fitur non-invasif.
- Mengembangkan dashboard interaktif sebagai proof-of-concept untuk implementasi.
- Membuktikan bahwa model dapat memberikan nilai bisnis yang signifikan di berbagai sektor.

Batasan & Langkah Selanjutnya

- Batasan: Performa model kemungkinan besar dibatasi oleh ketersediaan fitur. Presisi masih menjadi area untuk perbaikan.
- Rekomendasi #1: Untuk peningkatan performa di masa depan, fokus harus beralih ke penambahan data (Data Enhancement), seperti data klinis (kolesterol, tekanan darah).
- Rekomendasi #2: Melakukan A/B testing dashboard pada target pengguna (misal: agen asuransi) untuk mendapatkan umpan balik.

