PROJEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

Penerapan Algoritma XGBoost Classifier untuk Klasifikasi Penyakit Jantung Koroner

Disusun oleh

Nama: Lutfiana Deka Nurhayati

NIM: 22.11.4975

Kelas: 22 Informatika 7

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

1. PENDAHULUAN

Penyakit jantung koroner terjadi akibat penyumbatan atau penyempitan pembuluh darah koroner, yang disebabkan oleh kerusakan dinding pembuluh darah [1]. Penyakit ini termasuk salah satu penyebab kematian utama di dunia [2]. Gejala yang umum dialami meliputi nyeri dada di sebelah kiri, irama detak jantung dan nadi yang tidak teratur, serta sesak napas [3].

Serangan jantung yang terlambat ditangani dapat menyebabkan komplikasi serius. Oleh karena itu, diperlukan metode untuk mengidentifikasi seseorang yang berisiko terkena penyakit jantung secara lebih dini [4]. Salah satu pendekatan yang dapat digunakan untuk deteksi penyakit ini adalah menggunakan algoritma XGBoost Classifier, yang mampu mengklasifikasi data secara efisien dan akurat [5].

XGBoost merupakan pengembangan dari algoritma gradient boosting yang dirancang untuk meningkatkan kinerja prediksi melalui pengurangan overfitting. Algoritma ini memanfaatkan model pohon regresi yang lebih teratur, sehingga optimal dalam menangani data kategorikal dan kelas yang tidak seimbang [6]. Sehingga di harapkan alogritma ini mampu menghasilkan akurasi yang tinggi [7].

Penerapan algoritma XGBoost dalam deteksi penyakit jantung koroner diharapkan mampu membantu tenaga medis dan pasien untuk mengidentifikasi penyakit ini secara lebih dini, sehingga dapat dilakukan penanganan yang cepat dan tepat.

2. PROFILE DATASET

Dataset yang digunakan merupakan datset yang berisi faktor-faktor umum penyebab penyakit jantung yang menggabungkan 5 dataset terkenal: Cleveland, Hungarian, Switzerland, Long Beach VA, dan Statlog (Heart). Dataset ini memiliki 11 fitur umum terkait penyakit jantung dan 1190 baris data. Dataset ini diunggah oleh Mexwell di Kaggle https://www.kaggle.com/datasets/mexwell/heart-disease-dataset.

Berikut ini merupakan deskripsi dari masing-masing fitur yang terdapat pada dataset:

1. Age : Usia.
2. Sex : Jenis kelamin (1 = male, 0 = female).
3. Chest Pain Type : Jenis nyeri dada yang dialami individu:
    - 1: Angina tipikal (typical angina)
    - 2: Angina atipikal (atypical angina)
    - 3: Nyeri dada non-anginal (non-anginal pain)
    - 4: Asimptomatik (asymptomatic)
4. Resting Blood Pressure: Tekanan darah sistolik saat istirahat (dalam mm Hg)
5. Serum Cholesterol: Kadar kolesterol dalam darah (mg/dl)
6. Fasting Blood Sugar: Level gula darah puasa lebih dari 120 mg/dl, (1 = true; 0 = false).
7. Resting Electrocardiogram Results: Hasil elektrokardiogram saat istirahat:
    - 0: Normal
    - 1: Menunjukkan kelainan gelombang ST-T (termasuk inversi gelombang T dan/atau elevasi atau depresi ST lebih dari 0.05 mV)
    - 2: Menunjukkan hipertrofi ventrikel kiri yang pasti atau mungkin berdasarkan kriteria Estes
8. Maximum Heart Rate Achieved: Detak jantung maksimum
9. Exercise Induced Angina: Apakah individu mengalami angina,(1 = Yes, 0= No).
10. Oldpeak (ST Depression): Depresi.

11. The Slope of the Peak Exercise ST Segment: Kemiringan segmen ST pada puncak latihan:
    - 1: Meningkat (upsloping)
    - 2: Datar (flat)
    - 3: Menurun (downsloping)
12. Target (class): Kelas diagnosis penyakit jantung:
    - 1: Heart Disease (terkena penyakt jantung)
    - 0: Normal (tidak ada penyakit jantung)

3. DATA PREPROCESSING

Preprocessing yang dialkukan adalah fitur selection dan normalisasi data. Alasan dilakukannya tahapan tersebut adalah:

1. Fitur selection : Tahapan ini digunakan untk memilih fitur yang dianggap relevan dengan target.
2. Normalisasi: Tahapan ini dilakukan karena terdapat rentang nilai antar fitur yang sangat bervariasi. Normalisasi diperlukan agar setiap fitur memiliki skala yang seragam, sehingga model dapat berfungsi lebih baik,

4. EXPLORATORY DATA ANALYSIS

Pada analisis EDA dilakukan untuk memeriksa beberapa hal penting, seperti mengecek adanya missing value pada setiap fitur, mengecek tipe data setiap fitur, melihat tampilan data, bentuk data (jumlah baris dan kolom), dan korelasi antar fitur.

Pada proses EDA yang dihasilkan adalah:

1 Data terdiri dari 1190 baris dan 12 fitur.

```
[162]  data.shape

(1190, 12)
```

2 Tipe data terdiri dari numerik (integer dan float)

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  1190 non-null   int64
 1   sex                  1190 non-null   int64
 2   chest pain type      1190 non-null   int64
 3   resting bp s         1190 non-null   int64
 4   cholesterol          1190 non-null   int64
 5   fasting blood sugar  1190 non-null   int64
 6   resting ecg          1190 non-null   int64
 7   max heart rate       1190 non-null   int64
 8   exercise angina      1190 non-null   int64
 9   oldpeak              1190 non-null   float64
 10  ST slope             1190 non-null   int64
 11  target               1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
```

3 Tidak ditemukan missing value

```
data.isnull().sum()
```

	0
age	0
sex	0
chest pain type	0
resting bp s	0
cholesterol	0
fasting blood sugar	0
resting ecg	0
max heart rate	0
exercise angina	0
oldpeak	0
ST slope	0
target	0

dtype: int64

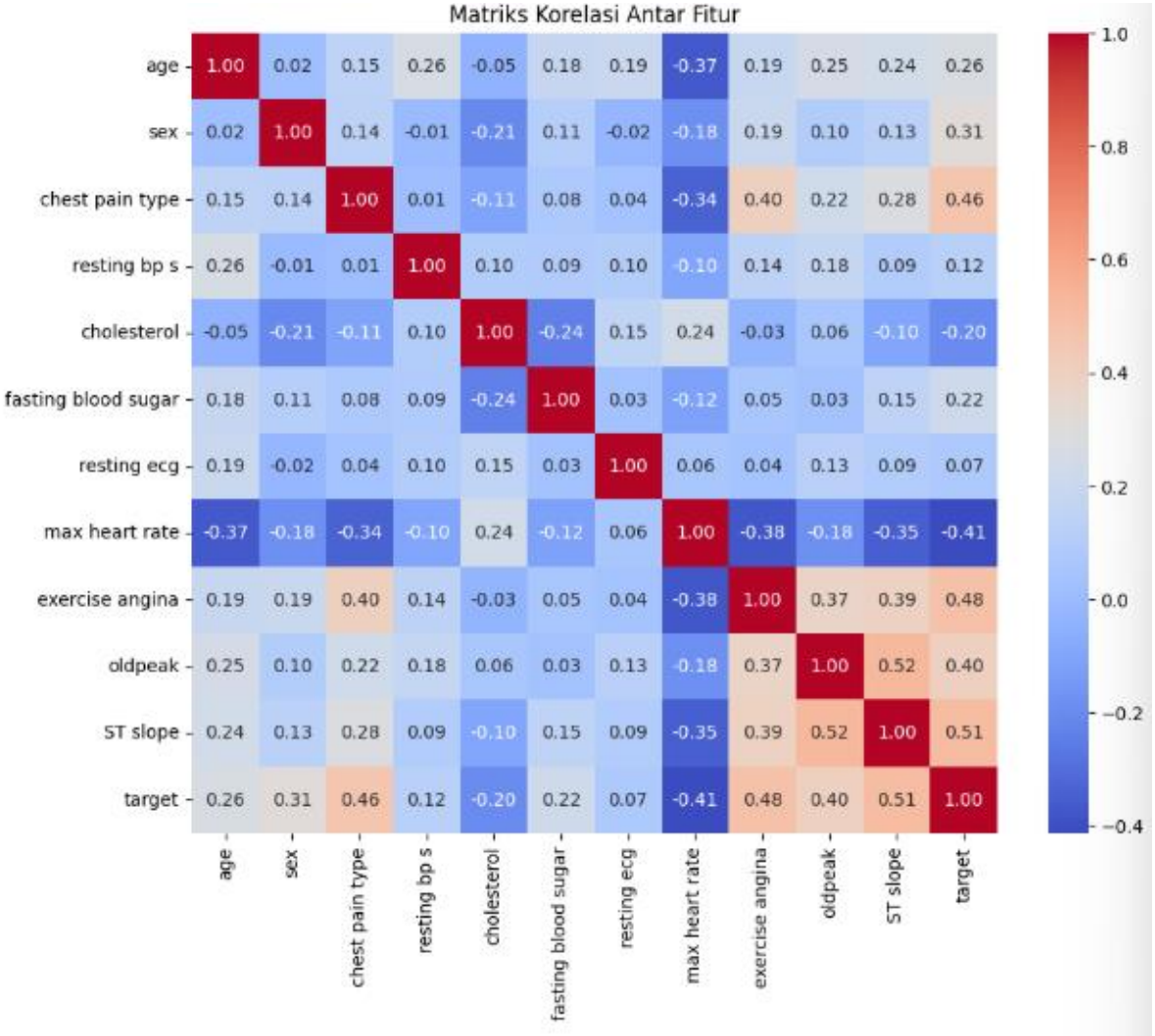
4 Data target (class) yang tidak terlalu jomplang, sehingga balancing data tidak terlalu diperlukan.

```
] data['target'].value_counts()
```

target	count
1	629
0	561

dtype: int64

5 Korelasi antar fitur yang akan digunakan untuk proses fitur selection.



5. SELEKSI FITUR

Pada proses seleksi fitur, digunakan matriks korelasi sebagai metode untuk mempertimbangkan fitur yang relevan. Fitur yang dipertimbangkan memiliki nilai korelasi di luar rentang < -0.1 dan > 0.1. Fitur dengan nilai korelasi antara -0.1 dan 0.1 dianggap memiliki hubungan yang lemah atau tidak signifikan terhadap target variabel. Oleh karena itu, fitur-fitur tersebut akan dihapus, karena dianggap kurang memiliki hubungan yang kuat dengan target. Seleksi fitur ini bertujuan untuk meningkatkan kinerja model dengan mempertahankan fitur-fitur yang memiliki korelasi yang lebih kuat dengan target variabel.

```
age                   0.262029
sex                   0.311267
chest pain type       0.460127
resting bp s          0.121415
cholesterol          -0.198366
fasting blood sugar   0.216695
resting ecg           0.073059
max heart rate       -0.413278
exercise angina       0.481467
oldpeak               0.398385
ST slope              0.505608
dtype: float64
```

Pada hasil di atas, fitur "resting ecg" dianggap memiliki korelasi yang lemah dengan fitur target, sehingga fitur tersebut akan dihapus. Sehingga menyisakan fitur dibawaah ini:

```
age                   0.262029
sex                   0.311267
chest pain type       0.460127
resting bp s          0.121415
cholesterol          -0.198366
fasting blood sugar   0.216695
max heart rate       -0.413278
exercise angina       0.481467
oldpeak               0.398385
ST slope              0.505608
dtype: float64
```

6. MODELING

   Pada analisis ini digunakan model Xtreme Gradiant Boosting, dengan pembagian data 80:20, menghasilkan skor akurasi 93.

   Link github model: https://github.com/LutfianaDeka/UAS_Data-Mining

   Link Launchinpad: https://launchinpad.com/project/penerapan-algoritma-xgboost-classifier-untuk-klasifikasi-penyakit-jantung-koroner-b0e608e

   Link ipnyb: https://colab.research.google.com/drive/1BTu4e1XO40-NW47JT-CE9xm-a2fAfbMQ?usp=sharing

7. EVALUASI MODEL

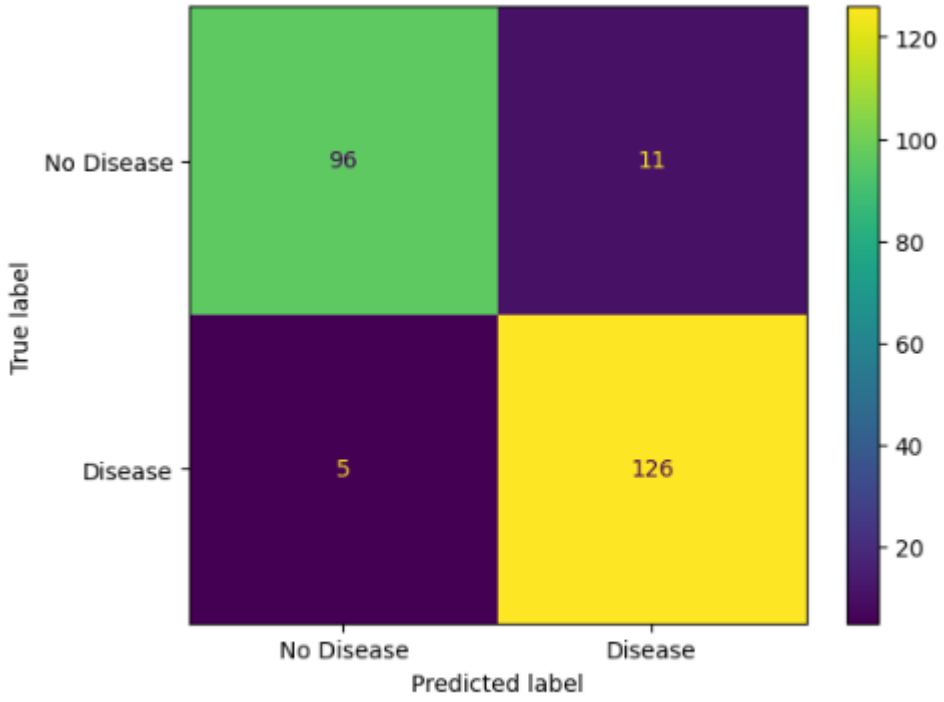
```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.90      0.92       107
           1       0.92      0.96      0.94       131

    accuracy                           0.93       238
   macro avg       0.94      0.93      0.93       238
weighted avg       0.93      0.93      0.93       238

Akurasi Model: 0.93
Precision: 0.93
Recall: 0.93
F1 Score: 0.93
```

Berdasarkan evaluasi menggunakan confusion matrix, model menunjukkan hasil yang sangat baik dengan akurasi sebesar 93%, precision 93%, recall 93%, dan F1 score 93%.



Model berhasil mengkasfifikasi dengan benar Disease sebanyak 126 data, No Disease dengan benar sebanyak 96 data, sedangkan kesalahan False Positif sebanyak 11 data dan False Negatif sebanyak 5 data.

8. ANALISA DAN PEMBAHASAN

Model yang menggunakan XGBoost (Extreme Gradient Boosting) dengan akurasi 93% menunjukkan hasil yang cukup baik. Pada model ini, pemilihan fitur dan juga proses pre-processing cukup berpengaruh pada kinerja model.

9. KESIMPULAN

Berdasarkan hasil percobaan menggunakan algoritma XGBoost pada klasifikasi penyakit jantung, berhasil memperoleh akurasi yang baik, yakni 93%. Selain itu, model ini juga menunjukkan performa yang konsisten dengan precision, recall, dan F1 score masing-masing sebesar 93%. Sehingga dapat disimpulkan bahwa model bekerja dengan baik dan memiliki kesalahan yang rendah.

10. Referensi

[1] W. Margareth, L. P. Tondang, and H. Yemima, “Warning : Your Fat Gonna Kills You,” *Pros. Semin. Nas. Penelit. dan Pengabdi. Kpd. Masy.*, vol. 1, no. 1, pp. 274–276, 2023, [Online]. Available: https://www.who.int/news-

[2] M. C. Untoro, L. Rizta, A. Perdana, N. A. Wijaya, and N. Ferdiyanto, “Penerapan K-Means Clustering pada Imbalance Dataset Gejala Penyakit Jantung,” *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 5, no. 1, pp. 1–7, 2023, doi: 10.28926/ilkomnika.v5i1.455.

[3] K. Kevin, “Diagnosa Penyakit Jantung Menggunakan Metode Certainty Factor,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 1, pp. 93–106, 2022, doi: 10.33365/jatika.v3i1.1866.

[4] H. Hidayat, A. Sunyoto, and H. Al Fatta, “Klasifikasi Penyakit Jantung

Menggunakan Random Forest Clasifier," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 7, no. 1, pp. 31–40, 2023, doi: 10.47970/siskom-kb.v7i1.464.

[5] D. M. Pratiwi and L. Mufidah, "Perbandingan Metode Decision Tree Classifier dan XGBoost Classifier Dalam Memprediksi Penyakit Jantung," pp. 991–1000, 2024.

[6] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.

[7] Y. Amelia, "Perbandingan Metode Machine Learning Untuk Mendeteksi Penyakit Jantung," *IDEALIS Indones. J. Inf. Syst.*, vol. 6, no. 2, pp. 220–225, 2023, doi: 10.36080/idealis.v6i2.3043.