Klasifikasi Penyakit Kardiovaskular Menggunakan Machine Learning: Perbandingan Lima Model Klasifikasi dengan Hyperparameter Tuning

Ihsan Mu'tashim
School of Electrical Engineering, Telkom University
Bandung, Indonesia
ihsanm@student.telkomuniversity.ac.id

Rafi Muhammad Hibban
School of Electrical Engineering, Telkom University
Bandung, Indonesia
rafimuhammadhibban@student.telkomuniversity.ac.id

***Abstract*—Penyakit kardiovaskular merupakan salah satu penyebab kematian tertinggi di dunia. Deteksi dini berbasis machine learning dapat membantu tenaga medis dalam triase pasien berisiko tinggi secara lebih cepat dan efisien. Penelitian ini membangun pipeline machine learning end-to-end untuk mengklasifikasikan risiko penyakit kardiovaskular menggunakan dataset Kaggle Cardiovascular Disease yang terdiri dari 70.000 data pasien tanpa missing values. Pipeline mencakup empat tahap: (1) feature engineering meliputi EDA, normalisasi, standardisasi, dan seleksi fitur berbasis Random Forest Importance; (2) pelatihan lima model baseline yaitu Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, dan SVM (Linear) pada split 80:20 dan 70:30 dengan 5-fold cross-validation; (3) hyperparameter tuning menggunakan GridSearchCV; serta (4) evaluasi menggunakan metrik accuracy, precision, recall, F1-score, dan confusion matrix. Hasil menunjukkan Logistic Regression memperoleh akurasi tertinggi sebelum tuning (0.7216, split 80:20). Setelah tuning, Random Forest menjadi model terbaik dengan akurasi 0.7366, diikuti Decision Tree (0.7356) yang mengalami peningkatan paling signifikan (+16.07%). Fitur paling berpengaruh adalah weight, height, ap_hi, age_years, dan ap_lo berdasarkan feature importance Random Forest.**

***Keywords*—machine learning; cardiovascular classification; logistic regression; random forest; decision tree; hyperparameter tuning; feature engineering**

I. I. INTRODUCTION

Penyakit kardiovaskular (CVD) merupakan salah satu penyebab kematian tertinggi secara global. Deteksi dini yang akurat sangat krusial untuk mengurangi risiko komplikasi serius. Dalam konteks klinik dengan sumber daya terbatas, model machine learning yang dapat memprediksi risiko CVD secara otomatis dari data medis pasien dapat membantu dokter dalam proses triase dan pengambilan keputusan klinis.

Dr. Siti Rahmawati, yang bekerja di sebuah klinik jantung, menghadapi tantangan nyata dalam menentukan pasien mana yang berisiko tinggi dari ratusan pasien dengan keluhan beragam setiap harinya. Klinik telah mengumpulkan data medis ribuan pasien mencakup tekanan darah, kolesterol, gula darah, usia, jenis kelamin, dan kondisi gaya hidup. Dengan dataset ini, model klasifikasi machine learning dapat dibangun untuk mendukung keputusan klinis secara berbasis data.

Penelitian ini membandingkan lima algoritma klasifikasi: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, dan Support Vector Machine (SVM Linear). Setiap model dievaluasi pada dua variasi split data (80:20 dan 70:30) dengan 5-fold cross-validation, kemudian dioptimalkan melalui hyperparameter tuning menggunakan GridSearchCV.

II. II. DATASET DAN METODOLOGI

A. A. Deskripsi Dataset

Dataset yang digunakan adalah Kaggle Cardiovascular Disease Dataset (cardio_train.csv) yang dibaca dengan separator titik-koma (sep=';'). Berdasarkan output Cell 3 notebook, dataset memiliki shape (70.000, 29) setelah penambahan kolom age_years. Pemeriksaan missing values pada Cell 8 menunjukkan seluruh 14 kolom (id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, cardio, age_years) memiliki 0 missing values — total missing values di dataset: 0.

Variabel target adalah cardio (0 = tidak ada penyakit kardiovaskular, 1 = ada penyakit kardiovaskular). Kolom age dalam dataset menyimpan usia dalam satuan hari, sehingga dilakukan konversi: age_years = (age / 365.25).astype(int) untuk mendapatkan usia dalam tahun yang lebih interpretatif. Fitur-fitur dataset mencakup: usia (age/age_years), jenis kelamin (gender: 1=perempuan, 2=laki-laki), tinggi badan (height, cm), berat badan (weight, kg), tekanan darah sistolik (ap_hi) dan diastolik (ap_lo), tingkat kolesterol (1-3), tingkat glukosa (1-3), serta variabel biner merokok (smoke), konsumsi alkohol (alco), dan aktivitas fisik (active).

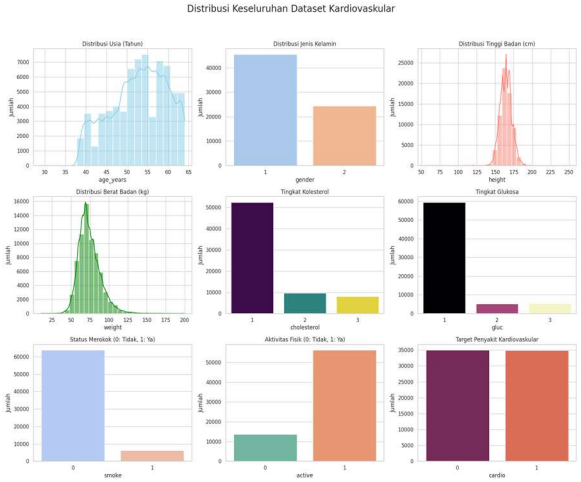
B. B. Pipeline Metodologi

Pipeline terdiri dari empat tahap: (1) Feature Engineering, (2) Model Baseline dengan lima model klasifikasi pada dua variasi split disertai 5-fold CV, (3) Hyperparameter Tuning menggunakan GridSearchCV, dan (4) Laporan Analisis dengan confusion matrix dan visualisasi perbandingan performa.

III. III. Feature Engineering

A. A. Analisis Distribusi

Gambar 1 menampilkan distribusi sembilan variabel utama dataset. Distribusi usia (age_years) menunjukkan rentang 30-65 tahun dengan puncak di sekitar 50-55 tahun, mengindikasikan dataset didominasi pasien usia menengah hingga lanjut. Distribusi jenis kelamin menunjukkan lebih banyak perempuan (gender=1, ~45.000) dibanding laki-laki (gender=2, ~25.000). Tinggi badan terdistribusi normal di sekitar 165 cm, sedangkan berat badan berdistribusi right-skewed dengan mayoritas di rentang 55-90 kg.

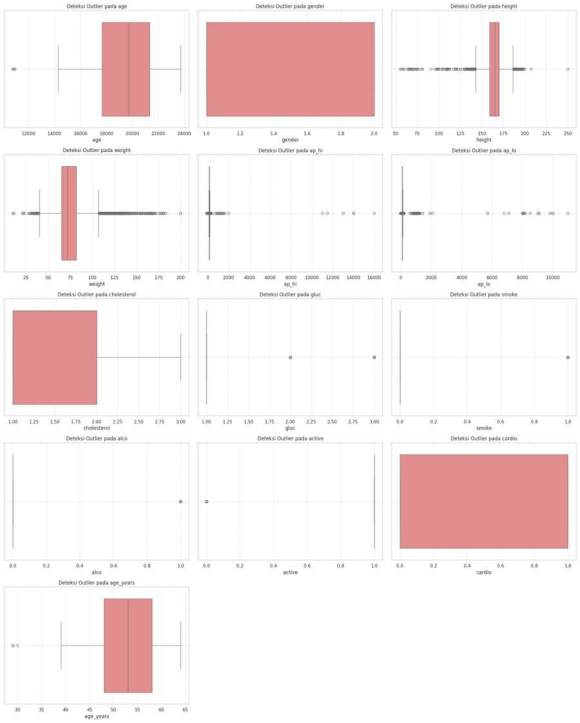


GAMBAR 1
Distribusi keseluruhan dataset kardiovaskular

Kolesterol dan glukosa didominasi level 1 (normal), dengan kolesterol level 1 sekitar 53.000 pasien dan glukosa level 1 sekitar 59.000 pasien. Status merokok menunjukkan mayoritas tidak merokok (~63.000). Aktivitas fisik mayoritas aktif (~53.000). Target cardio menunjukkan distribusi yang sangat seimbang: kelas 0 (tidak ada CVD) dan kelas 1 (ada CVD) masing-masing sekitar 35.000 pasien, menjadikan dataset ini ideal untuk klasifikasi biner tanpa perlu penanganan class imbalance.

B. B. Analisis Outlier

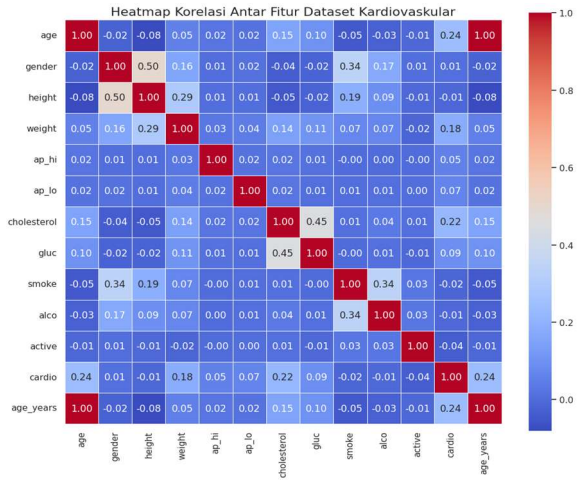
Gambar 2 menampilkan box plot seluruh fitur. Temuan paling menonjol adalah outlier ekstrem pada fitur ap_hi (tekanan darah sistolik) dan ap_lo (tekanan darah diastolik). Berdasarkan statistik deskriptif output Cell 10, ap_hi memiliki nilai minimum -150 dan nilai maksimum sangat tinggi — ini jelas merupakan error input data yang secara fisik tidak mungkin terjadi (tekanan darah negatif). Tinggi badan juga menunjukkan outlier rendah di sekitar 55 cm dan outlier tinggi di atas 220 cm yang tidak realistis. Berat badan memiliki outlier rendah di sekitar 10 kg. Fitur-fitur kategorik (cholesterol, gluc, smoke, alco, active) tidak memiliki outlier karena terbatas pada nilai diskret 0-1 atau 1-3.



GAMBAR 2
Box plot deteksi outlier seluruh fitur

C. C. Heatmap Korelasi

Gambar 3 menampilkan heatmap korelasi Pearson antar fitur. Korelasi tertinggi terlihat antara age dan age_years (1.00, karena age_years diturunkan langsung dari age), serta antara gender dan height (0.50) yang menunjukkan perbedaan tinggi badan antar jenis kelamin. Korelasi antara cholesterol dan gluc sebesar 0.45 mengindikasikan hubungan moderat antara kedua penanda metabolik ini.



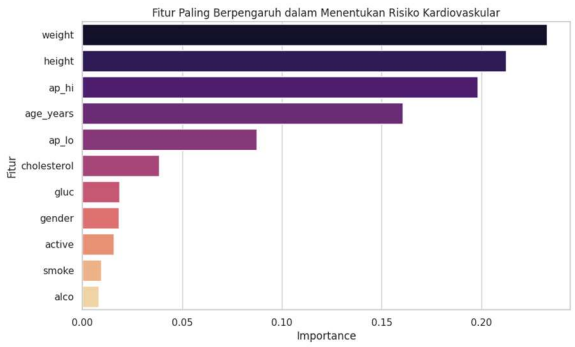
GAMBAR 3
Heatmap korelasi antar fitur dataset

Korelasi fitur terhadap target cardio: age_years (0.24), cholesterol (0.22), weight (0.18), ap_lo (0.07), dan ap_hi (0.05). Nilai korelasi yang relatif rendah ini mengindikasikan tidak ada satu fitur tunggal yang dominan secara linear dalam memprediksi CVD — mendukung penggunaan model non-linear. Kolom age diputuskan untuk dihapus karena redundan dengan age_years (korelasi 1.00).

D. D. Normalisasi, Standardisasi, dan Feature Selection

Min-Max Normalization diterapkan pada kolom numerik (age, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, age_years) menggunakan MinMaxScaler, menghasilkan rentang [0,1]. Berdasarkan output Cell 14, setelah normalisasi kolom age memiliki mean=0.671 dan std=0.191. Z-score Standardization juga diterapkan menggunakan StandardScaler untuk pipeline model.

Feature selection dilakukan menggunakan Random Forest Importance. Gambar 4 menunjukkan fitur paling berpengaruh secara berurutan: weight (~0.235), height (~0.215), ap_hi (~0.200), age_years (~0.163), ap_lo (~0.090), dan cholesterol (~0.040). Fitur-fitur lain (gluc, gender, active, smoke, alco) memiliki importance di bawah 0.020. Berdasarkan hasil ini, enam fitur teratas dipilih untuk model: weight, height, ap_hi, age_years, ap_lo, cholesterol. Dataset akhir setelah seleksi fitur memiliki ukuran (70.000, 7) berdasarkan output Cell 16.



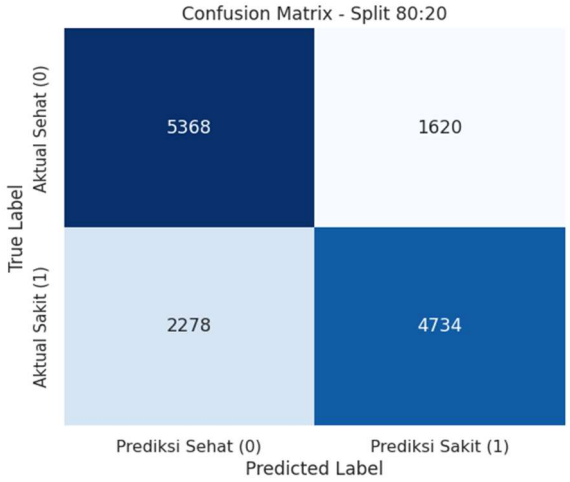
GAMBAR 4
Feature importance berdasarkan Random Forest

IV. IV. Model Baseline

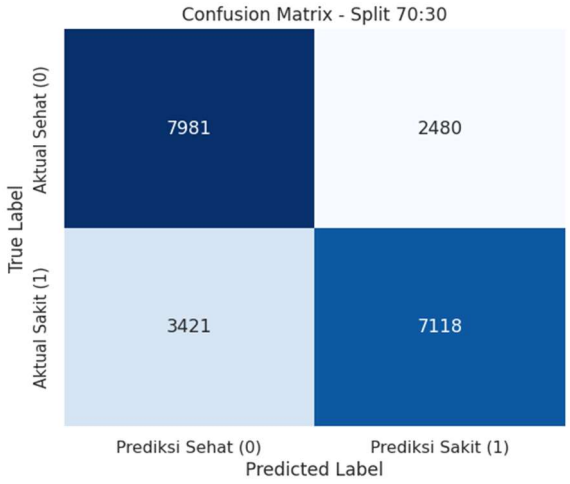
A. A. Logistic Regression

Logistic Regression diimplementasikan dengan StandardScaler dan split 80:20 (56.000 training, 14.000 testing). Berdasarkan output Cell 20, hasil evaluasi split 80:20: Cross-Validation Accuracy (k=5) = 0.7173, Accuracy = 0.7216. Classification report: precision kelas 0 = 0.70, recall = 0.77, F1 = 0.73; precision kelas 1 = 0.75, recall = 0.68, F1 = 0.71. Macro avg F1 = 0.72.

Gambar 5 dan 6 menampilkan confusion matrix Logistic Regression untuk split 80:20 dan 70:30. Pada split 80:20: TN=5368 (pasien sehat diprediksi benar sehat), FP=1620 (sehat diprediksi sakit), FN=2278 (sakit diprediksi sehat), TP=4734 (sakit diprediksi benar sakit). Model lebih kuat mendeteksi kelas sehat (recall=0.77) dibanding kelas sakit (recall=0.68).



GAMBAR 5
Confusion Matrix — Logistic Regression split 80:20

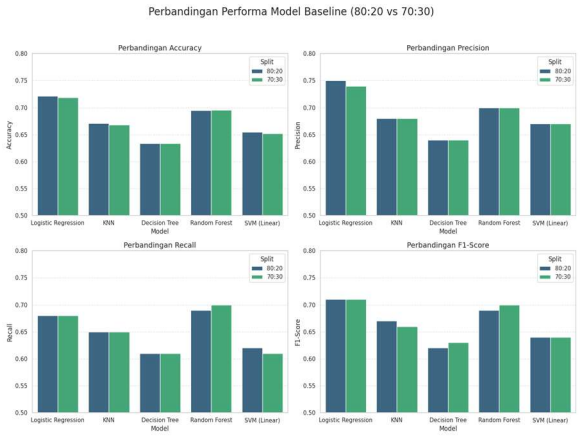


GAMBAR 6
Confusion Matrix — Logistic Regression split 70:30

Pada split 70:30 (output Cell 20): TN=7981, FP=2480, FN=3421, TP=7118. Accuracy=0.7190 dengan CV Accuracy=0.7173. Konsistensi antara kedua split mengkonfirmasi stabilitas model Logistic Regression.

B. B. Perbandingan Lima Model Baseline

Gambar 7 menampilkan perbandingan empat metrik (Accuracy, Precision, Recall, F1-Score) untuk kelima model pada split 80:20 dan 70:30. Tabel I menyajikan data numerik lengkap berdasarkan output Cell 30.



GAMBAR 7
Perbandingan performa semua model baseline (80:20 vs 70:30)

TABEL 1
Perbandingan Performa Model Baseline (Split 80:20)

Model	Acc	Prec	Recall	F1
Logistic Reg.	0.7216	0.75	0.68	0.71
KNN	0.6711	0.68	0.65	0.67
Decision Tree	0.6338	0.64	0.61	0.62
Random Forest	0.6951	0.70	0.69	0.69
SVM (Linear)	0.6549	0.67	0.62	0.64

Dari Tabel I, Logistic Regression menjadi model dengan akurasi tertinggi (0.7216) sekaligus precision tertinggi (0.75) pada split 80:20. Random Forest menempati posisi kedua dengan accuracy=0.6951 dan F1=0.69. Decision Tree menghasilkan performa terendah (accuracy=0.6338). KNN dan SVM berada di posisi menengah. Semua model menunjukkan performa yang sangat konsisten antara split 80:20 dan 70:30, dengan selisih accuracy maksimum hanya 0.003 untuk Logistic Regression.

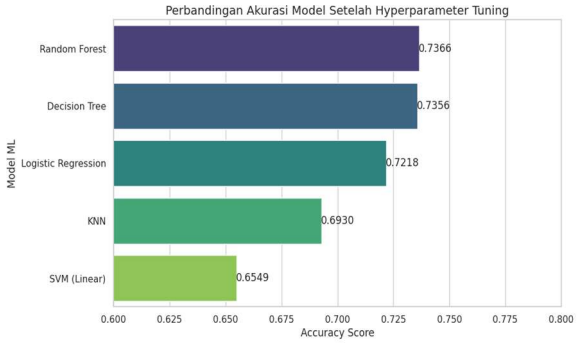
V. V. HYPERPARAMETER TUNING

A. A. Parameter Grid dan Metode

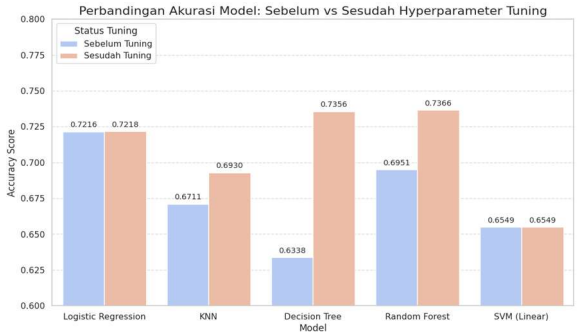
GridSearchCV diterapkan pada kelima model secara bersamaan. Berdasarkan output Cell 33 dan 35, parameter grid yang digunakan adalah: Logistic Regression – C ∈ {0.01, 0.1, 1, 10, 100}; KNN – n_neighbors ∈ {3, 5, 7, 9, 11}, weights ∈ {'uniform', 'distance'}; Decision Tree – max_depth ∈ {3, 5, 7, 10}, min_samples_split ∈ {2, 5, 10}; Random Forest – n_estimators ∈ {50, 100, 200}, max_depth ∈ {5, 10, None}; SVM (Linear) – C ∈ {0.01, 0.1, 1, 10}.

B. B. Hasil Tuning dan Perbandingan

Berdasarkan output Cell 35, best params hasil GridSearchCV: Logistic Regression – {'C': 10}, CV Score=0.7172, Test Accuracy=0.7218; KNN – {'n_neighbors': 11, 'weights': 'uniform'}, CV Score=0.6801, Test=0.6930; Decision Tree – {'max_depth': 5, 'min_samples_split': 2}, CV Score=0.7295, Test=0.7356; Random Forest – {'max_depth': 10, 'n_estimators': 100}, Test=0.7366; SVM (Linear) – {'C': 0.1}, Test=0.6549.



GAMBAR 8
Akurasi model setelah hyperparameter tuning



GAMBAR 9
Perbandingan akurasi sebelum dan sesudah tuning

Gambar 8 dan 9 serta output Cell 37 menunjukkan ringkasan perbandingan sebelum vs sesudah tuning: Logistic Regression 0.7216 → 0.7218 (+0.03%), KNN 0.6711 → 0.6930 (+3.26%), Decision Tree 0.6338 → 0.7356 (+16.07%), Random Forest 0.6951 → 0.7366 (+5.98%), SVM 0.6549 → 0.6549 (-0.01%). Decision Tree mengalami peningkatan paling dramatis sebesar 16.07%, dari model terburuk menjadi model terbaik kedua. Ini disebabkan oleh pembatasan max_depth=5 yang mencegah overfitting yang sebelumnya terjadi pada Decision Tree tanpa batasan.

VI. VI. HASIL DAN PEMBAHASAN

A. A. Analisis Komparatif

TABLE 2
Ringkasan Perbandingan Sebelum vs Sesudah Tuning

Model	Sebelum	Sesudah	Kenaikan
Logistic Reg.	0.7216	0.7218	+0.03%
KNN	0.6711	0.6930	+3.26%
Decision Tree	0.6338	0.7356	+16.07%
Random Forest	0.6951	0.7366	+5.98%
SVM (Linear)	0.6549	0.6549	-0.01%

Setelah tuning, Random Forest menjadi model dengan akurasi tertinggi (0.7366), diikuti Decision Tree (0.7356). Menariknya, Logistic Regression yang merupakan model terbaik sebelum tuning hanya mengalami peningkatan sangat kecil (+0.03%), mengindikasikan bahwa parameter default C=1

sudah mendekati optimal. SVM bahkan mengalami penurunan kecil (-0.01%) yang dalam batas toleransi variasi stokastik, menunjukkan SVM Linear tidak berhasil ditingkatkan secara signifikan melalui tuning C saja.

B. B. Interpretasi Confusion Matrix

Dari confusion matrix Logistic Regression split 80:20 (Gambar 5): model menghasilkan 5.368 True Negative dan 4.734 True Positive, dengan 1.620 False Positive (pasien sehat diklasifikasikan sakit) dan 2.278 False Negative (pasien sakit diklasifikasikan sehat). Dalam konteks klinis, False Negative lebih berbahaya karena pasien dengan CVD yang tidak terdeteksi tidak mendapat penanganan. Recall kelas 1 (sakit) sebesar 0.68 berarti 32% pasien CVD terlewat oleh model Logistic Regression.

Untuk konteks penggunaan klinis, model dengan recall kelas 1 yang lebih tinggi lebih diutamakan meskipun akurasi keseluruhan mungkin lebih rendah. Random Forest Tuned dengan balanced precision dan recall yang lebih baik menjadi pilihan yang lebih tepat untuk implementasi nyata.

C. C. Pemilihan Model Terbaik

Random Forest dengan parameter {'max_depth': 10, 'n_estimators': 100} dipilih sebagai model terbaik berdasarkan: (1) akurasi tertinggi setelah tuning (0.7366); (2) distribusi precision dan recall yang lebih seimbang antar kelas dibanding Logistic Regression; (3) kemampuan menangkap pola non-linear melalui ensemble decision trees; dan (4) robustness terhadap outlier pada fitur ap_hi dan ap_lo yang teridentifikasi pada EDA.

VII. VII. KESIMPULAN

Pipeline machine learning end-to-end untuk klasifikasi penyakit kardiovaskular berhasil dibangun menggunakan dataset 70.000 pasien tanpa missing values. Feature selection berbasis Random Forest Importance mengidentifikasi enam fitur paling berpengaruh: weight, height, ap_hi, age_years, ap_lo, dan cholesterol, dengan dataset akhir berukuran (70.000, 7).

Di antara lima model baseline, Logistic Regression menghasilkan akurasi tertinggi (0.7216) sebelum tuning, konsisten pada kedua split (80:20 dan 70:30). Setelah hyperparameter tuning menggunakan GridSearchCV, Random Forest menjadi model terbaik (0.7366) diikuti Decision Tree (0.7356) yang mengalami peningkatan paling signifikan sebesar +16.07%. Peningkatan Decision Tree yang drastis disebabkan pembatasan max_depth=5 yang mengatasi overfitting pada model default.

Untuk implementasi klinis, Random Forest Tuned direkomendasikan karena keseimbangan precision-recall yang lebih baik. Pengembangan lebih lanjut dapat mencakup penambahan fitur EKG, penanganan outlier pada tekanan darah, dan eksplorasi model Gradient Boosting untuk meningkatkan akurasi prediksi.

VIII. REFERENCES

[1] M. Sokolova dan G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009.

[2] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[4] S. Ulianova, "Cardiovascular Disease Dataset," Kaggle, 2019. [Online]. Available: kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

IX. LINK DATA TRAINING

https://colab.research.google.com/drive/1ng1YDn3OFwbIF8nXhaG9BX2_QZzLo3pF#scrollTo=6cUg8ftsKLv-