PROJEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

PERBANDINGAN METODE LDA, SVM, DAN *GRADIENT BOOSTING* DALAM KLASIFIKASI PENYAKIT JANTUNG

Dosen Pengampu

Kusnawi, S.Kom, M.Eng

Disusun oleh

21.11.4185

Luthfia Ridho Damayanti

BDDM-IF03

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

BAB I PENDAHULUAN

I. Latar Belakang

Jantung adalah organ vital yang berfungsi sebagai pemompa darah untuk memenuhi kebutuhan oksigen dan nutrisi ke seluruh tubuh[1]. Jika pembuluh darah mengalami penyempitan, maka fungsi jantung akan mengalami gangguan sehingga menyebabkan penyakit jantung[7]. Jantung sebagai salah organ terpenting dalam tubuh memiliki resiko kematian jika ada kelainan yang terjadi pada jantung[2]. Penyakit jantung terjadi karena penyumbatan sebagaian atau total dari suatu lebih pembuluh darah, akibat dari adanya penyumbatan maka dengan sendirinya suplai dan kebutuhan darah, penyakit jantung merupakan penyebab kematian tertinggi kedua setelah stroke[3]. Oleh karena itu sangat penting untuk mendeteksi penyakit jantung sejak dini dan mengelolanya dengan baik. Upaya pencegahan, diagnosis dini, dan perawatan yang tepat waktu dapat mengurangi angka kematian akibat penyakit jantung dengan meningkatkan kualitas hidup pasien.

Dalam upaya mendeteksi penyakit jantung secara efektif, metode klasifikasi memiliki peran yang signifikan. Linear Discriminant Analysis (LDA) adalah salah satu teknik yang banyak digunakan dalam pengenalan pola statistik, dengan memanfaatkan proyeksi linear untuk memaksimalkan pemisahan antar-kelas sekaligus meminimalkan penyimpangan dalam kelompok yang sama[4]. Di sisi lain, Support Vector Machines (SVM) adalah algoritma machine learning yang dirancang untuk memisahkan dua kelas dengan menentukan hyperplane terbaik yang memaksimalkan margin antara kelas[6]. Selain itu, Gradient Boosting, seperti XGBoost, dikenal sebagai salah satu algoritma ensemble learning yang sangat andal dalam menyelesaikan masalah klasifikasi maupun regresi[5]. Kombinasi keunggulan ketiga metode ini diharapkan dapat memberikan wawasan lebih dalam tentang cara mendeteksi penyakit jantung secara akurat dan efisien.

Proyek ini tidak hanya membangun model klasifikasi untuk deteksi penyakit jantung, tetapi juga membandingkan kinerja Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), dan Gradient Boosting untuk menentukan algoritma yang memiliki akurasi terbaik. Pendekatan ini meliputi langkah-langkah seperti pengambilan data, preprocessing data, eda, seleksi fitur, implementasi algoritma, evaluasi model berdasarkan metrik tertentu, dan penyusunan kesimpulan. Hasil proyek diharapkan dapat memberikan rekomendasi algoritma yang paling unggul untuk mendukung deteksi dini penyakit jantung secara lebih efektif.

I. Tujuan

1) Mengembangkan model klasifikasi berbasis machine learning untuk memprediksi penyakit jantung berdasarkan fitur-fitur relevan yang telah dipilih.
2) Menganalisis dan membandingkan performa tiga metode klasifikasi, yaitu Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), dan Gradient Boosting (GB), dalam mendeteksi penyakit jantung pada dataset medis.
3) Menilai akurasi, presisi, recall, dan F1-score dari masing-masing model untuk menentukan metode yang paling efektif dalam klasifikasi penyakit jantung.
4) Mengeksplorasi kelebihan dan kekurangan dari setiap metode klasifikasi, baik dari segi kompleksitas model, interpretabilitas, maupun kemampuan prediksi.

II. Metode

Proyek ini menggunakan pendekatan kuantitatif dengan eksperimen untuk membandingkan tiga algoritma klasifikasi, yaitu Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), dan Gradient Boosting (GB), dalam klasifikasi penyakit jantung. Langkah-langkah metode yang dilakukan meliputi:

1) Pengambilan Data
   Dataset yang digunakan dalam proyek ini diambil pada platfrom public yaitu Kaggle.
2) Preprocessing
   Preprocessing ini dilakukan untuk memeriksa data pada dataset yang dipilih dan mengidentifikasi atau memperbaiki kesalahan yang ditemukan pada dataset tersebut, sehingga dapat dilanjutkan pada langkah berikutnya.
3) Exploratory Data Analysis (EDA)
   Analisis data eksplorasi dilakukan untuk memahami distribusi data, hubungan antar fitur, dan pola penting yang ada dalam dataset. Visualisasi dalam proyek ini adalah scatter plot untuk memberikan wawasan lebih mendalam tentang dataset.
4) Seleksi Fitur
   Seleksi fitur dilakukan untuk mengidentifikasi atribut yang paling berpengaruh dalam deteksi penyakit jantung.
5) Modelling
   Tiga algoritma klasifikasi, yaitu LDA, SVM, dan Gradient Boosting, diterapkan menggunakan library scikit-learn. Parameter utama pada setiap model akan dioptimalkan untuk mencapai performa terbaik.
   1. LDA: Memproyeksikan data ke ruang linear untuk memaksimalkan pemisahan antar-kelas.

2. SVM: Menentukan hyperplane terbaik untuk memisahkan data menjadi dua kelas.
3. Gradient Boosting: Membuat model ensemble berbasis pohon keputusan untuk meningkatkan akurasi prediksi.

6) Evaluasi Model

Kinerja model dievaluasi menggunakan metrik berikut:

1. Akurasi: Mengukur seberapa tepat prediksi model.
2. Precision, Recall, dan F1-Score: Memberikan analisis lebih mendalam tentang kinerja model terhadap data minoritas.
3. Confusion Matrix: Menampilkan distribusi prediksi yang benar dan salah.

BAB II PROFILE DATASET

I. Sumber Dataset

Dataset yang digunakan dalam proyek ini merupakan kumpulan data yang tersedia secara public milik desalegngeb yang dapat diakses melalui platform Kaggle berikut https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/notebook dataset ini yang nantinya akan digunakan untuk proses pengolahan klasifikasi.

II. Informasi Dataset

Dataset ini berisi 303 baris dan 14 kolom, dengan semua kolom memiliki jumlah nilai non-null yang sama (303 non-null).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Berikut adalah rincian dari setiap kolom/fitur:

No	Fitur	Keterangan	Tipe Data
1	age	Menyimpan usia	Integer (int64)
2	sex	(1 = laki-laki; 0 = Perempuan)	Integer (int64)
3	cp	Jenis nyeri dada (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)	Integer (int64)
4	trestbps	Tekanan darah	Integer (int64)
5	chol	Kolesterol	Integer (int64)
6	fbs	Gula darah > 120 mg/dl (1 = true; 0 = false)	Integer (int64)
7	restecg	Hasil elektrokardiografi (0 = normal; 1 = abnormality ; 2 = hypertrophy)	Integer (int64)
8	thalach	Detak jantung maksimum	Integer (int64)

9	exang	Latihan diinduksi angina (1 = yes; 0 = no)	Integer (int64)
10	oldpeak	Depresi ST	Float (float64)
11	slope	Segmen ST latihan puncak (0 = upsloping; 1 = flat; 2 = downsloping)	Integer (int64)
12	ca	Jumlah kapal utama oleh flourosopy	Integer (int64)
13	thal	Kelainan darah (0 = error; 1 = fixed defect, 2 = normal, 3 = reversable defect)	Integer (int64)
14	target	(0 = tidak ada penyakit/sehat; 1 = penyakit)	Integer (int64)

III. Karakteristik Data

1) Jumlah Data
   Dataset terdiri dari 303 sampel (baris data) dengan 13 fitur (kolom) dan 1 target.
2) Tipe Data
   a) Integer (int64): Sebagian besar kolom, seperti age, sex, dan lainnya, menggunakan tipe data integer.
   b) Float (float64): Kolom oldpeak memiliki nilai desimal, karena merepresentasikan perubahan segmen ST yang bersifat kontinu.
3) Kualitas Data
   Tidak ada missing values, sehingga tidak diperlukan imputasi atau penanganan nilai yang hilang.
4) Distribusi Target
   Kolom target adalah variabel dependen yang menunjukkan apakah pasien memiliki penyakit jantung (1) atau tidak (0).

BAB III DATA PREPROCESSING

Tahap preprocessing data dilakukan untuk mempersiapkan dataset agar dapat digunakan dalam proses pemodelan. Teknik-teknik berikut diterapkan:

I. Pemeriksaan Missing Values

Dilakukan identifikasi nilai yang hilang (missing values) dalam dataset untuk memastikan kualitas data. Kolom yang memiliki missing values akan dipertimbangkan untuk penanganan lebih lanjut, seperti imputasi atau penghapusan.

```
Jumlah 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
```

Pemeriksaan nilai yang hilang dilakukan untuk memastikan kualitas data yang baik sebelum digunakan dalam proses analisis dan pemodelan. Dataset dengan missing values dapat menyebabkan bias dalam model atau mengurangi akurasi prediksi. Dalam penelitian ini, tidak ditemukan missing values, sehingga data dapat langsung digunakan tanpa perlu penanganan tambahan.

II. Pemisahan Fitur dan Target

Dataset dipisahkan menjadi:

a) Fitur (X): Variabel independen yang digunakan untuk memprediksi.
b) Target (y): Variabel dependen yang menunjukkan label atau output yang akan diprediksi.

```
Fitur (X):
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope
0   63    1   3       145   233    1        0      150      0      2.3      0
1   37    1   2       130   250    0        1      187      0      3.5      0
2   41    0   1       130   204    0        0      172      0      1.4      2
3   56    1   1       120   236    0        1      178      0      0.8      2
4   57    0   0       120   354    0        1      163      1      0.6      2

   ca  thal
0   0     1
1   0     2
2   0     2
3   0     2
4   0     2

Target (y):
target
1    165
0    138
Name: count, dtype: int64
```

Pemisahan fitur dan target untuk membedakan variabel yang digunakan sebagai input (predictor) dengan output yang ingin diprediksi.

III. Penanganan Kolom Kategorikal

```
Tipe Data Setelah Konversi Kategorikal:
sex         category
cp          category
restecg     category
slope       category
ca          category
thal        category
dtype: object
```

Kolom yang bersifat kategorikal seperti sex, cp, restecg, slope, ca, dan thal dikonversi menjadi tipe data category. Langkah ini bertujuan untuk:

a) Mengurangi penggunaan memori.
b) Mempermudah proses encoding atau transformasi selanjutnya.

IV. Pembagian Dataset

Dataset dibagi menjadi data latih (training set) dan data uji (testing set) untuk mengevaluasi performa model secara objektif.

a) Training Set: Digunakan untuk melatih model, sehingga dapat mempelajari pola dari data.
b) Testing Set: Digunakan untuk menguji model pada data yang belum pernah dilihat sebelumnya, sehingga mengevaluasi kemampuan generalisasi model.

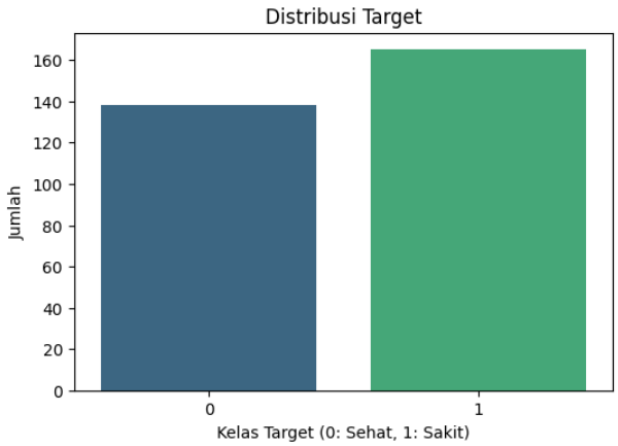
```
Ukuran Data Latih dan Uji:
Data latih: 242 baris, 13 kolom
Data uji: 61 baris, 13 kolom
```

Metode ini dilakukan untuk mengurangi risiko overfitting karena model diuji pada data yang tidak terlihat selama pelatihan.

BAB IV EXPLORATORY DATA ANALYSIS

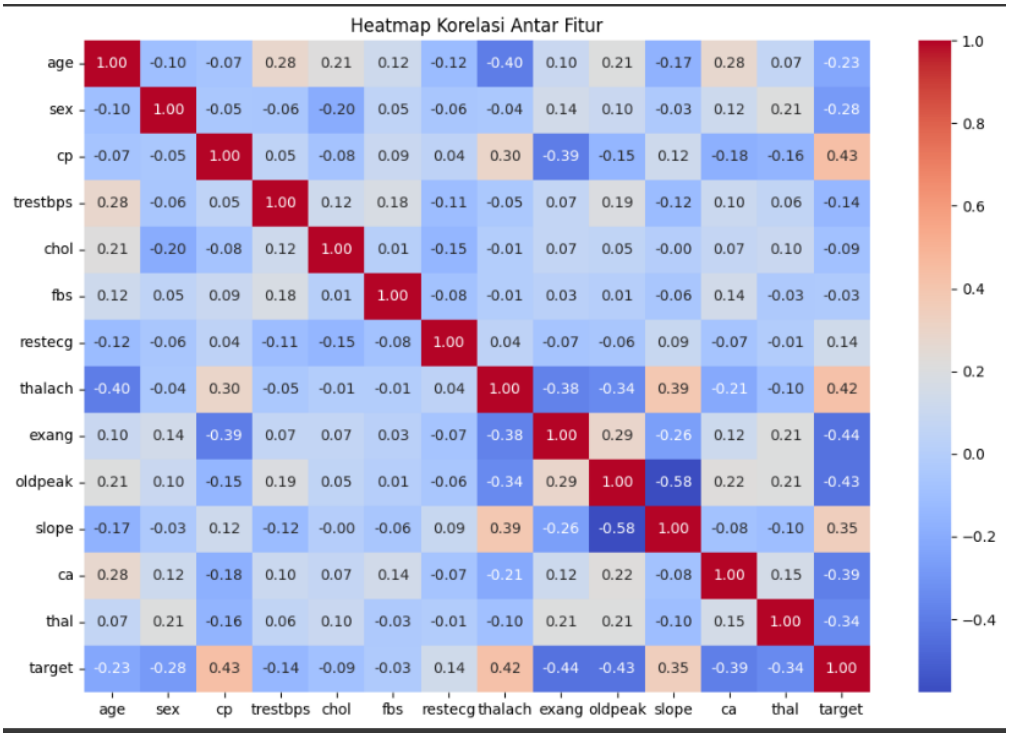
Analisis data eksplorasi dilakukan untuk memahami distribusi data, hubungan antar fitur, dan pola penting yang ada dalam dataset. Visualisasi dalam proyek ini adalah scatter plot, heatmap dan lainnya untuk memberikan wawasan lebih mendalam tentang dataset. Berikut hasil yang didapatkan:

I. Visualisasi Distribusi Target



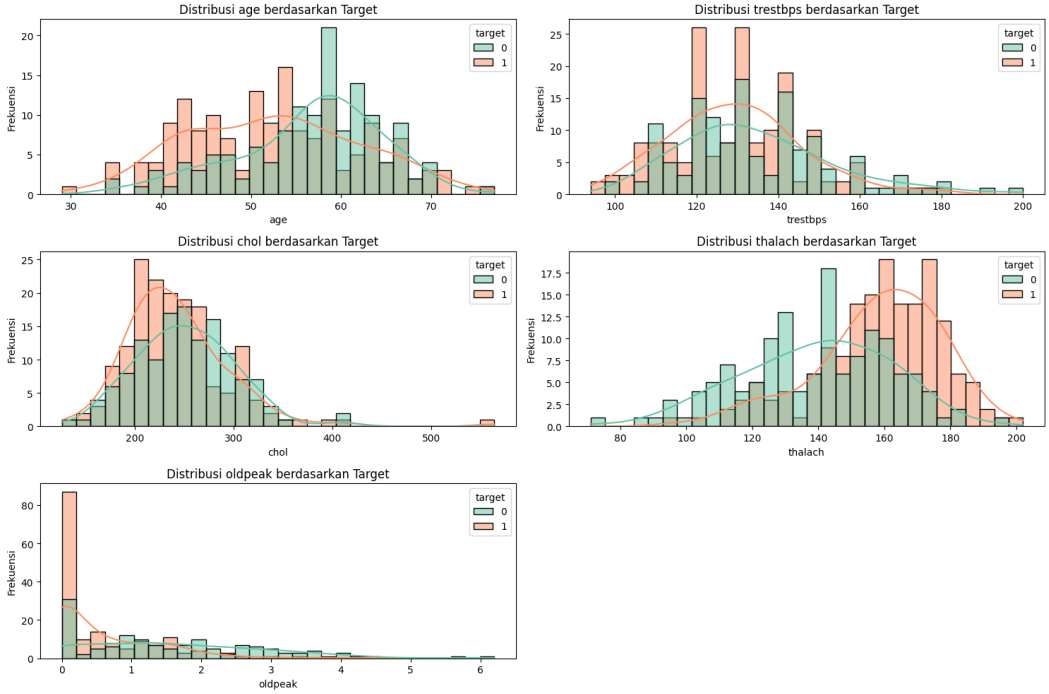
Kelas 1 (Sakit) memiliki jumlah yang sedikit lebih besar dibandingkan dengan kelas 0 (Sehat). Perbedaan jumlah tidak terlalu signifikan, sehingga dataset ini cukup seimbang, dan tidak perlu melakukan penanganan khusus seperti oversampling atau undersampling untuk mengatasi ketidakseimbangan kelas.

II. Visualisasi Korelasi Antar Fitur



Heatmap korelasi menunjukkan hubungan antar fitur dan target dalam dataset. Fitur yang memiliki korelasi positif signifikan dengan target adalah cp (chest pain) (+0.43) dan thalach (maximum heart rate achieved) (+0.42), sementara fitur dengan korelasi negatif signifikan adalah exang (exercise induced angina) (-0.44), oldpeak (ST depression) (-0.43), dan ca (number of major vessels) (-0.39). Sebagian besar fitur lainnya memiliki korelasi rendah terhadap target, yang menunjukkan bahwa model mungkin memerlukan feature engineering atau pendekatan non-linear untuk meningkatkan performa. Selain itu, beberapa fitur seperti ca-thal (+0.55) dan oldpeak-slope (-0.58) menunjukkan korelasi antar fitur yang cukup tinggi, yang dapat memicu masalah multikolinearitas jika tidak ditangani dengan baik. Oleh karena itu, disarankan untuk fokus pada fitur-fitur dengan korelasi signifikan terhadap target, serta mempertimbangkan regularisasi atau teknik pengurangan dimensi untuk mengatasi multikolinearitas sebelum proses modeling.

III. Distribusi Fitur Numerik



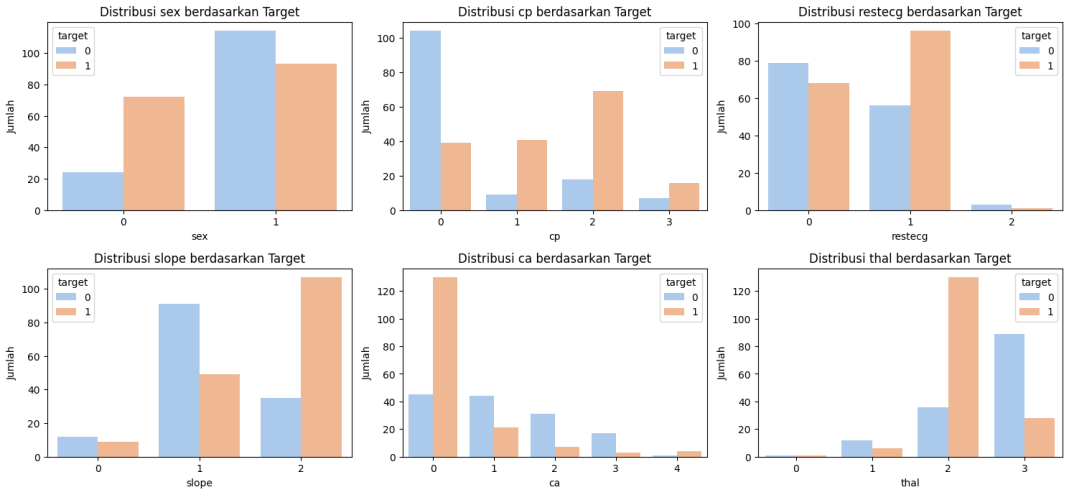
Visualisasi ini tersebut distribusi beberapa fitur numerik (age, trestbps, chol, thalach, oldpeak) berdasarkan target klasifikasi (0: Sehat, 1: Sakit). Berikut penjelasannya disetiap fitur:

1) Age (Usia):
   a) Distribusi usia pasien dengan target 1 (Sakit) cenderung lebih tinggi pada rentang 50–60 tahun.
   b) Pasien dengan target 0 (Sehat) lebih tersebar pada rentang usia 30–70 tahun, menunjukkan distribusi yang lebih merata.
2) Trestbps (Tekanan Darah Saat Istirahat):
   a) Distribusi pada kedua target terlihat cukup mirip, dengan mayoritas pasien memiliki tekanan darah pada rentang 110–140 mmHg.

b) Tidak ada perbedaan pola yang signifikan antara target 0 dan 1.

3) Chol (Kolesterol Total):
   Sebagian besar pasien, baik dengan target 0 maupun 1, memiliki kadar kolesterol pada rentang 200–300 mg/dL.
4) Thalach (Denyut Jantung Maksimum):
   Pasien dengan target 1 (Sakit) cenderung memiliki nilai thalach yang lebih tinggi (140–170).
5) Oldpeak (Depresi ST):
   a) Pasien dengan target 1 (Sakit) cenderung memiliki nilai oldpeak yang lebih rendah (sekitar 0–1).
   b) Sebaliknya, pasien dengan target 0 (Sehat) menunjukkan distribusi yang lebih lebar dengan nilai oldpeak lebih tinggi (>2).

IV. Distribusi Fitur Kategorikal

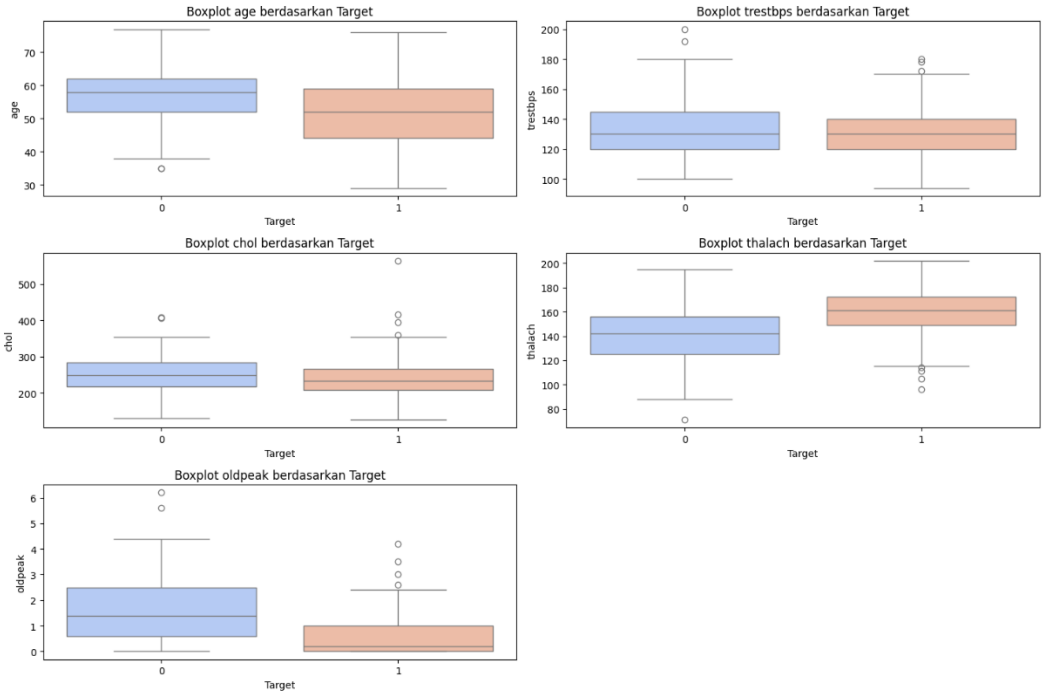


Berdasarkan grafik hasil distribusi, berikut adalah ulasan hasil exploratory data analysis (EDA) dari distribusi variabel-variabel terhadap target:

1) Distribusi Sex Berdasarkan Target:
   a) Kode 1 pada variabel sex (kemungkinan pria) memiliki jumlah yang lebih besar dibandingkan kode 0 (kemungkinan wanita).
   b) Pria (kode 1) yang memiliki target 1 (positif penyakit) lebih banyak dibandingkan wanita.
   c) Namun, pada target 0 (negatif penyakit), pria juga mendominasi.
2) Distribusi CP (Chest Pain Type) Berdasarkan Target:
   a) Tipe nyeri dada (kode 2) memiliki korelasi yang lebih tinggi terhadap target 1 (positif penyakit) dibandingkan tipe lainnya.
   b) Tipe nyeri dada 0 (tidak ada gejala) lebih sering terjadi pada pasien dengan target 0 (negatif penyakit).
3) Distribusi RestECG (Electrocardiographic Results) Berdasarkan Target:
   a) RestECG dengan nilai 1 (hasil tidak normal) lebih banyak dikaitkan dengan target 1 (positif penyakit).
   b) Sementara itu, hasil RestECG normal (nilai 0) cukup berimbang, meskipun lebih condong pada target 0 (negatif penyakit).

4) Distribusi Slope Berdasarkan Target:
   a) Slope dengan nilai 2 (upsloping) paling sering muncul pada pasien dengan target 1 (positif penyakit).
   b) Slope dengan nilai 1 (flat) lebih sering terkait dengan target 0 (negatif penyakit).
5) Distribusi CA (Number of Major Vessels) Berdasarkan Target:
   a) Jumlah pembuluh darah besar bernilai 0 mendominasi pasien dengan target 1 (positif penyakit).
   b) Jumlah pembuluh darah yang lebih banyak (nilai 1, 2, 3, atau 4) cenderung lebih banyak ditemukan pada target 0 (negatif penyakit).
6) Distribusi Thal Berdasarkan Target:
   a) Thal dengan nilai 2 (fixed defect) sangat sering dikaitkan dengan target 1 (positif penyakit).
   b) Thal dengan nilai 3 (reversible defect) lebih sering dikaitkan dengan target 0 (negatif penyakit).

V. Outlier Detection



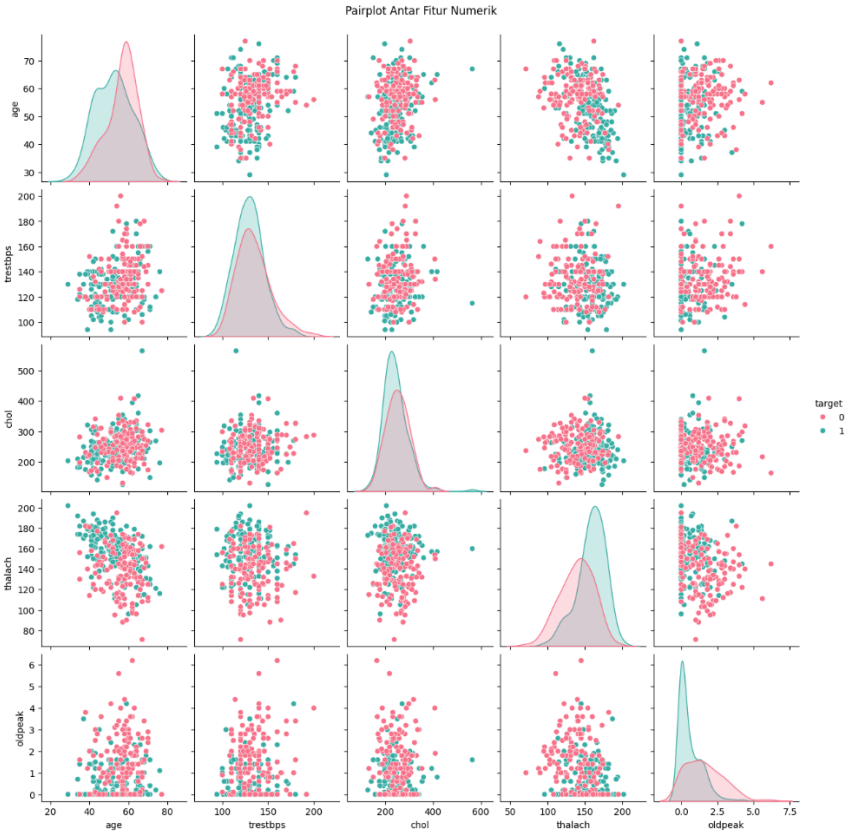
Berdasarkan boxplot hasil deteksi, berikut adalah ulasan singkat dan detail terkait analisis eksplorasi (EDA) terhadap variabel-variabel yang dipengaruhi oleh target:

1) Boxplot Age Berdasarkan Target:
   a) Distribusi usia antara target 0 (tidak memiliki penyakit) dan target 1 (memiliki penyakit) cukup mirip, dengan median usia di kisaran 55-60 tahun.
   b) Usia pada target 1 sedikit lebih tersebar ke bawah (lebih muda) dibandingkan target 0.
2) Boxplot Trestbps (Resting Blood Pressure) Berdasarkan Target:

a) Distribusi tekanan darah cenderung serupa antara target 0 dan 1, dengan median sekitar 130-140 mmHg.
b) Tidak ada perbedaan signifikan dalam persebaran nilai tekanan darah antara kedua target.

3) Boxplot Chol (Serum Cholesterol) Berdasarkan Target:
   a) Nilai kolesterol pada target 0 dan 1 memiliki median yang mirip, tetapi target 1 menunjukkan lebih banyak outlier di atas 300.
   b) Persebaran kolesterol cenderung lebih lebar pada target 1.
4) Boxplot Thalach (Maximum Heart Rate Achieved) Berdasarkan Target:
   a) Median thalach untuk target 1 lebih tinggi dibandingkan target 0, menunjukkan bahwa pasien dengan target 1 cenderung memiliki detak jantung maksimum yang lebih tinggi.
   b) Persebaran thalach pada target 0 lebih rendah dibandingkan target 1, dengan beberapa outlier yang rendah.
5) Boxplot Oldpeak Berdasarkan Target:
   a) Median oldpeak (ST depression) pada target 0 lebih tinggi dibandingkan target 1, menunjukkan pasien dengan target 0 lebih cenderung memiliki depresi segmen ST yang lebih besar.
   b) Persebaran oldpeak pada target 0 lebih luas dibandingkan target 1.

VI. Hubungan Antar Fitur



Pairplot tersebut menunjukkan distribusi dan hubungan antar fitur numerik (age, trestbps, chol, thalach, oldpeak) berdasarkan target klasifikasi (0: Sehat, 1: Sakit). Berikut analisisnya:

1) Distribusi Univariate:
   a) Age: Pasien dengan target 1 (Sakit) cenderung terkonsentrasi pada usia 50–60 tahun, sedangkan distribusi pasien 0 (Sehat) lebih merata.
   b) Thalach: Distribusi pasien 1 (Sakit) menunjukkan nilai denyut jantung maksimum yang lebih tinggi dibanding pasien 0 (Sehat).
   c) Oldpeak: Pasien 0 (Sehat) lebih tersebar pada nilai oldpeak yang lebih tinggi, sementara pasien 1 (Sakit) cenderung terfokus pada nilai rendah (sekitar 0–1).
2) Hubungan Antar Fitur:
   a) Tidak terdapat pola hubungan linier yang sangat kuat antara sebagian besar fitur, seperti age vs trestbps atau chol vs thalach, menunjukkan fitur-fitur ini bersifat independen secara relatif.
   b) Kombinasi thalach vs oldpeak memperlihatkan pemisahan yang lebih jelas antara target 0 dan 1, yang menunjukkan potensi kontribusi signifikan dari kedua fitur ini dalam membedakan target.
3) Pemisahan Berdasarkan Target:

   Pasien dengan target 1 (Sakit) terlihat lebih terdistribusi pada area tertentu dalam plot antar fitur, seperti pada kombinasi thalach vs age dan oldpeak vs chol, yang menandakan bahwa fitur ini dapat memberikan informasi yang bermanfaat untuk klasifikasi.

BAB V SELEKSI FITUR

Pada proyek ini dilakukan beberapa metode seleksi fitur untuk mengidentifikasi fitur terbaik yang memiliki hubungan signifikan dengan target variabel. Metode yang digunakan adalah sebagai berikut;

I. Metode Chi-square

```
Fitur yang dipilih (Chi-square):
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'thalach', 'exang', 'oldpeak',
       'slope', 'ca'],
      dtype='object')
```

Metode Chi-square mengevaluasi ketergantungan antara fitur dan target berdasarkan distribusi chi-square. Metode ini lebih cocok untuk fitur kategorikal, hasilnya mungkin kurang optimal untuk fitur numerik, seperti trestbps dan chol, yang memiliki nilai kontinu.

II. Metode Mutual Information

```
Fitur yang dipilih (Mutual Information):
Index(['cp', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope',
       'ca', 'thal'],
      dtype='object')
```

Mutual Information mengukur ketergantungan non-linear antara fitur dan target, sehingga mampu menangkap hubungan kompleks yang mungkin terlewat oleh metode linear seperti Chi-square. Hasil seleksi menunjukkan fitur seperti fbs dan thal dipilih, meskipun korelasi langsungnya dengan target tidak signifikan. Hal ini menunjukkan potensi hubungan non-linear.

III. Evaluasi Korelasi terhadap Target

```
Korelasi fitur terhadap target:
target      1.000000
cp          0.433798
thalach     0.421741
slope       0.345877
restecg     0.137230
fbs        -0.028046
chol       -0.085239
trestbps   -0.144931
age        -0.225439
sex        -0.280937
thal       -0.344029
ca         -0.391724
oldpeak    -0.430696
exang      -0.436757
Name: target, dtype: float64
```

Metode ini menunjukkan bahwa fitur seperti cp, thalach, dan slope memiliki korelasi positif yang cukup kuat terhadap target, sedangkan exang memiliki korelasi negatif yang signifikan.

IV. Seleksi berdasarkan Threshold Korelasi

```
Fitur yang dipilih berdasarkan korelasi:
Index(['target', 'cp', 'thalach', 'slope', 'restecg', 'trestbps', 'age', 'sex',
       'thal', 'ca', 'oldpeak', 'exang'],
      dtype='object')
```

Setelah membandingkan ketiga metode seleksi fitur, metode Mutual Information dipilih sebagai pendekatan akhir. Alasan pemilihan ini adalah:

1) Kemampuan untuk menangkap hubungan non-linear, yang sesuai dengan kompleksitas dataset penyakit jantung.
2) Fitur yang dipilih mencakup variabel penting seperti fbs dan thal, yang mungkin memiliki kontribusi signifikan terhadap model.

BAB VI MODELLING

Pada proyek ini menggunakan tiga model alogaritma untuk mengetahui akurasi yang paling baik diantara ketiga model tersebut. Berikut adalah ketiga model yang digunakan;

I. Model LDA (Linear Discriminant Analysis)

LDA adalah metode statistik yang digunakan untuk memisahkan atau mengklasifikasikan objek berdasarkan fitur-fitur input dengan memaksimalkan jarak antar kelas dan meminimalkan variansi dalam kelas.

Kode implementasi:

```
# Model LDA (Linear Discriminant Analysis)
def train_lda(X_train, y_train, X_test, y_test):
    print("LDA Model")
    model = LinearDiscriminantAnalysis()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    return y_test, y_pred, "LDA"
```

II. Model Support Vector Machine (SVM)

SVM adalah algoritma supervised learning yang digunakan untuk klasifikasi dengan mencari hyperplane terbaik yang memisahkan data dari kelas yang berbeda. Pada eksperimen ini, digunakan kernel linear.

Kode implementasi:

```
# Model SVM
def train_svm(X_train, y_train, X_test, y_test):
    print("SVM Model")
    model = SVC(kernel='linear', C=1.0, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    return y_test, y_pred, "SVM"
```

III. Model Gradient Boosting Classifier

Gradient Boosting adalah metode ensemble learning yang menggabungkan prediksi dari beberapa model lemah untuk menghasilkan model yang kuat. Algoritma ini bekerja dengan membangun model secara bertahap untuk meminimalkan error.

Kode implementasi:

```
# Model Gradient Boosting
def train_gradient_boosting(X_train, y_train, X_test, y_test):
    print("Gradient Boosting Model")
    model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    return y_test, y_pred, "Gradient Boosting"
```

Link model: https://github.com/luthfiaard/PERBANDINGAN-METODE-LDA-SVM-DAN-GRADIENT-BOOSTING-DALAM-KLASIFIKASI-PENYAKIT-JANTUNG

Link file ipnyb: https://colab.research.google.com/drive/1klmrqb66bxm39LqJUClWob4wQeV5BS1o?usp=sharing

BAB VII EVALUASI MODEL

I. Evaluasi Model

Evaluasi model pada proyek ini dilakukan berdasarkan beberapa metrik kinerja yaitu:

1) Accuracy: Proporsi prediksi yang benar dari keseluruhan data.
2) Precision: Proporsi prediksi benar dari total prediksi positif.
3) Recall: Proporsi prediksi benar dari total kejadian aktual positif.
4) F1-Score: Harmonik rata-rata antara precision dan recall.
5) Confusion Matrix: Matriks yang menunjukkan distribusi prediksi benar dan salah.

II. Hasil Evaluasi

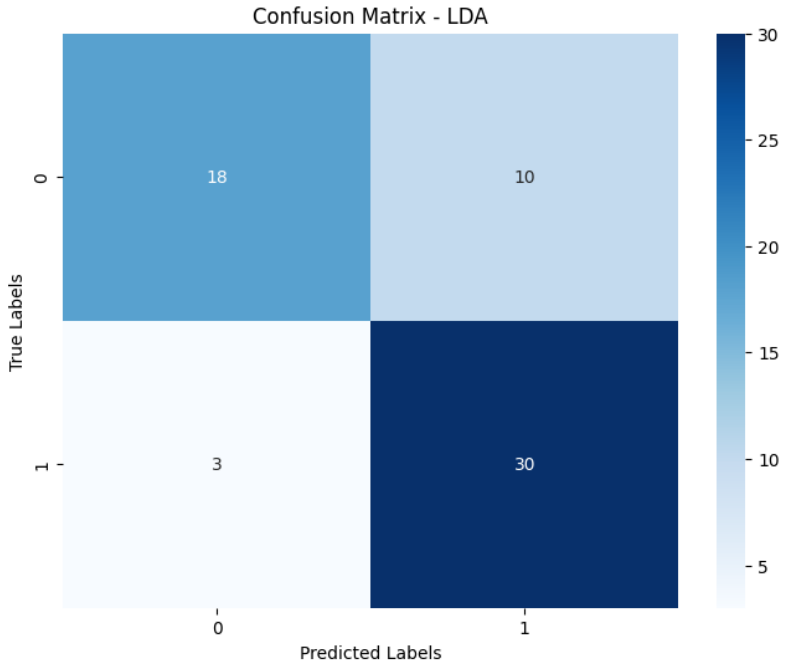
1) Evaluasi Model LDA

   Classification Report:

LDA Model
Classification Report - LDA:

	precision	recall	f1-score	support
0	0.86	0.64	0.73	28
1	0.75	0.91	0.82	33
accuracy			0.79	61
macro avg	0.80	0.78	0.78	61
weighted avg	0.80	0.79	0.78	61

   Confusion Matrix:



2) Evaluasi Model SVM

Classification Report:

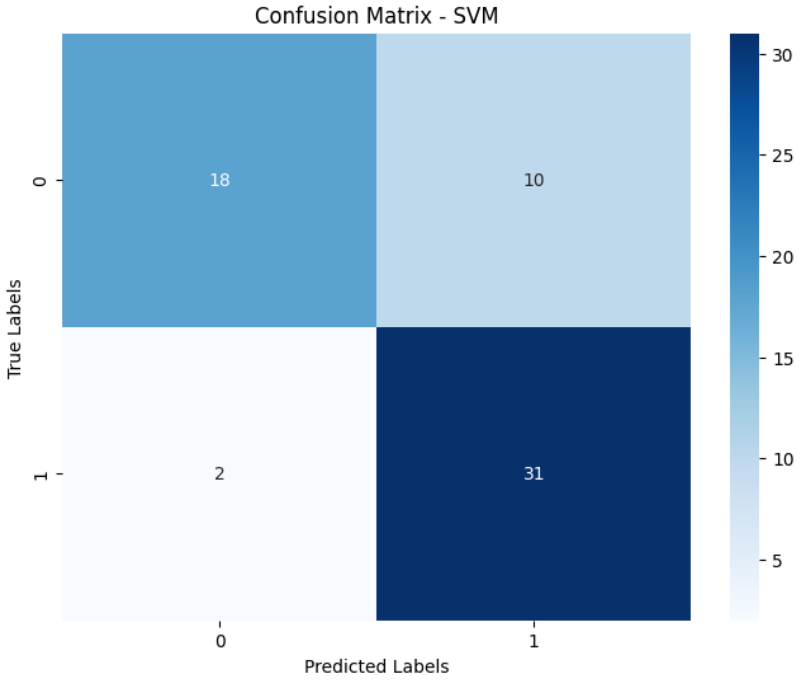
```
SVM Model
Classification Report - SVM:

              precision    recall  f1-score   support

           0       0.90      0.64      0.75        28
           1       0.76      0.94      0.84        33

    accuracy                           0.80        61
   macro avg       0.83      0.79      0.79        61
weighted avg       0.82      0.80      0.80        61
```

Confusion Matrix:



3) Evaluasi Model GB

Classification Report:

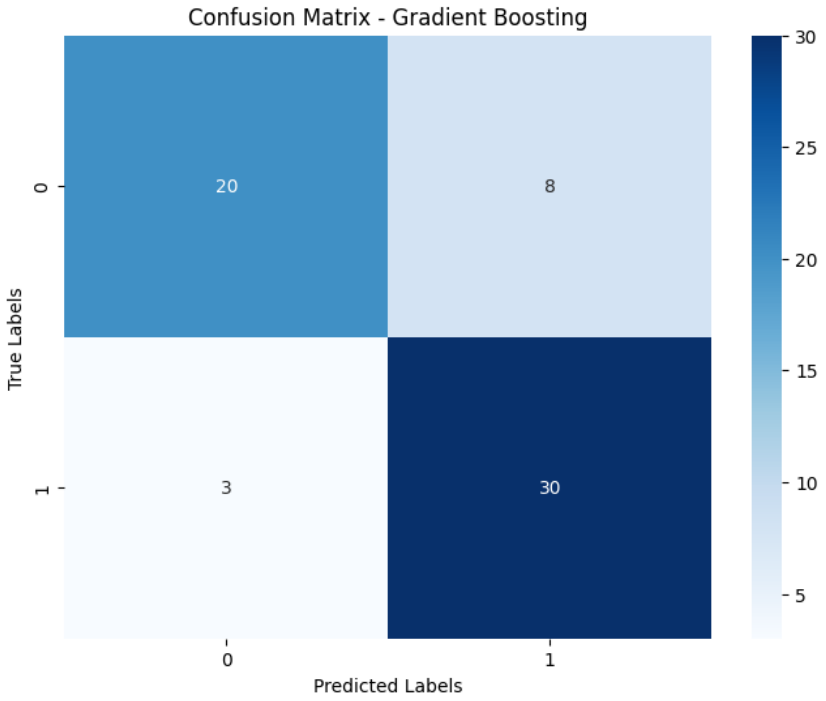
```
Gradient Boosting Model
Classification Report - Gradient Boosting:

              precision    recall  f1-score   support

           0       0.87      0.71      0.78        28
           1       0.79      0.91      0.85        33

    accuracy                           0.82        61
   macro avg       0.83      0.81      0.81        61
weighted avg       0.83      0.82      0.82        61
```

Confusion Matrix:



III. Interpretasi Metrics Evaluasi

1) LDA (Linear Discriminant Analysis):
   a) Accuracy: 0.79 menunjukkan bahwa model LDA mampu mengklasifikasikan 79% sampel dengan benar.
   b) Precision:
      1. Kelas 0(sehat): 0.86 mengindikasikan bahwa dari semua prediksi kelas 0, 86% benar.
      2. Kelas 1(sakit): 0.75 berarti bahwa dari semua prediksi kelas 1, 75% benar.
   c) Recall:
      1. Kelas 0(sehat): 0.64 menunjukkan bahwa model hanya mampu mendeteksi 64% dari total kelas 0.
      2. Kelas 1(sakit): 0.91 berarti model mampu mendeteksi 91% dari total kelas 1.
   d) F1-Score:
      1. Kelas 0(sehat): 0.73 menunjukkan keseimbangan antara presisi dan recall untuk kelas 0.
      2. Kelas 1(sakit): 0.82 menunjukkan keseimbangan antara presisi dan recall untuk kelas 1.

2) SVM (Support Vector Machine):
   a) Accuracy: 0.80 menunjukkan bahwa model SVM mampu mengklasifikasikan 80% sampel dengan benar.
   b) Precision:
      1. Kelas 0: 0.90 berarti 90% prediksi kelas 0 benar.
      2. Kelas 1(sakit): 0.76 berarti 76% prediksi kelas 1 benar.

c) Recall:
    1. Kelas 0(sehat): 0.64 menunjukkan hanya 64% dari total kelas 0 yang terdeteksi.
    2. Kelas 1(sakit): 0.94 berarti 94% dari total kelas 1 terdeteksi.
d) F1-Score:
    1. Kelas 0(sehat): 0.75 menunjukkan keseimbangan moderat antara presisi dan recall.
    2. Kelas 1(sakit): 0.84 menunjukkan keseimbangan yang baik antara presisi dan recall.

3) Gradient Boosting:
a) Accuracy: 0.82 menunjukkan bahwa model Gradient Boosting mampu mengklasifikasikan 82% sampel dengan benar.
b) Precision:
    1. Kelas 0(sehat): 0.87 berarti 87% prediksi kelas 0 benar.
    2. Kelas 1(sakit): 0.79 berarti 79% prediksi kelas 1 benar.
c) Recall:
    1. Kelas 0(sehat): 0.71 menunjukkan hanya 71% dari total kelas 0 yang terdeteksi.
    2. Kelas 1(sakit): 0.91 berarti 91% dari total kelas 1(sakit) terdeteksi.
d) F1-Score:
    1. Kelas 0(sehat): 0.78 menunjukkan keseimbangan moderat antara presisi dan recall.
    2. Kelas 1(sakit): 0.85 menunjukkan keseimbangan yang baik antara presisi dan recall.

BAB VII ANALISA DAN PEMBAHASAN

Dari ketiga model yang diuji, Gradient Boosting memiliki performa terbaik dalam hal akurasi (0.82), diikuti oleh SVM (0.80) dan LDA (0.79). Berikut analisis lebih lanjut terhadap masing-masing model:

1. **LDA:**
   a. LDA cenderung kurang optimal dalam mendeteksi kelas 0, terlihat dari recall yang hanya mencapai 64%. Hal ini mungkin disebabkan oleh asumsi distribusi data yang kurang sesuai dengan kenyataan, karena LDA mengasumsikan data bersifat Gaussian.
   b. Precision untuk kelas 0 lebih tinggi daripada kelas 1(sakit), menunjukkan model lebih konservatif dalam memprediksi kelas 0.
2. **SVM:**
   a. SVM menunjukkan kinerja yang baik pada kelas 1(sakit) dengan recall mencapai 94%, yang berarti model sangat efektif dalam mendeteksi kelas 1(sakit).
   b. Namun, presisi untuk kelas 1(sakit) hanya 76%, menunjukkan bahwa model menghasilkan beberapa prediksi false positive untuk kelas ini.
3. **Gradient Boosting:**
   a. Gradient Boosting memiliki kombinasi terbaik antara precision dan recall, menghasilkan F1-score tertinggi di kedua kelas.
   b. Model ini lebih robust dalam menangani data yang tidak terdistribusi secara linear, yang mungkin menjadi alasan utama keunggulannya dibanding LDA dan SVM.

BAB IX KESIMPULAN

Berdasarkan eksperimen yang dilakukan dengan tiga model berbeda yaitu LDA, SVM, dan Gradient Boosting, dapat disimpulkan bahwa:

1. Gradient Boosting adalah model terbaik untuk data ini dengan akurasi 82%, diikuti oleh SVM (80%) dan LDA (79%).
2. Gradient Boosting mampu menangani data dengan distribusi yang kompleks, menghasilkan performa terbaik di kedua kelas, dengan F1-score 0.78 untuk kelas 0 (sehat) dan 0.85 untuk kelas 1(sakit).
3. SVM memberikan hasil yang memuaskan, terutama dalam mendeteksi kelas 1 (sakit) dengan recall 94%. Namun, presisi untuk kelas ini lebih rendah dibandingkan Gradient Boosting.
4. LDA, meskipun sederhana, tetap kompetitif dengan akurasi yang hanya sedikit lebih rendah dari model lainnya. Kelemahannya terletak pada asumsi distribusi data yang tidak selalu sesuai dengan kenyataan.

BAB X REFERENSI

[1] David Galih P., Muhammad Luthfi A., Muhammad Farhan J., dan Shulun Dwisiwi P., 2022. Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network.

[2] Deo Haganta D., Yuni Widiastiwi, dan Mayanda Mega S., 2022. Perbandingan Model Descision Tree, Naïve Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung.

[3] Donny Maulana dan Rezayadi Yahya. 2019. Implementasi Alogaritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung di Indonesia Menggunakan Rapid Miner.

[4] Ibnu Rashad, R Rizal Isnanto, dan Catur Edi Widodo. 2022. Klasifikasi Penyakit Jantung Menggunakan Algoritma Analisis Diskriminan Linier.

[5] Lutfi dan Muhammad Rafli Aulia Rojani. 2024. Analisis Penggunaan Teknik Oversampling Pada Extreme Gradient Boosting (XGBoost) Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Penyakit Jantung.

[6] Muhammad Dion F.T., Herliyani Hasanah, dan Tri Djoko S., 2023. Perbandingan Alogaritma Support Vecto Machine (SVM) dan Neural Network untuk Klasifikasi Penyakit Jantung.

[7] Syafitri Hidayatul A.A., Yuita Arum S., dan Achmad Arwan., 2018. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes.