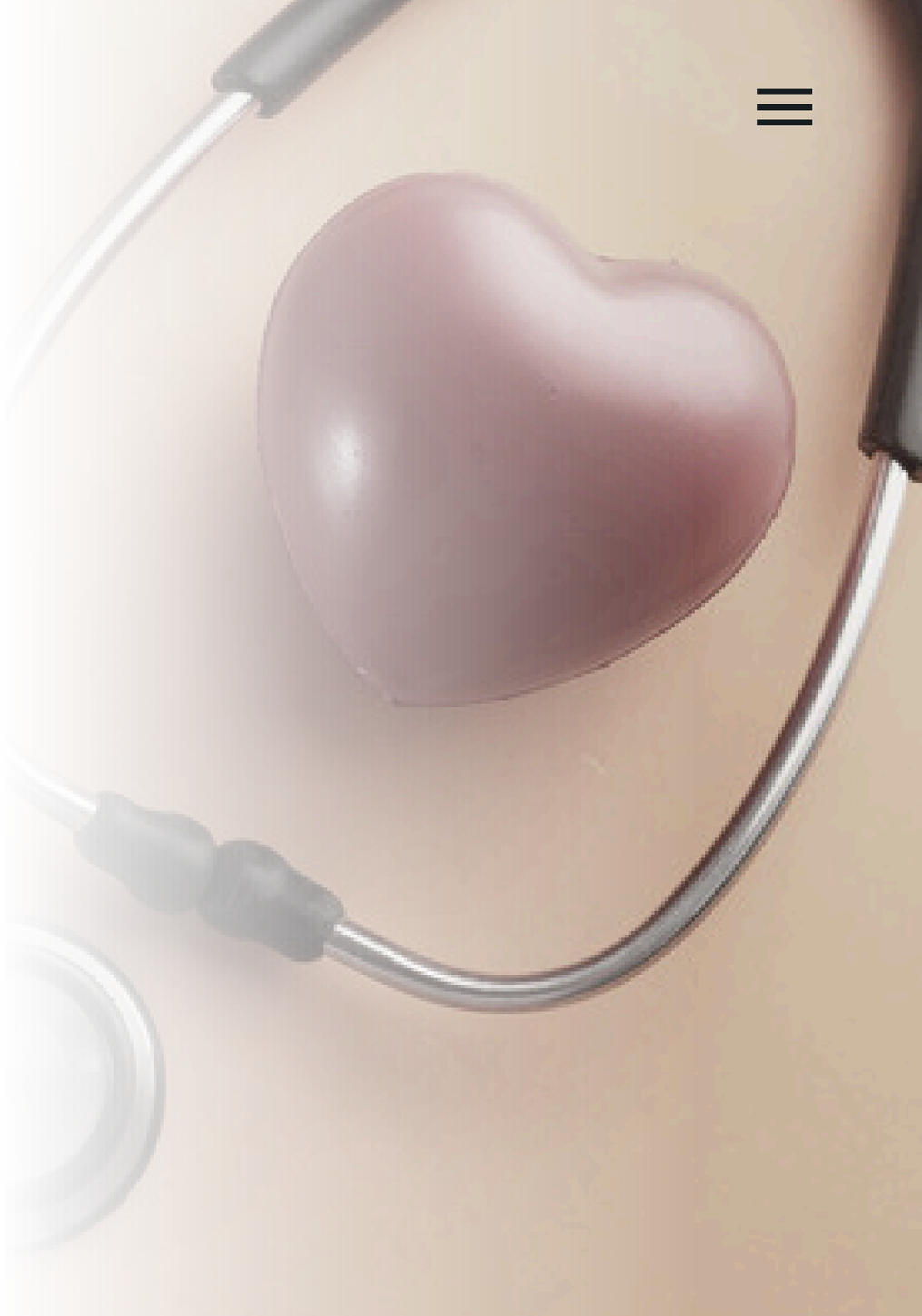
SISTEM PERINGATAN DINI UNTUK SKRINING RISIKO PENYAKIT JANTUNG



2702243016 - Ricky Rudiansyah
2201848545 - Hafid Nur Shiddiq
2702268555 - Muhammad Rafly
2702362320 - Shandy Shulton Shihab
2702359175 - Dio Obriel Saragih

Anggota

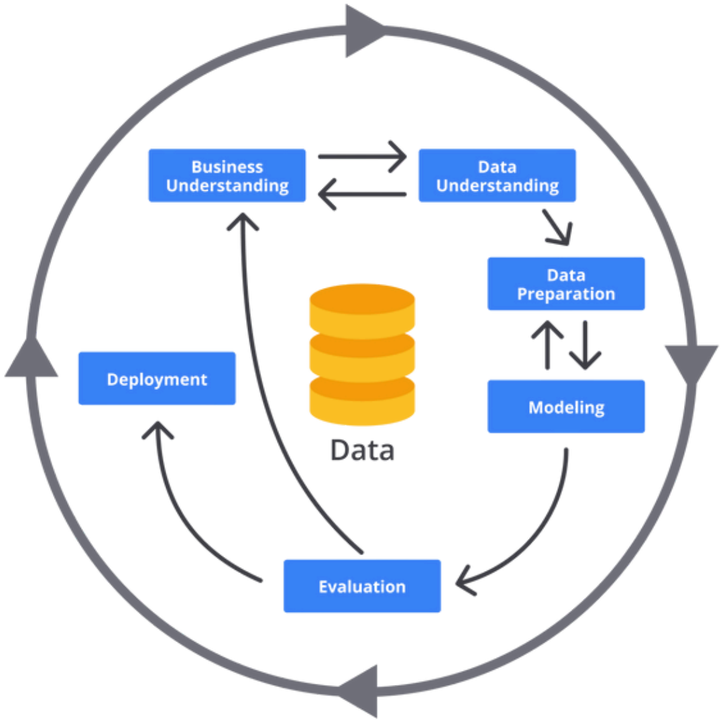
Latar Belakang dan Tujuan Proyek

- 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

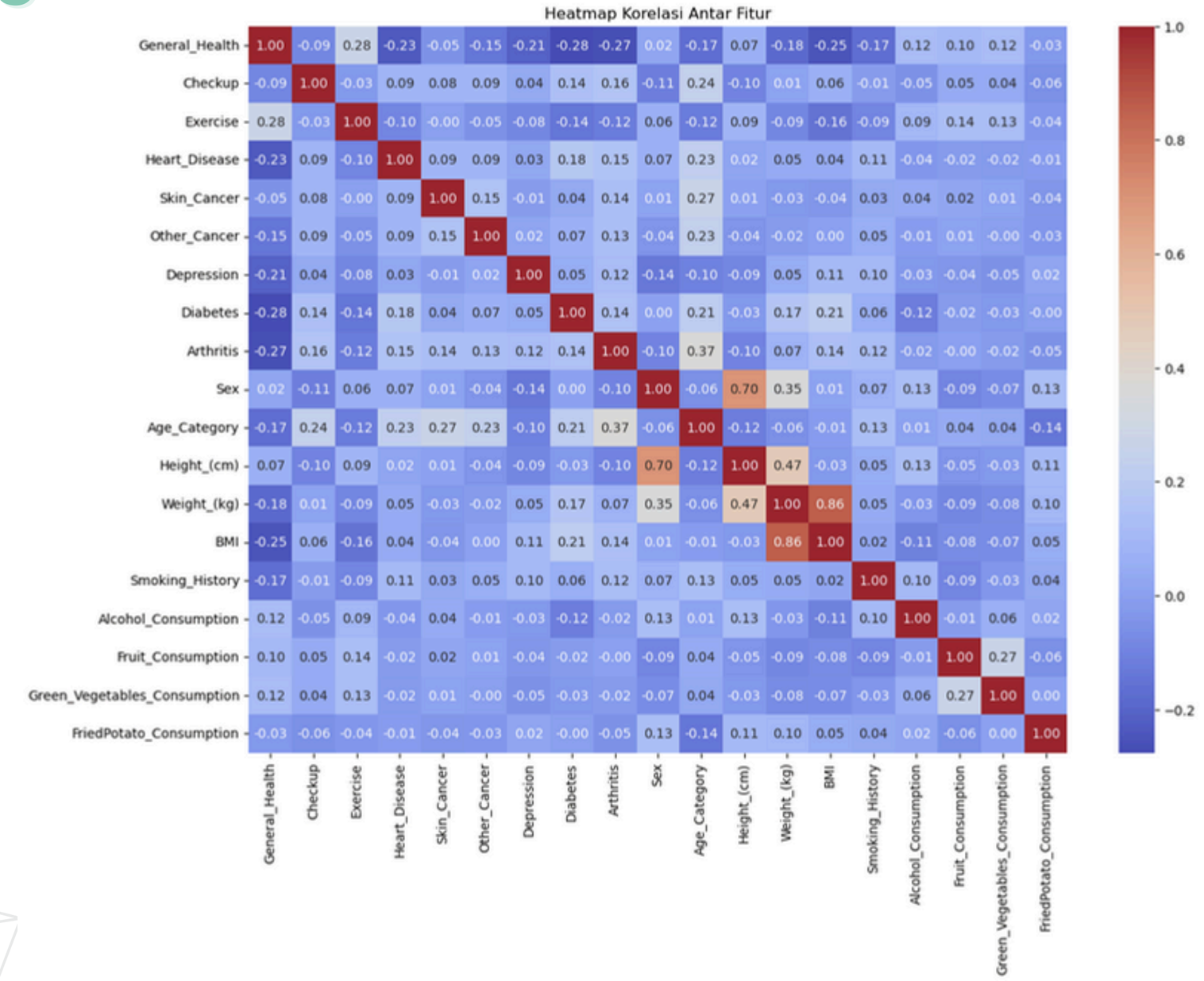
- Sumber Data: Dataset CDC BRFSS 2022 (Behavioral Risk Factor Surveillance System).
- Lingkup: Subset yang telah dikurasi, berisi lebih dari 300.000 responden di Amerika Serikat.
- Konteks: Merupakan survei kesehatan telepon berkelanjutan terbesar di dunia, memberikan data yang sangat kredibel mengenai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

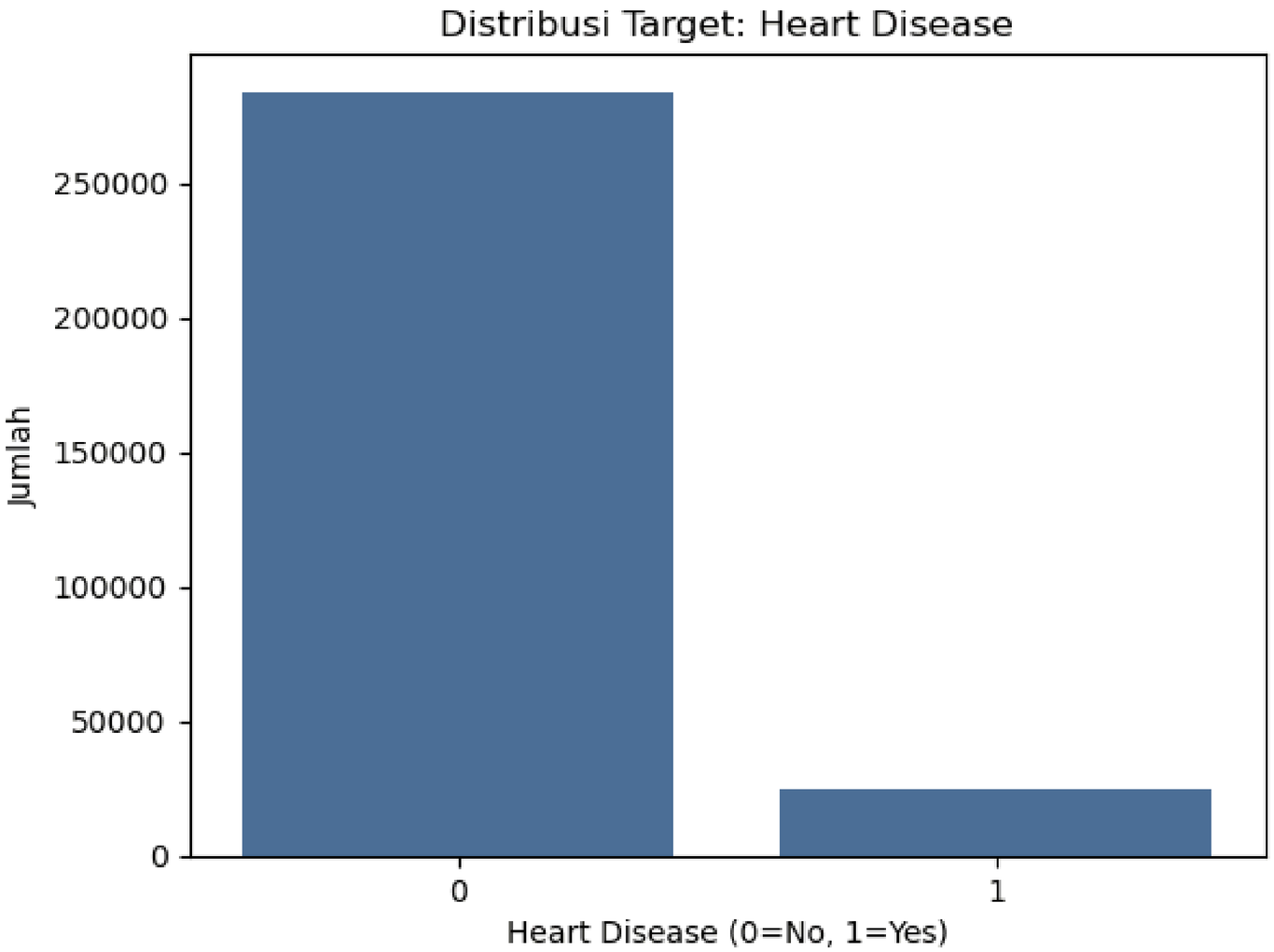
BRFSS™
Behavioral Risk Factor Surveillance System

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).





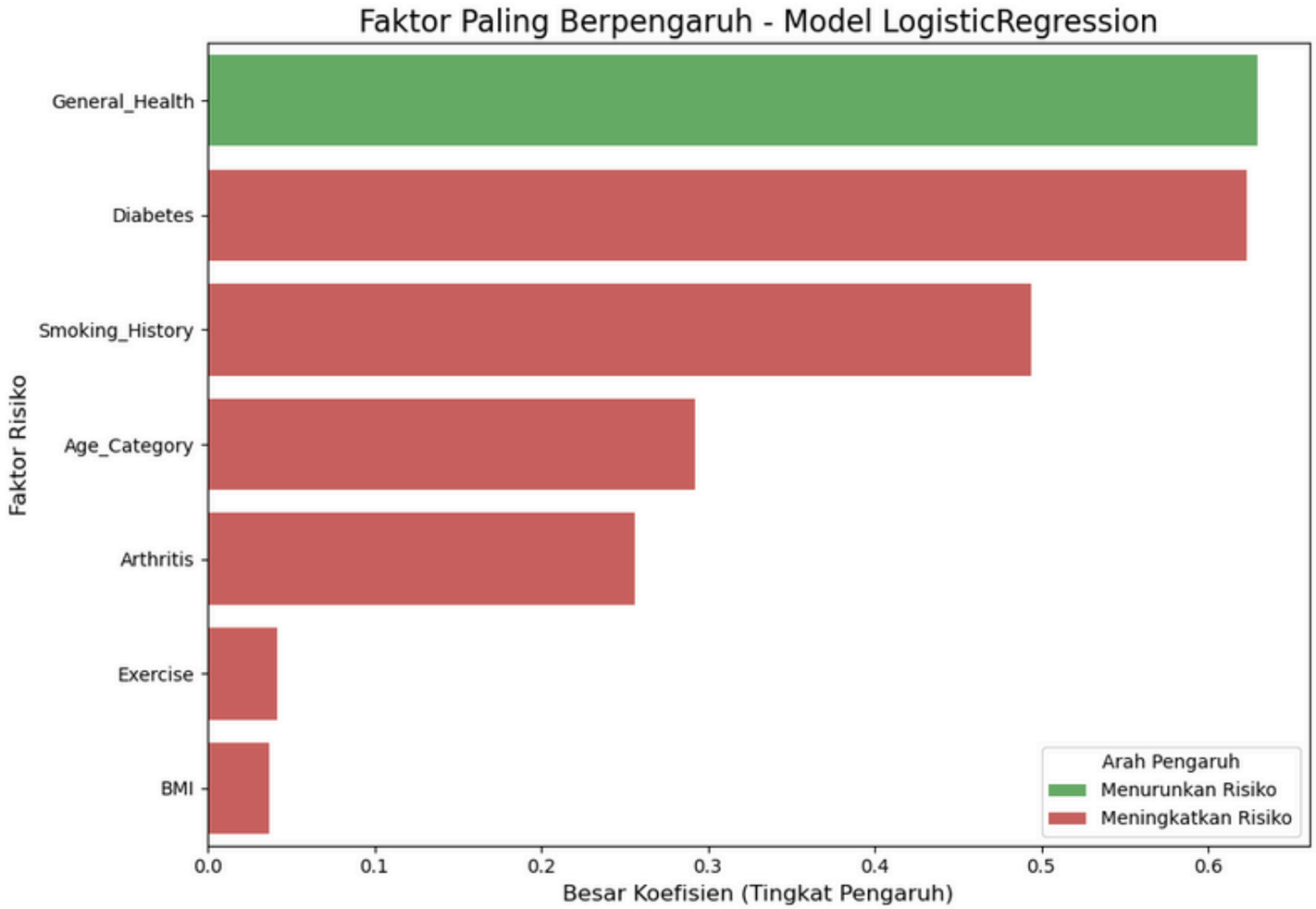
Data Understanding

- 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.

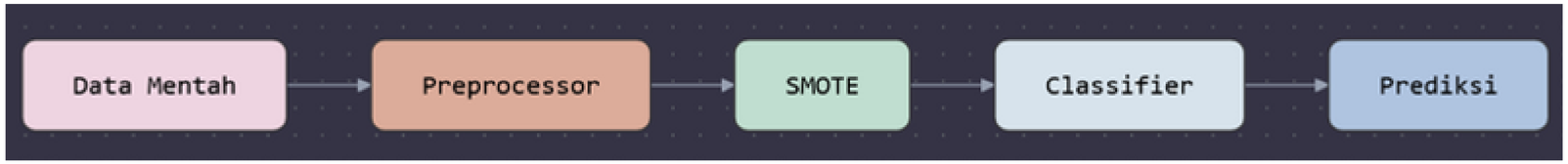


Modeling

- Penanganan Imbalance: Menggunakan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menyeimbangkan data training secara sintetis.

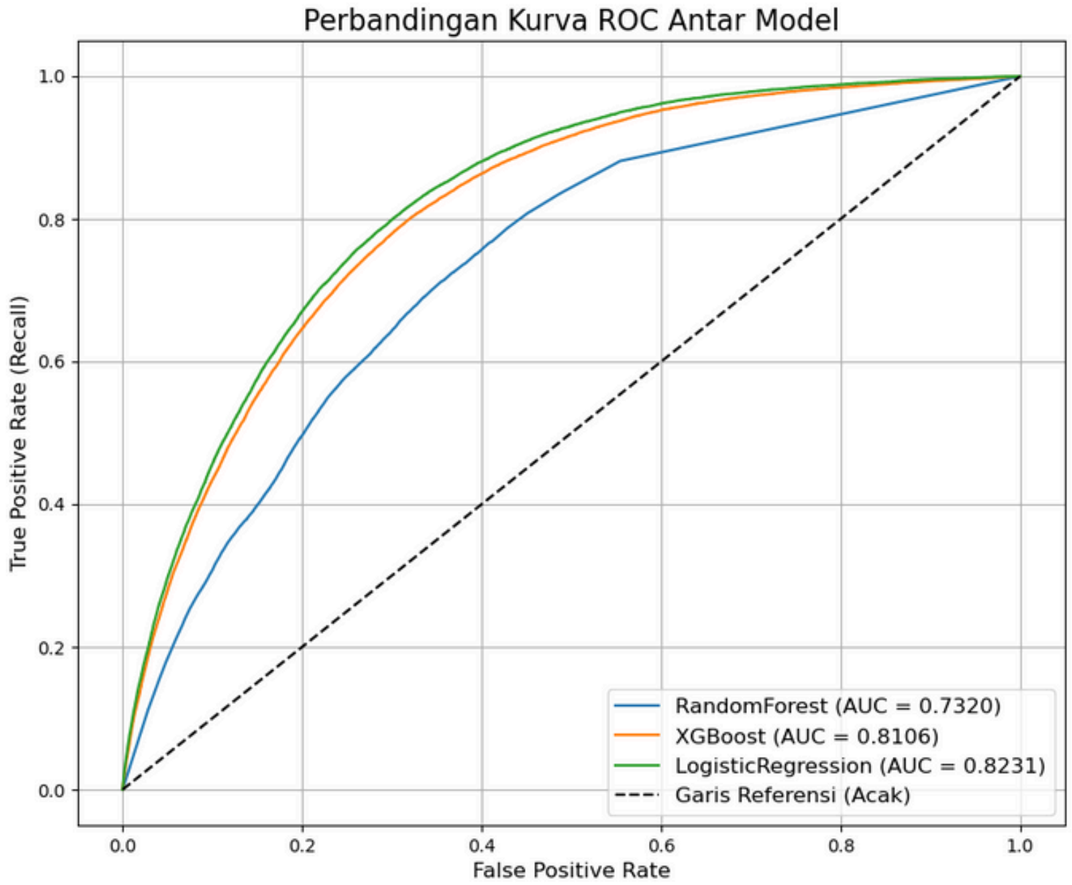
- Pipeline Terpadu: Menggabungkan semua langkah (preprocessing, SMOTE, dan classifier) ke dalam satu Scikit-learn Pipeline untuk memastikan proses yang konsisten dan bebas dari kebocoran data.

- 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)

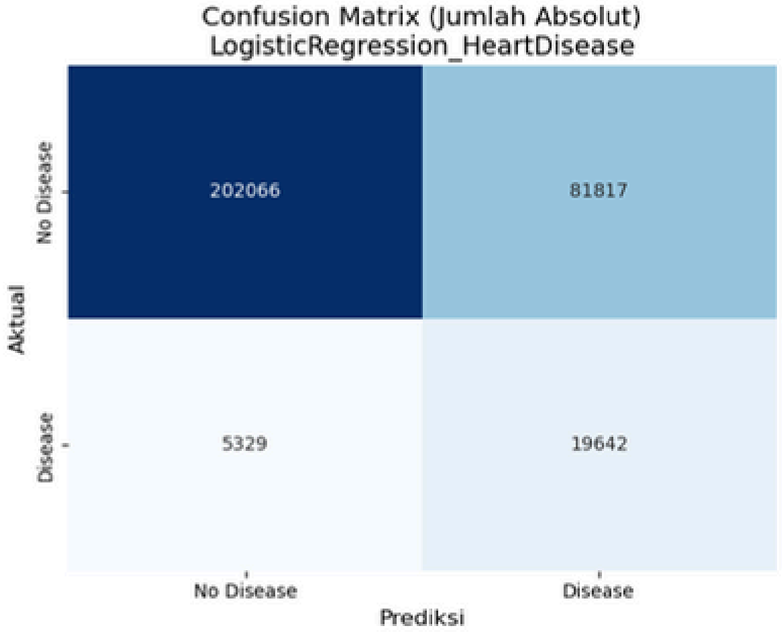
Evaluasi: Perbandingan Kinerja Model

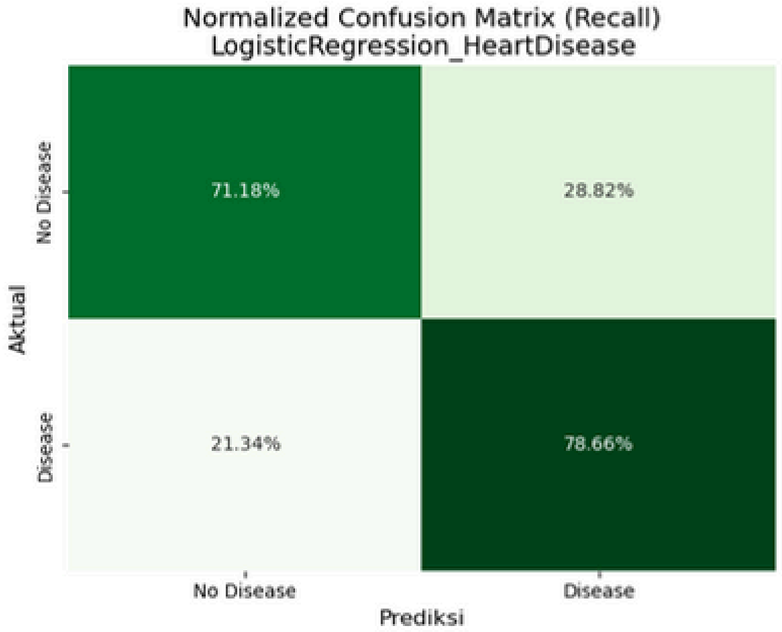


Fokus pada Recall (kemampuan menemukan kasus positif) dan ROC AUC (kemampuan membedakan kelas).

- 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%).

Evaluasi: Analisis Model Terpilih

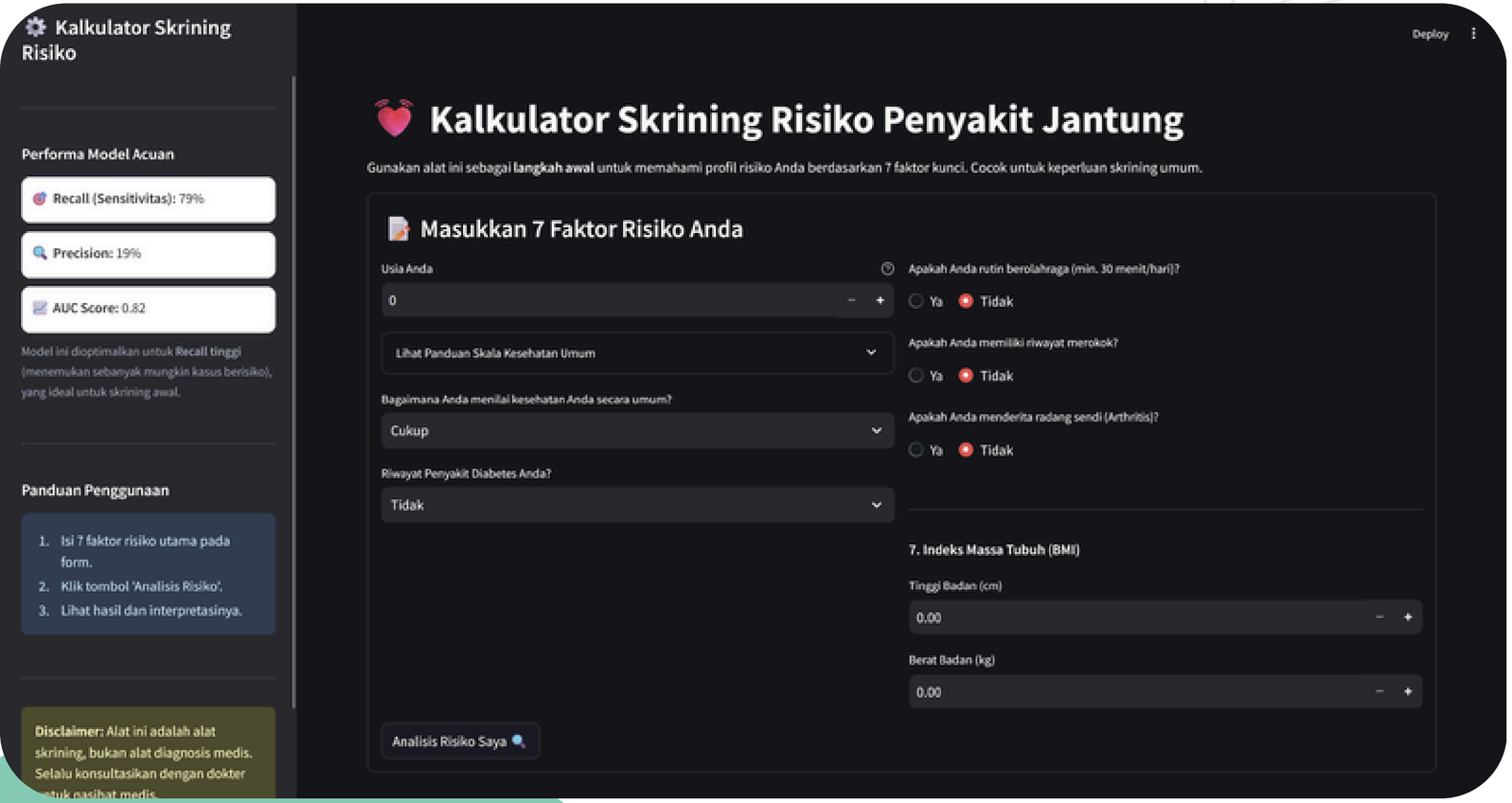




- True Positives (TP): Berhasil mengidentifikasi [jumlah] pasien sakit.
- True Negatives (TN): Berhasil mengidentifikasi [jumlah] pasien sehat.
- False Positives (FP): [Jumlah] pasien sehat yang keliru diidentifikasi berisiko (sumber presisi rendah).
- False Negatives (FN): Hanya [jumlah] pasien sakit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

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

Deployment

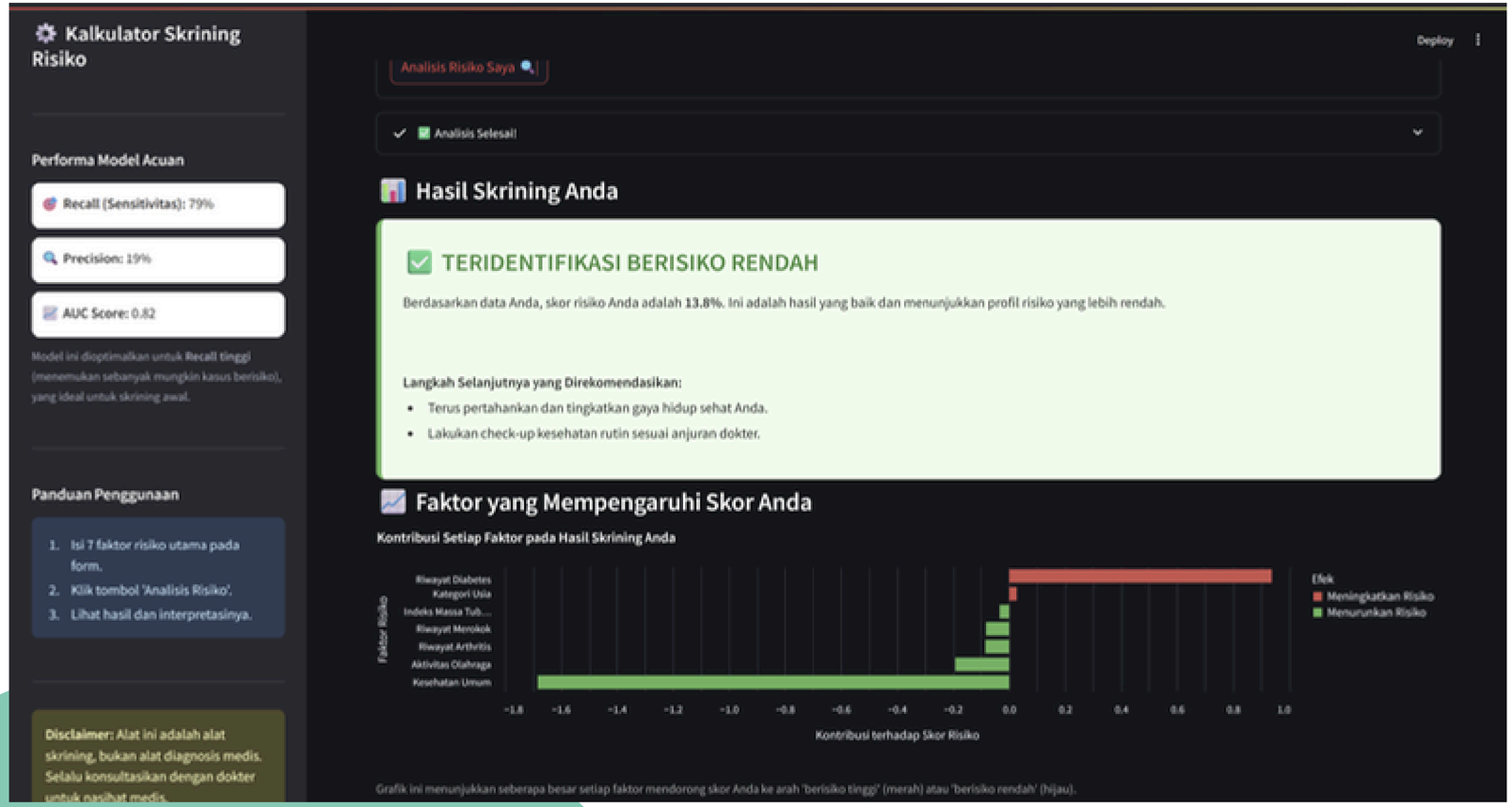


- Produk: Sebuah dashboard interaktif yang dibangun menggunakan Streamlit.

- 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.

THANK YOU