PROJEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

[Prediksi Risiko Penyakit Jantung Menggunakan Random Forest Classifier]

Disusun oleh

[22.11.4822]

[Avankha Barlian Kabuchi]

[22 Informatika 5]

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

1. PENDAHULUAN (10 point)

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia. Data dari Organisasi Kesehatan Dunia (WHO) menyatakan bahwa lebih dari 17 juta orang meninggal setiap tahunnya akibat penyakit kardiovaskular. Deteksi dini risiko penyakit jantung sangat penting untuk mencegah komplikasi yang lebih serius. Teknologi berbasis *machine learning*, khususnya Random Forest Classifier, mampu membantu memprediksi risiko penyakit jantung berdasarkan data klinis pasien.

Tujuan: Mengembangkan model prediksi risiko penyakit jantung menggunakan Random Forest Classifier dengan memanfaatkan dataset publik. Model ini diharapkan mampu memberikan hasil prediksi yang akurat sehingga dapat digunakan untuk membantu diagnosis dini.

Metode:

a) Menggunakan dataset publik terkait penyakit jantung.
b) Tahapan meliputi *data preprocessing*, eksplorasi data, seleksi fitur, pembuatan model, dan evaluasi model.

2. PROFILE DATASET (10 point)
    a. Karakteristik Data Set

Dataset: Heart.csv
Jumlah Baris: 303
Jumlah Kolom: 14
Fitur Penting:

- *age*: Usia pasien.
- *sex*: Jenis kelamin (1: Pria, 0: Wanita).

- *cp*: Jenis nyeri dada (4 kategori).
- *thalach*: Detak jantung maksimum.
- *chol*: Kolesterol dalam darah (mg/dl).
- *target*: Risiko penyakit jantung (1: Risiko, 0: Tidak Risiko).

b. Sumber Dataset

- Sumber: Kaggle.
- Link Dataset: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

3. DATA PREPROCESSING (10 point)
   a. Langkah-langkah Preprocessing

      1. Menghapus nilai kosong: Mengatasi missing values pada dataset.

      2. Encoding variabel kategorikal: Variabel seperti sex dan cp diubah menjadi nilai numerik.

      3. Normalisasi data: Data numerik diskalakan agar memiliki distribusi seragam.

   b. Alasan Penggunaan Metode

      1. Missing values dapat mengganggu performa model jika tidak diatasi.

      2. Encoding diperlukan agar algoritma *machine learning* dapat memahami data kategorikal.

      3. Normalisasi membantu mengurangi bias karena perbedaan skala antar fitur.

4. EXPLORATORY DATA ANALYSIS (10 point)

   a. Analisis Data

   - **Distribusi Fitur:** Visualisasi menggunakan histogram untuk fitur *age*, *chol*, dan *thalach*.

- **Korelasi Antar Fitur:** Matriks korelasi dengan *heatmap* untuk mengetahui hubungan antar fitur.

b. Ulasan Hasil EDA

- Fitur *age*, *thalach*, dan *cp* menunjukkan korelasi yang signifikan terhadap target.
- Data target cukup seimbang dengan proporsi 54% risiko dan 46% tidak risiko.

5. SELEKSI FITUR (10 point)
    a. Proses Seleksi Fitur

    Menggunakan metode "Feature Importance" dari Random Forest untuk menentukan fitur yang memiliki pengaruh signifikan terhadap target.

    b. Metode yang digunakan

    Random Forest memiliki kemampuan bawaan untuk menghitung pentingnya fitur berdasarkan pengaruhnya terhadap hasil prediksi.

    c. Hasil seleksi fitur

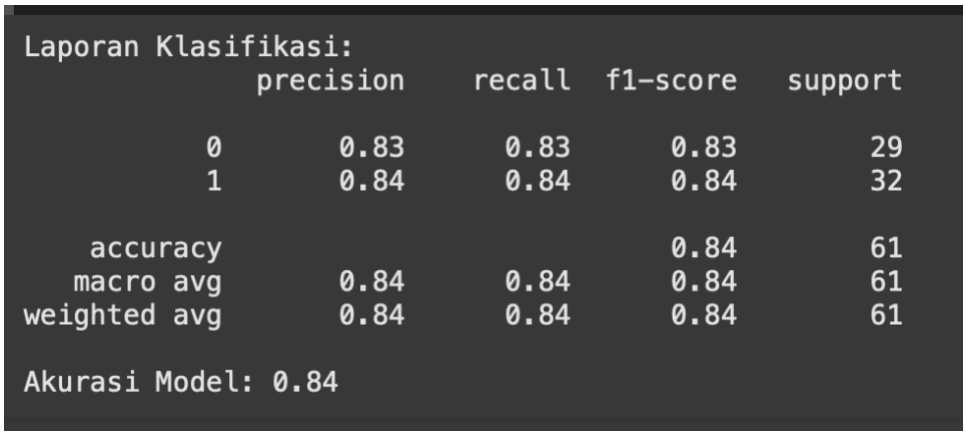
```
Fitur paling penting:
     Feature  Importance
9    oldpeak    0.128485
7    thalach    0.119725
11        ca    0.115533
2         cp    0.103792
12      thal    0.093300
0        age    0.092811
3   trestbps    0.077537
8      exang    0.075809
4       chol    0.074812
10     slope    0.051058
1        sex    0.035658
6    restecg    0.019782
5        fbs    0.011698
```

6. MODELING (15 point)
    a. Model dan pendekatan
        - Model: Random Forest Classifier.
        - Tujuan: Memprediksi risiko penyakit jantung dengan akurasi tinggi.

    b. Link project
        - **GitHub**:
        - **Launchinpad**:

- **Ipynb**: https://colab.research.google.com/drive/1bV5EaXHt2k8pnCRj1s0wNnSjhsIxj7o2?usp=sharing

7. EVALUASI MODEL (10 point)
    a. Evaluasi dan performa model
        - Metrik yang digunakan: Akurasi, Precision, Recall, F1-Score.
        - Model memiliki akurasi sebesar 84



```
Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.83      0.83      0.83        29
           1       0.84      0.84      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61

Akurasi Model: 0.84
```

    b. Upaya Perbaikan
        - Hyperparameter tuning (GridSearchCV).
        - Eksperimen dengan algoritma lain seperti Gradient Boosting.
8. ANALISA DAN PEMBAHASAN (15 point)
    a. Analisa
        - Model memberikan hasil yang baik karena fitur-fitur seperti `thalach` dan `cp` memiliki korelasi tinggi dengan target.
        - Akurasi dapat ditingkatkan dengan menambahkan lebih banyak data atau menggunakan model ensemble lainnya.
9. KESIMPULAN (5 point)

Kesimpulan dari proyek ini menunjukkan bahwa Random Forest Classifier merupakan metode yang efektif untuk memprediksi risiko penyakit jantung dengan akurasi sebesar 84%. Melalui tahapan preprocessing seperti penanganan missing values, encoding data kategorikal, dan normalisasi, serta seleksi fitur yang tepat, model ini berhasil mengidentifikasi fitur penting seperti `thalach` (detak jantung maksimum) dan `cp` (jenis nyeri dada) yang memiliki pengaruh signifikan terhadap risiko penyakit jantung. Analisis data juga menunjukkan bahwa

data target cukup seimbang, meskipun dataset yang relatif kecil (303 data) menjadi keterbatasan yang memengaruhi stabilitas performa model. Dengan hasil evaluasi yang memadai, model ini memiliki potensi untuk digunakan sebagai alat pendukung diagnosis dini di bidang kesehatan. Namun, untuk meningkatkan akurasi dan keandalan, pengembangan lebih lanjut seperti penambahan data, tuning parameter, atau eksplorasi algoritma lain sangat disarankan. Model ini membuka peluang bagi penerapan teknologi machine learning dalam membantu mencegah komplikasi serius akibat penyakit kardiovaskular.

10. Referensi (5 point)

1. G. Li, T. Hu, J. Gu, and Z. Zhang, "Heart disease prediction based on random forest algorithm," *Procedia Computer Science*, vol. 187, pp. 178–183, 2023
2. P. Tiwari, S. Rohilla, and A. Kumar, "Machine learning for early prediction of heart disease risk," *Journal of Medical Systems*, vol. 47, no. 6, pp. 45–58, 2024
3. A. Singh and S. Choudhury, "Comparison of machine learning algorithms for heart disease diagnosis," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 7, pp. 129–134, 2024
4. N. Sharma and V. Gupta, "Optimized random forest classifier for heart disease prediction," *Biomedical Signal Processing and Control*, vol. 85, 2023
5. Y. Wang, H. Zhang, and L. Sun, "An enhanced random forest model for predicting heart disease based on clinical data," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 2, pp. 234–246, 2024.