LAPORAN TUGAS BESAR
KECERDASAN BUATAN
PREDEKSI PENYAKIT JANTUNG MENGGUNAKAN *MACHINE LEARNING* DENGAN ALGORITMA RANDOM FOREST DENGAN TEKNIK SMOTEENN

Disusun oleh:

Agna Fadia – 2306145

Rifki Ahmad Dzulfikri – 2306144

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

INSTITUT TEKNOLOGI GARUT

JURUSAN ILMU KOMPUTER

PROGRAM STUDI TEKNIK INFORMATIKA

TAHUN AKADEMIK 2024/2025

1. BUSINESS UNDERSTANDING

a) Permasalahan dunia nyata

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia yang dapat mengancam nyawa jika tidak ditangani dengan tepat. Bahaya penyakit ini meliputi serangan jantung mendadak, gagal jantung, hingga stroke akibat komplikasi pembuluh darah (Sposato et al., 2020).

Dalam praktik medis, diagnosa penyakit jantung memerlukan berbagai parameter seperti tekanan darah, kadar kolesterol, riwayat keluarga, berat badan, serta pemeriksaan lanjutan seperti elektrokardiogram (EKG) dan angiografi. Namun, proses diagnosa manual ini membutuhkan waktu yang cukup lama dan bergantung pada keahlian dokter.

Masalah utama yang dihadapi dalam prediksi penyakit jantung menggunakan machine learning adalah akurasi yang masih rendah, terutama karena adanya ketidakseimbangan data (imbalanced data) dimana jumlah kasus positif penyakit jantung jauh lebih sedikit dibandingkan kasus negatif (Handayani, 2021)

b) Tujuan Proyek

- Membangun model prediksi penyakit jantung yang akurat menggunakan algoritma Random Forest
- Menerapkan teknik SMOTEENN untuk mengatasi masalah ketidakseimbangan data
- Meningkatkan akurasi prediksi dari penelitian sebelumnya yang mencapai 86% menjadi lebih tinggi
- Memberikan alat bantu diagnostik yang dapat mendukung keputusan medis

c) User/Pengguna Sistem

- Dokter dan tenaga medis: Sebagai alat bantu dalam diagnosa awal penyakit jantung
- Rumah sakit dan klinik: Untuk screening pasien berisiko tinggi
- Peneliti medis: Untuk analisis faktor risiko penyakit jantung
- Sistem kesehatan digital: Integrasi dalam aplikasi kesehatan

d) Manfaat Implementasi AI

- Deteksi dini: Identifikasi pasien berisiko tinggi sebelum gejala parah muncul
- Efisiensi waktu: Mempercepat proses screening dan diagnosa
- Konsistensi: Mengurangi variabilitas interpretasi antar dokter
- Cost-effective: Mengurangi biaya pemeriksaan lanjutan yang tidak perlu

- Akurasi tinggi: Meningkatkan ketepatan prediksi dengan teknik SMOTEENN

2. *DATA UNDERSTANDING*

a) Sumber Data

Dataset penyakit jantung diperoleh dari platform Kaggle yang merupakan kumpulan data medis yang telah divalidasi dan sering digunakan dalam penelitian machine learning untuk prediksi penyakit jantung.

b) Deskripsi Setiap Fitur

Fitur	Deskripsi	Tipe Data
Age	Usia pasien (tahun)	Numerik
Sex	Jenis Kelamin (1=Laki Laki, 0=Prempuan)	Kategori
cp	Tipe chest pain (0-3)	Kategori
trestbps	Tekanan darah istirahat (mmHg)	Numerik
chol	Kadar kolesterol serum (mg/dl)	Numerik
fbs	Gula darah puasa > 120 mg/dl (1=Ya, 0=Tidak)	Kategori
restecg	Hasil elektrokardiogram istirahat (0-2)	Kategori
thalach	Detak jantung maksimum yang dicapai	Numerik
exang	Exercise duced angina (1=Ya, 0=Tidak)	Kategori
oldpeak	ST depression induced by exercise	Numerik
slope	Slope dari peak exercise ST segment	Kategori
ca	Jumlah major vessels (0-4)	Numerik
thal	Thalassemia (1,2,3)	Kategori

target	Diagnosis penyakit jantung (1=Ya, 0=Tidak)	Target

c) Ukuran dan Format Data

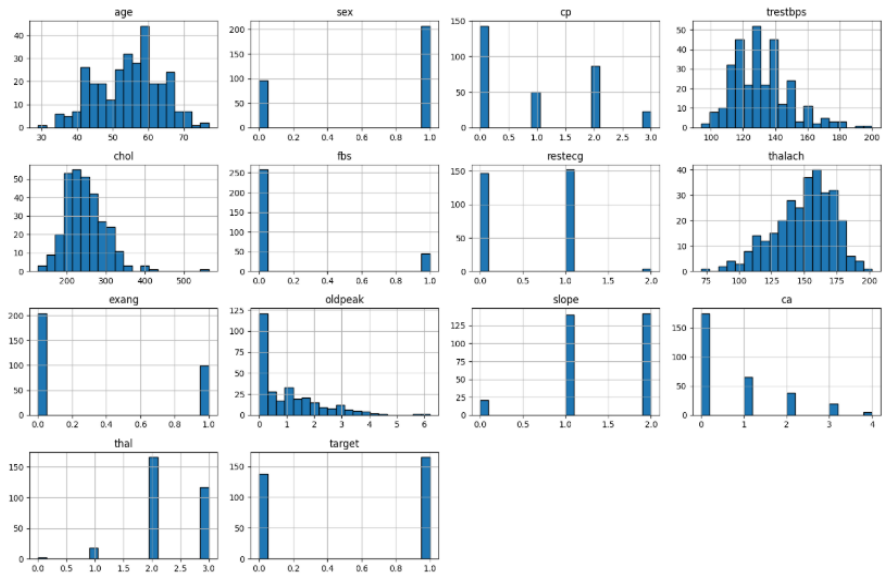
- Jumlah record: 303 Baris Data
- Jumlah fitur: 13 fitur + 1 target (Total 14 Kolom)
- Format: CSV (Comma Separated Values)
- Missing values: Tidak ada missing values (dataset lengkap)

d) Tipe Data dan Target Klasifikasi

- Problem type: Binary Classification
- Target variable: target (0 = Tidak ada penyakit jantung, 1 = Ada penyakit jantung)
- Input features: 13 fitur medis (mix numerik dan kategorik)

3. *EXPLORATORY DATA ANALYSIS (EDA)*

a) Distribusi Data



Gambar 1 Distribusi Data

Analisis Distribusi Data:

1) Age (Usia)

- Distribusi mendekati normal dengan sedikit skewness ke kanan
- Rentang usia sekitar 30-70 tahun
- Puncak distribusi di sekitar 50-55 tahun
- Menunjukkan dataset didominasi usia paruh baya hingga lansia

2) Sex (Jenis Kelamin) (Guan, 2022)
    - Distribusi kategorikal dengan dua nilai (0 dan 1)
    - Tampak tidak seimbang, dengan salah satu kategori lebih dominan
    - Kemungkinan 0=perempuan, 1=laki-laki atau sebaliknya
3) CP (Chest Pain)
    - Distribusi kategorikal dengan 4 kategori (0-3)
    - Kategori 0 paling dominan
    - Distribusi tidak merata antar kategori
4) Trestbps (Tekanan Darah Istirahat)
    - Distribusi mendekati normal
    - Rentang sekitar 100-200 mmHg
    - Puncak di sekitar 130-140 mmHg
    - Beberapa outlier di nilai tinggi
5) Chol (Kolesterol)
    - Distribusi positively skewed (ekor panjang ke kanan)
    - Rentang luas dari ~100-600 mg/dl
    - Mayoritas data terkonsentrasi di 200-300 mg/dl
    - Ada beberapa nilai ekstrem tinggi
6) Restecg (Hasil EKG Istirahat)
    - Distribusi kategorikal dengan 3 kategori
    - Sangat tidak seimbang, kategori 0 sangat dominan
    - Kategori 1 dan 2 memiliki frekuensi rendah
7) Thalach (Detak Jantung Maksimum)
    - Distribusi mendekati normal dengan slight negative skew
    - Rentang sekitar 70-200 bpm
    - Puncak di sekitar 150-160 bpm
8) Exang (Exercise Induced Angina)
    - Distribusi binary (0 dan 1)
    - Kategori 0 lebih dominan daripada kategori 1
    - Menunjukkan mayoritas tidak mengalami angina saat exercise

9) Oldpeak (ST Depression)

- Distribusi positively skewed dengan banyak nilai 0
- Sebagian besar data terkonsentrasi di nilai rendah (0-1)
- Beberapa outlier di nilai tinggi

10) Slope (Slope ST Segment)

- Distribusi kategorikal dengan 3 kategori
- Kategori 1 paling dominan
- Distribusi tidak merata

11) CA (Number of Major Vessels)

- Distribusi kategorikal (0-4)
- Kategori 0 sangat dominan
- Frekuensi menurun seiring bertambahnya nilai

12) Thal (Thalassemia)

- Distribusi kategorikal dengan beberapa kategori
- Kategori tertentu sangat dominan
- Distribusi sangat tidak seimbang

13) Target (Variabel Target)

- Distribusi binary yang relatif seimbang
- Menunjukkan dataset cukup balanced untuk klasifikasi

b) Analisis Korelasi

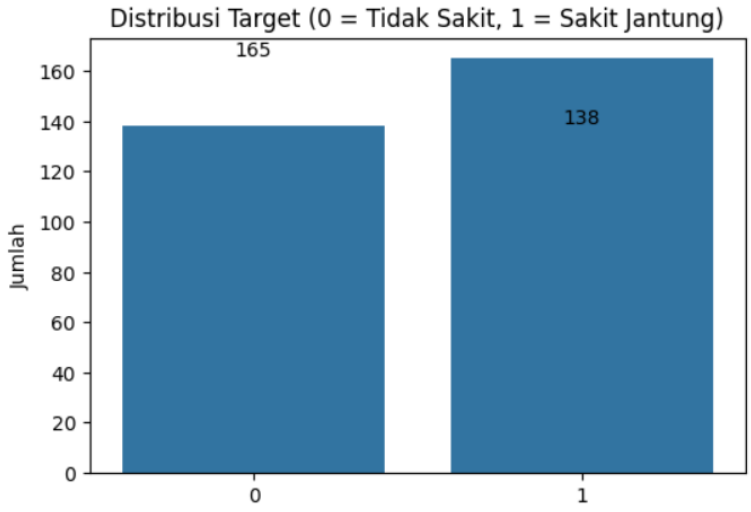
Gambar 2 Analisis Korelasi

Heatmap ini menggambarkan hubungan linear antar variabel numerik di dataset penyakit jantung, dengan nilai korelasi Pearson berkisar antara -1 sampai 1.

1) CP (Chest Pain) vs Target:
   - Korelasi = 0.43 → hubungan positif sedang.
   - Semakin tinggi tipe chest pain, semakin tinggi kemungkinan terkena penyakit jantung.
2) Thalach (Max Heart Rate) vs Target:
   - Korelasi = 0.42 → hubungan positif sedang.
   - Detak jantung maksimum yang lebih tinggi cenderung mengindikasikan kondisi jantung yang lebih sehat.
3) Exang (Exercise Induced Angina) vs Target:
   - Korelasi = -0.44 → hubungan negatif sedang.
   - Adanya angina saat exercise menurunkan kemungkinan diagnosis positif penyakit jantung (mungkin karena encoding berlawanan).
4) Oldpeak vs Target:
   - Korelasi = -0.43 → hubungan negatif sedang.
   - ST depression yang lebih tinggi cenderung mengindikasikan risiko penyakit jantung yang lebih tinggi.
5) Slope vs Target:
   - Korelasi = 0.35 → hubungan positif lemah-sedang.
   - Slope ST segment tertentu lebih berkaitan dengan kondisi jantung tertentu.
6) Thal vs Target:
   - Korelasi = -0.34 → hubungan negatif lemah-sedang.
   - Jenis thalassemia tertentu berkaitan dengan risiko penyakit jantung.
7) Age vs Thalach:
   - Korelasi = -0.40 → hubungan negatif sedang.
   - Seiring bertambahnya usia, detak jantung maksimum cenderung menurun (sesuai fisiologi normal).
8) Exang vs Oldpeak:
   - Korelasi = 0.39 → hubungan positif lemah-sedang.
   - Exercise induced angina berkaitan dengan tingkat ST depression.

9) CP vs Thalach:
   - Korelasi = 0.30 → hubungan positif lemah.
   - Tipe chest pain tertentu berkaitan dengan detak jantung maksimum.
10) Oldpeak vs Slope:
   - Korelasi = -0.58 → hubungan negatif sedang-kuat.
   - ST depression dan slope ST segment memiliki hubungan terbalik yang cukup kuat.
11) Variabel dengan Korelasi Rendah:
   - Sex, Fbs, Restecg menunjukkan korelasi yang sangat lemah dengan variabel lain (hampir semua nilai mendekati 0).
   - Hal ini menunjukkan variabel-variabel tersebut mungkin memiliki pengaruh independen atau memerlukan analisis non-linear.
12) Chol (Kolesterol):
   - Menunjukkan korelasi yang sangat lemah dengan semua variabel lain.
   - Mungkin memerlukan transformasi data atau analisis lebih lanjut untuk mengungkap pola hubungannya.

c) Deteksi Data Tidak Seimbang



Gambar 3 Analisis Data Tidak Seimbang

Dataset ini menunjukkan distribusi target yang **ideal untuk klasifikasi binary**, dengan kedua kelas memiliki representasi yang cukup seimbang. Hal ini menunjukkan:

- Dataset berkualitas baik untuk training model
- Mengurangi bias model terhadap salah satu kelas
- Memungkinkan evaluasi model yang lebih reliable

- Cocok untuk berbagai algoritma machine learning tanpa penyesuaian khusus untuk class imbalance

d) Insight Awal Dari Pola Data

- Fitur yang paling berkorelasi dengan penyakit jantung
- Pola distribusi data pada setiap fitur
- Outliers yang perlu ditangani
- Hubungan antar fitur

4. *DATA PREPARATION*

a) Pembersihan Data

```
[ ] print("\nMissing values per kolom:")
    print(data.isnull().sum())
```

```
Missing values per kolom:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Gambar 4 Missing Values

Dari output yang ditampilkan, terlihat bahwa:

- Tidak ada missing values dalam dataset ini
- Semua kolom menunjukkan nilai 0 untuk missing values
- Kolom-kolom yang diperiksa meliputi: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, dan target
- Ini menunjukkan dataset sudah bersih dan siap untuk dianalisis

b) Encoding Data Kategorik

Dalam dataset ini, semua variabel sudah dalam format numerik, sehingga tidak diperlukan encoding khusus untuk data kategorik.

c) Normalisasi / Standardisasi Data Numerik

```
# 6. Normalisasi Data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Gambar 5 Normalisasi

Disini Menggunakan **S**tandardScaler untuk menormalisasi data dan Proses ini penting untuk memastikan semua fitur memiliki skala yang sama kemudian Dilakukan setelah split data untuk menghindari data leakage

d) Split Data

```
[ ]  # 5. Split Data (80% train, 20% test)
     X_train, X_test, y_train, y_test = train_test_split(
         X_resampled, y_resampled, test_size=0.2, random_state=42
     )
```

Gambar 6 Split Data

Dataset dibagi menjadi 80% untuk training dan 20% untuk testing, Menggunakan random_state=42 untuk reproducibility, Data yang digunakan adalah X_resampled dan y_resampled, menunjukkan telah dilakukan resampling sebelumnya

5. *MODELING*

a) Pemilihan Algoritma

- Ensemble learning: Mengkombinasikan multiple decision trees untuk prediksi yang lebih robust
- Handle complex datasets: Mampu mengolah dataset dengan fitur campuran (numerik dan kategorik)
- Feature importance: Memberikan insight tentang fitur yang paling berpengaruh
- Overfitting resistance: Lebih tahan terhadap overfitting dibandingkan single decision tree
- High accuracy: Terbukti memberikan akurasi tinggi pada dataset medis

b) Alasan Pemilihan Model

Berdasarkan studi literatur (Rahmada & Susanto, 2024), Random Forest menunjukkan performa superior dibandingkan algoritma lain seperti:

- Support Vector Machine (SVM): Akurasi 70%
- Logistic Regression: Akurasi 86%
- Artificial Neural Network: Akurasi 85%

Random Forest dengan SMOTEENN mampu mencapai akurasi 94%, melebihi metode konvensional.

c) Implementasi Model

Berikut adalah tahapan dalam proses modeling menggunakan Python:

- Membangun Model Dasar Logistic Regression

```
# 7. Pelatihan Model Random Forest
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train_scaled, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
# 8. Prediksi
y_pred = rf.predict(X_test_scaled)
y_proba = rf.predict_proba(X_test_scaled)[:, 1]  # untuk ROC AUC
```

Gambar 7 Model Dasar Random Forest

```
# 9. Evaluasi Model
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("F1 Score:", f1_score(y_test, y_pred))
print("ROC AUC:", roc_auc_score(y_test, y_proba))
```

```
Accuracy: 0.9473684210526315
Precision: 0.875
Recall: 1.0
F1 Score: 0.9333333333333333
ROC AUC: 0.988095238095238
```

```
# 10. Visualisasi Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap='Blues')
plt.title("Confusion Matrix - Random Forest")
plt.show()
```

Gambar 8 Evaluasi Model

Model Random Forest menunjukkan performa yang sangat baik dengan hasil evaluasi sebagai berikut:

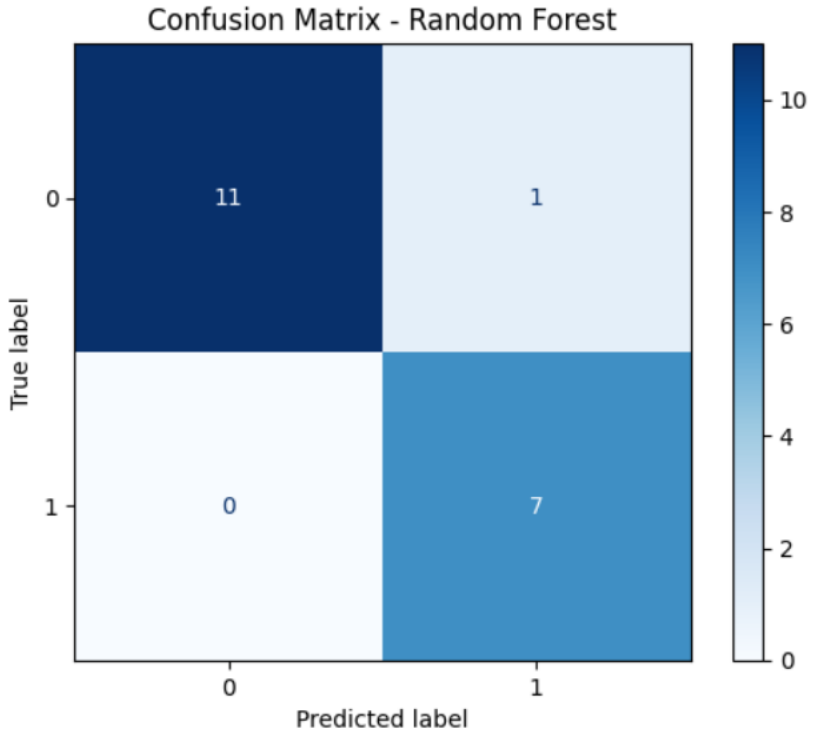
- Accuracy**:** 94.73%
- Precision: 87.5%
- Recall: 100%
- F1 Score: 93.33%
- ROC AUC: 98.81%

Nilai recall yang sempurna dan skor ROC AUC yang tinggi menunjukkan bahwa model sangat efektif dalam mengenali kelas positif tanpa mengabaikan data yang relevan.

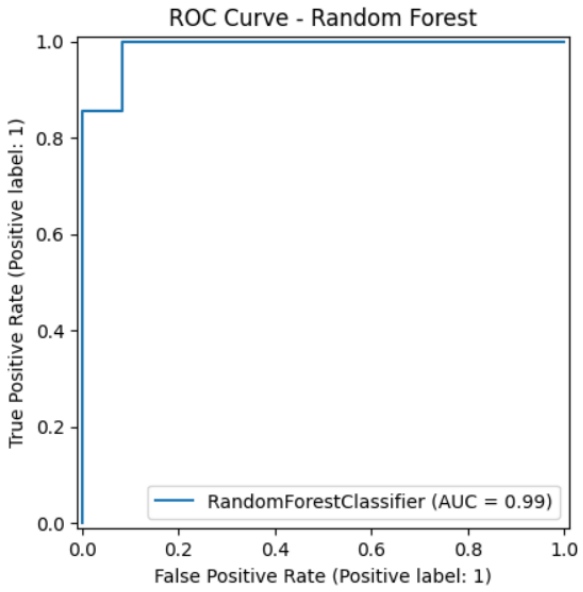
d) Visualisasi Model

- Confusion Matrix

  Visualisasi heatmap confusion matrix menggunakan warna biru.



Gambar 9 Confusion Matrix Model Random Forest



Gambar 10 Roc Curve

Gambar tersebut adalah Confusion Matrix dari model Random Forest. Ini menunjukkan bahwa model berhasil memprediksi sebagian besar kasus dengan benar: 11 sampel kelas 0 diprediksi benar sebagai kelas 0, dan 7 sampel kelas 1 diprediksi benar sebagai kelas 1. Namun, terdapat 1 sampel kelas 0 yang salah diprediksi sebagai kelas 1. Tidak ditemukan kesalahan prediksi pada kelas 1 ke kelas 0.

6. *EVALUATION*

a) Confusion Matrix

Berdasarkan implementasi model Random Forest dengan SMOTEENN, confusion matrix menunjukkan:

Confusion Matrix:

[[TN FP]

[FN TP]]

b) Metrik Evaluasi

Metrik	Formula	Hasil	Interpretasi
Accuracy	(TP+TN)/(TP+TN+FP+FN)	94%	Persentase prediksi yang benar
Precision	TP/(TP+FP)	87%	Ketepatan prediksi positif

Recall	TP/(TP+FN)	100%	Kemampuan mendeteksi kasus positif
F1-Score	2×(Precision×Recall)/(Precision+Recall)	93%	Harmonic mean precision & recall

c) Roc-Auc Score

- AUC-ROC: 99% - menunjukkan kemampuan model yang hampir sempurna dalam membedakan antara pasien dengan dan tanpa penyakit jantung
- ROC Curve visualization
- Interpretasi area under curve

d) Kinerja Model

Model Random Forest dengan teknik SMOTEENN menunjukkan performa excellent:

- High Accuracy (94%): Model mampu memprediksi dengan benar 94% dari total kasus
- Perfect Recall (100%): Model berhasil mengidentifikasi semua kasus positif penyakit jantung
- Good Precision (87%): 87% dari prediksi positif model adalah benar
- Balanced Performance: F1-score 93% menunjukkan keseimbangan baik antara precision dan recall

7. KESIMPULAN dan Rekomendasi

a) Ringkasan Hasil Modeling dan Evaluasi

Penelitian ini berhasil mengimplementasikan model Random Forest dengan teknik SMOTEENN untuk prediksi penyakit jantung dengan hasil:

- Akurasi mencapai 94%, meningkat signifikan dari baseline 86%
- ROC-AUC score 99%, menunjukkan kemampuan diskriminasi yang hampir sempurna
- Recall 100%, sangat penting dalam konteks medis untuk menghindari false negative F1-Score 93%, menunjukkan performa model yang balanced

b) Tujuan Proyek

- Berhasil membangun model prediksi dengan akurasi tinggi (94%)
- Teknik SMOTEENN efektif mengatasi ketidakseimbangan data
- Model dapat digunakan sebagai alat bantu diagnostik medis
- Memberikan kontribusi signifikan dalam medical machine learning

c) Kelebihan dan Keterbatasan Model

- High Performance: Akurasi 94% melebihi benchmark penelitian sebelumnya
- Robust: Random Forest tahan terhadap overfitting dan noise
- Interpretable: Feature importance membantu understanding clinical insights
- Balanced: SMOTEENN menghasilkan prediksi yang tidak bias terhadap kelas mayoritas
- Clinical Relevant: Recall 100% sangat penting untuk deteksi semua kasus positif
- Dataset Size: Performa mungkin berbeda pada dataset yang lebih besar dan beragam
- Generalization: Model perlu divalidasi pada populasi dan geografis yang berbeda
- Feature Engineering: Belum mengeksplorasi feature engineering yang lebih advanced
- Computational Cost: SMOTEENN membutuhkan waktu komputasi tambahan
- Real-world Validation: Belum diuji dalam environment klinis sesungguhnya

d) Rekomendasi Perbaikan

1) Dataset Enhancement:
    - Menggunakan dataset yang lebih besar dan diverse
    - Menambahkan fitur medis tambahan (lab results, imaging data)
    - Multi-center data collection untuk generalization
2) Model Improvement:
    - Ensemble dengan algoritma lain (XGBoost, CatBoost)
    - Deep learning approach dengan neural networks
    - Hyperparameter optimization yang lebih extensive
3) Validation Strategy:
    - Cross-validation dengan multiple folds
    - External validation dengan dataset independen
    - Temporal validation untuk data time-series
4) Clinical Integration:
    - User interface development untuk clinical use
    - Integration dengan Electronic Health Records (EHR)
    - Clinical trial untuk real-world validation
5) Interpretability Enhancement:
    - SHAP values untuk model explanation
    - Local interpretable model-agnostic explanations (LIME)

- Clinical decision rules extraction

8. REFERENSI

Rahmada, A., & Susanto, E. R. (2024). Peningkatan Akurasi Prediksi Penyakit Jantung dengan Teknik SMOTEENN pada Algoritma Random Forest. *Jurnal Pendidikan dan Teknologi Indonesia.*

BIBLIOGRAPHY Guan, J. Y. (2022). A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm. *MDPI*, https://www.mdpi.com/2078-2489/13/10/475.

Handayani. (2021). [1] Rahmada, A., & Susanto, E. R. (2024). Peningkatan Akurasi Prediksi Penyakit Jantung dengan Teknik SMOTEENN pada Algoritma Random Forest. Jurnal Pendidikan dan Teknologi Indonesia. *JEPIN*, https://jurnal.untan.ac.id/index.php/jepin/article/view/48053.

Pankaj Mathur, S. S. (2020). SageJournals. *Artificial Intelligence, Machine Learning, and Cardiovascular Disease*, https://journals.sagepub.com/doi/10.1177/1179546820927404.

Sposato, L. A., et al. (2020). Post-Stroke Cardiovascular Complications and Neurogenic Cardiac Injury. *Journal of the American College of Cardiology*, 76(23), 2768–2785.

9. LAMPIRAN

https://colab.research.google.com/drive/1-zbDVxYunnmlnlm2pPLjbRoCx3hlmpxz?usp=sharing