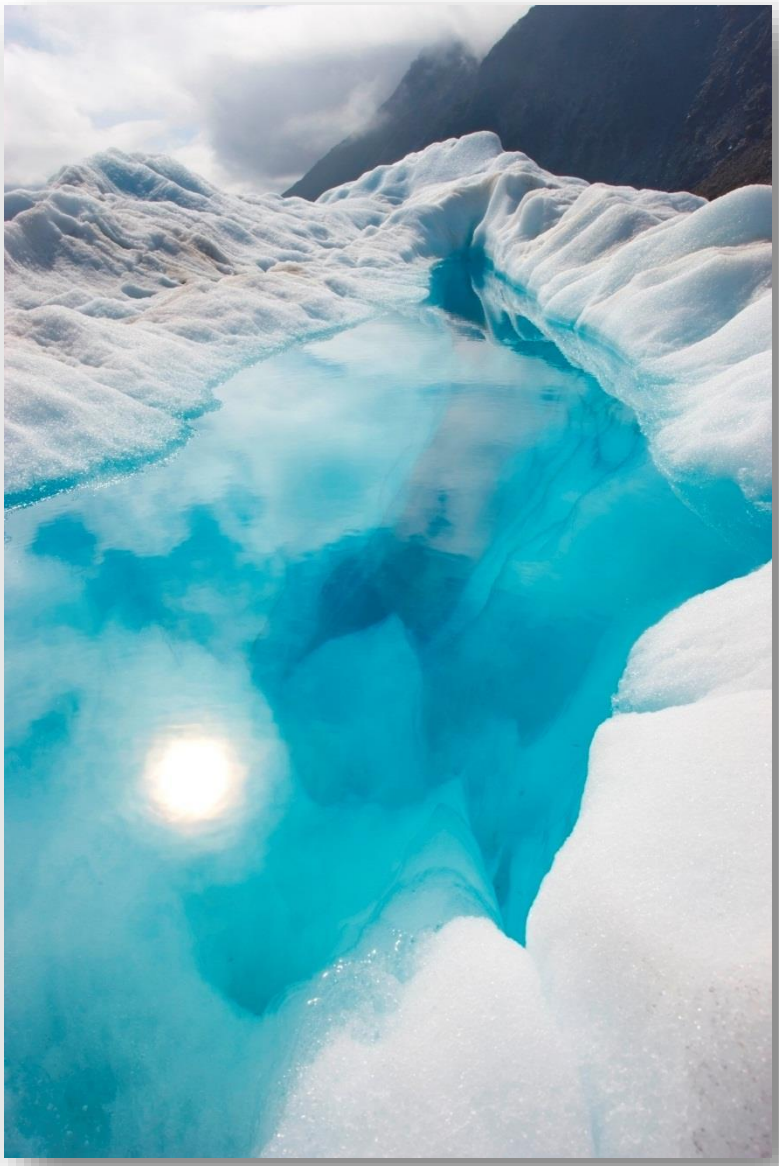
2023



BUKU KERJA/JOB SHEET

ASSOCIATE DATA SCIENTIST

Nama Peserta	:	Nurul Reny Agustin
Nomor Urut	:	

DAFTAR ISI

BUKTI 1-ADS

Kode Unit	:	J.62DMI00.004.1
Judul Unit	:	Mengumpulkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengumpulkan data untuk data science.

Langkah Kerja:

1) Menentukan kebutuhan data
2) Mengambil data
3) Mengintegrasikan data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengubah teks
  - Aplikasi basis data
  - Tools pengambilan data

1. KEBUTUHAN DATA

Instruksi Kerja:

- Identifikasi kebutuhan data sesuai tujuan teknis data science
- Periksa ketersediaan data berdasarkan kebutuhan data sesuai aturan yang berlaku
- Tentukan volume data berdasarkan kebutuhan data sesuai tujuan teknis data science

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score
```

```
#read data
df = pd.read_csv('heart.csv')
df
```

✓ 0.1s

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	NaN	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49.0	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37.0	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48.0	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54.0	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45.0	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68.0	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57.0	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57.0	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38.0	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

2. PENGAMBILAN DATA

Instruksi Kerja:

- Identifikasi metode dan tools pengambilan data sesuai tujuan teknis data science
- Tentukan tools pengambilan data sesuai tujuan teknis data science
- Siapkan tools pengambilan data sesuai tujuan teknis data science
- Jalankan proses pengambilan data sesuai dengan tools yang telah disiapkan

Cara pengumpulan data yang saya lakukan dengan mendownload dataset dari github yang sudah di sediakan oleh pelaksana. cara lain untuk pengumpulan data bisa menggunakan web scraping.

Sumber dataset: https://github.com/arubhasy/dataset/blob/main/heart.csv

Dalam pengolahan data ini saya menggunakan python untuk mengolah data heart.csv dan code editor yang saya gunakan adalah VS Code.

saya juga memakai beberapa library yang di sediakan oleh python dalam membantu saya membuat prediksi pada dataset ini.

3. PENGINTEGRASIAN DATA

Instruksi Kerja:

- Periksa integritas data sesuai tujuan teknis data sciene
- Integrasikan data sesuai tujuan teknis data science

DATASET STORY

Age = usia pasien (tahun)

sex = gender pasien(Male , Female)

ChestPainType = nyeri dada

`[TA: Typical Angina =  nyeri dada akibat aktivitas fisik ,`

`ATA: Atypical Angina = nyeri dada yang tidak berhubungan dengan jantung,`

`NAP: Non-Anginal Pain = mengacu pada nyeri yang mungkin dirasakan seseorang tanpa penyakit jantung,`

`ASY: Asymptomatic = kondisi ketika seseorang telah positif menderita suatu penyakit namun tidak menunjukkan gejala klinis apapun]`

RestingBP = tekanan darah [mm Hg]

Cholesterol = kolestrol

FastingBS = kadar gula darah puasa [1: if FastingBS > 120 mg/dl, 0: otherwise]

RestingECG = gula darah, r [1: if FastingBS > 120 mg/dl, 0: otherwise]

MaxHR = hasil EKG

`[Normal:  Normal,`

`ST: Hasil EKG menunjukkan adanya abnormalitas gelombang ST-T. Ini dapat berupa inversi gelombang T atau elevasi/depresi segmen ST > 0.05 mV,`

`LVH: Hasil electrocardiogram menunjukkan adanya hipertrofi ventrikel kiri yang mungkin atau pasti menurut kriteria Estes]`

ExerciseAngina = max denyut jantung antara 60 - 202

oldpeak = ST [Nilai numerik yang diukur pada depresi]

ST_Slope = kemiringan segmen ST pada EKG selama puncak latihan.

`[= Up: upsloping = naik,`

`Flat: flat = datar,`

`Down: downsloping = turun]`

HeartDisease = target [1: heart disease, 0: Normal]

BUKTI 2-ADS

Kode Unit	:	J.62DMI00.005.1
Judul Unit	:	Menelaah Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menelaah data untuk data science.

Langkah Kerja:

1) Menganalisis tipe dan relasi data
2) Menganalisis karakteristik data
3) Membuat laporan telaah data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Tools pengolahan data
  - Tools pembuat grafik

1. ANALISIS TIPE DAN RELASI DATA

Instruksi Kerja:

- Identifikasi tipe data yang terkumpul sesuai tujuan teknis
- Uraikan nilai atribut data yang terkumpul sesuai dengan batasan konteks bisnisnya
- Identifikasi relasi antar data yang terkumpul sesuai dengan tujuan teknis

```
df.info()
```
✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             911 non-null    float64
 1   Sex             908 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(2), int64(5), object(5)
memory usage: 86.2+ KB
```

Type data yang ada pada dataset ini yaitu : float, int, object. Dengan 12 kolom dan 918 baris data.

2. ANALISIS KARAKTERISTIK DATA

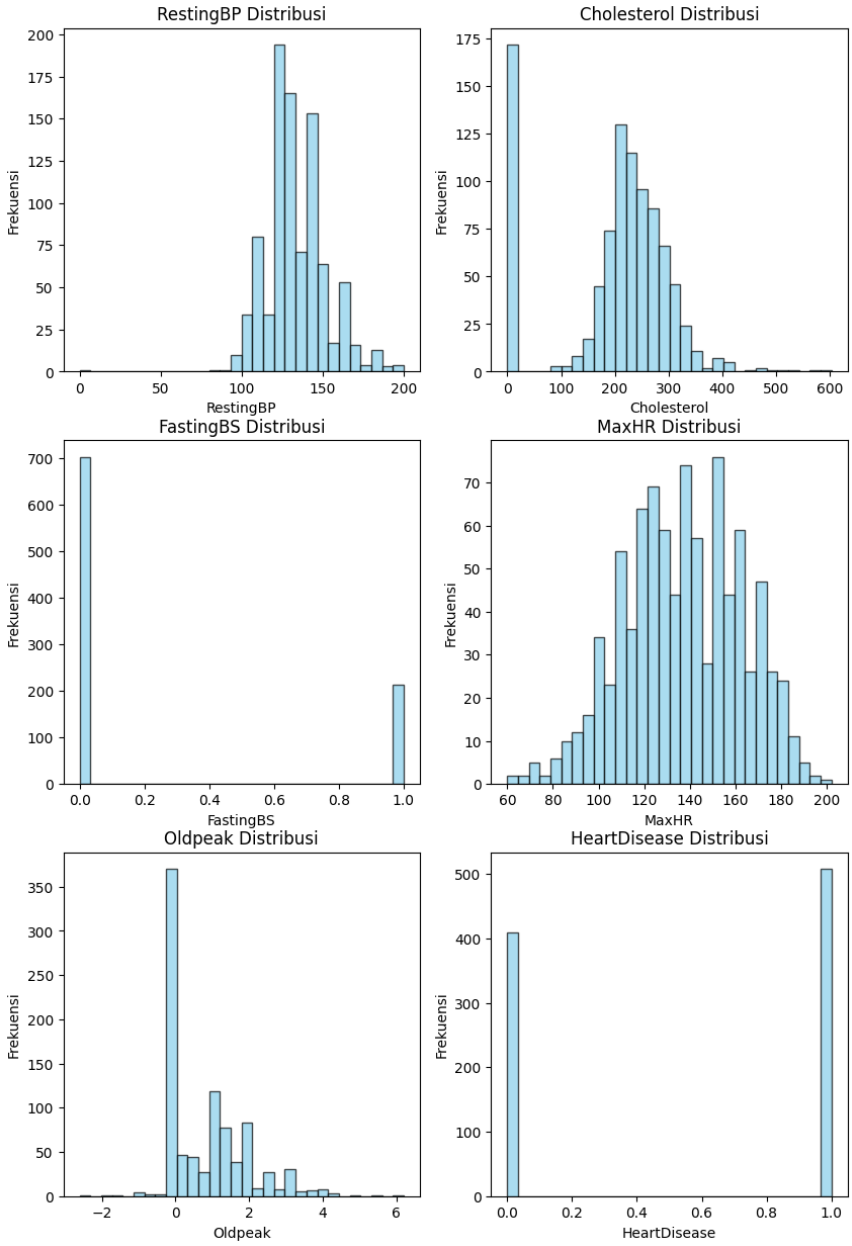
Instruksi Kerja:

- Sajikan karakteristik data yang terkumpul dengan deskripsi statistik dasar
- Sajikan karakteristik data yang terkumpul dengan visualisasi grafik
- Analisis karakteristik data dari hasil penyajian data untuk telaah data

```
df.describe().T
```
✓ 0.0s

	count	mean	std	min	25%	50%	75%	max
Age	911.0	54.102086	12.988393	0.0	47.00	54.0	60.0	177.0
RestingBP	918.0	132.396514	18.514154	0.0	120.00	130.0	140.0	200.0
Cholesterol	918.0	198.799564	109.384145	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	0.233115	0.423046	0.0	0.00	0.0	0.0	1.0
MaxHR	918.0	136.809368	25.460334	60.0	120.00	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.00	0.6	1.5	6.2
HeartDisease	918.0	0.553377	0.497414	0.0	0.00	1.0	1.0	1.0

```python
numerical_columns = df.select_dtypes(include=['int64', 'float64']).columns

# Plotting histograms
fig, axes = plt.subplots(len(numerical_columns) // 2 + len(numerical_columns) % 2, 2, figsize=(10, 15))

for i, column in enumerate(numerical_columns):
    ax = axes[i // 2, i % 2]
    df[column].plot(kind='hist', bins=30, ax=ax, alpha=0.7, color='skyblue', edgecolor='black')
    ax.set_title(f'{column} Distribusi')
    ax.set_xlabel(column)
    ax.set_ylabel('Frekuensi')

plt.show()
```



3. LAPORAN TELAAH DATA

Instruksi Kerja:

- Dokumentasikan hasil analisis dalam bentuk laporan sesuai dengan tujuan teknis
- Susun hipotesis berdasar hasil analisis sesuai tujuan teknis data science

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) mengalisis tipe dan relasi data; dan (2) menganalisis karakteristik data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat laporan telaah data; dapat diabaikan.

.....

.....

.....

.....

.....

BUKTI 3-ADS

Kode Unit	:	J.62DMI00.006.1
Judul Unit	:	Memvalidasi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memvalidasi data untuk data science.

Langkah Kerja:

1) Melakukan pengecekan kelengkapan data
2) Membuat rekomendasi kelengkapan data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengubah teks

1. PENGECEKAN KELENGKAPAN DATA

Instruksi Kerja:

- Sajikan penilaian kualitas data dari hasil telaah sesuai tujuan teknis data science
- Sajikan penilaian tingkat kecukupan data dari hasil telaah sesuai tujuan teknis data science

```
#cek missing value
df.isnull().sum()
```

✓ 0.0s

```
Age              7
Sex             10
ChestPainType    0
RestingBP        0
Cholesterol      0
FastingBS        0
RestingECG       0
MaxHR            0
ExerciseAngina   0
Oldpeak          0
ST_Slope         0
HeartDisease     0
dtype: int64
```

Terdapat data yang hilang atau missing value pada kolom Age dan Sex.

```python
#cek data yang duplikat
jumlah_duplikat = df.duplicated().sum()

# Mencetak jumlah data duplikat
print('data duplikat =', jumlah_duplikat)
```

✓ 0.0s

```
data duplikat = 0
```

2. REKOMENDASI KELENGKAPAN DATA

Instruksi Kerja:

- Susun rekomendasi hasil penilaian kualitas sesuai tujuan teknis data science
- Susun rekomendasi hasil penilaian kecukupan data sesuai tujuan teknis data science

```python
# Mengisi nilai yang hilang dalam kolom Age dengan median
df['Age'] = df['Age'].fillna(df['Age'].median())

# Mengisi nilai yang hilang dalam kolom Sex dengan mode
df['Sex'] = df['Sex'].fillna(df['Sex'].mode()[0])

df.isnull().sum()
```

✓ 0.0s

```
Age              0
Sex              0
ChestPainType    0
RestingBP        0
Cholesterol      0
FastingBS        0
RestingECG       0
MaxHR            0
ExerciseAngina   0
Oldpeak          0
ST_Slope         0
HeartDisease     0
dtype: int64
```

```
#mengubah type data pada kolom Age
df['Age'] = df['Age'].astype('int')
```

✓ 0.0s

```
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int32
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int32(1), int64(5), object(5)
memory usage: 82.6+ KB
```

BUKTI 4-ADS

Kode Unit	:	J.62DMI00.007.1
Judul Unit	:	Menentukan Objek Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilah dan memilih data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

1) Memutuskan kriteria dan teknik pemilihan data
2) Menentukan atrributes (columns) dan records (row) data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi spreadsheet
  - Aplikasi notepad plus
  - Aplikasi SQL (Structured Query Language)

1. KRITERIA DAN TEKNIK PEMILIHAN DATA

Instruksi Kerja:

- Identifikasi kriteria pemilihan data sesuai dengan tujuan teknis dan aturan yang berlaku
- Tetapkan teknik pemilihan data sesuai dengan kriteria pemilihan data

```
# Target Variable
df['HeartDisease'].unique()
```
✓ 0.0s

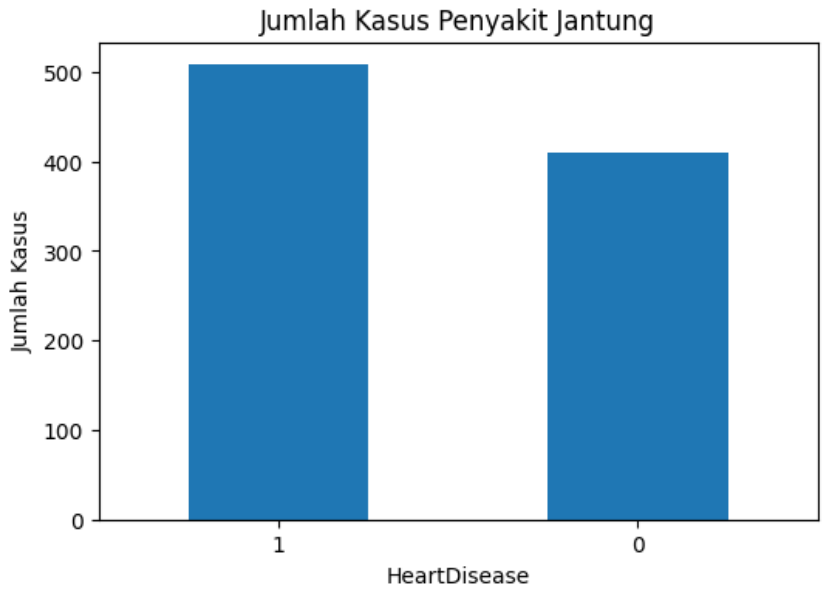
```
array([0, 1], dtype=int64)
```

```
df['HeartDisease'].value_counts()
```
✓ 0.0s

```
1    508
0    410
Name: HeartDisease, dtype: int64
```

```python
value_counts = df['HeartDisease'].value_counts()

# Membuat plot visualisasi
plt.figure(figsize=(6, 4))
value_counts.plot(kind='bar')
plt.title('Jumlah Kasus Penyakit Jantung')
plt.xlabel('HeartDisease')
plt.ylabel('Jumlah Kasus')
plt.xticks(rotation=0)
plt.show()
```

✓ 0.3s



2. ATTRIBUTES (COLUMNS) DAN RECORDS (ROW) DATA

Instruksi Kerja:

- Identifikasi attributes (columns) data sesuai dengan kriteria pemilihan data
- Identifikasi records (row) data sesuai dengan kriteria pemilihan data

Variabel dependent atau Target pada dataset ini yaitu ’HeartDisease’ karena untuk memprediksi apakah seorang pasien mengidap penyakit jantung atau tidak.

BUKTI 5-ADS

Kode Unit	:	J.62DMI00.008.1
Judul Unit	:	Membersihkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membersihkan data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

1) Melakukan pembersihan data yang kotor
2) Membuat laporan dan rekomendasi hasil membersihkan data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi spreadsheet
  - Aplikasi text editor
  - Aplikasi SQL (Structured Query Language)

1. PEMBERSIHAN DATA KOTOR

Instruksi Kerja:

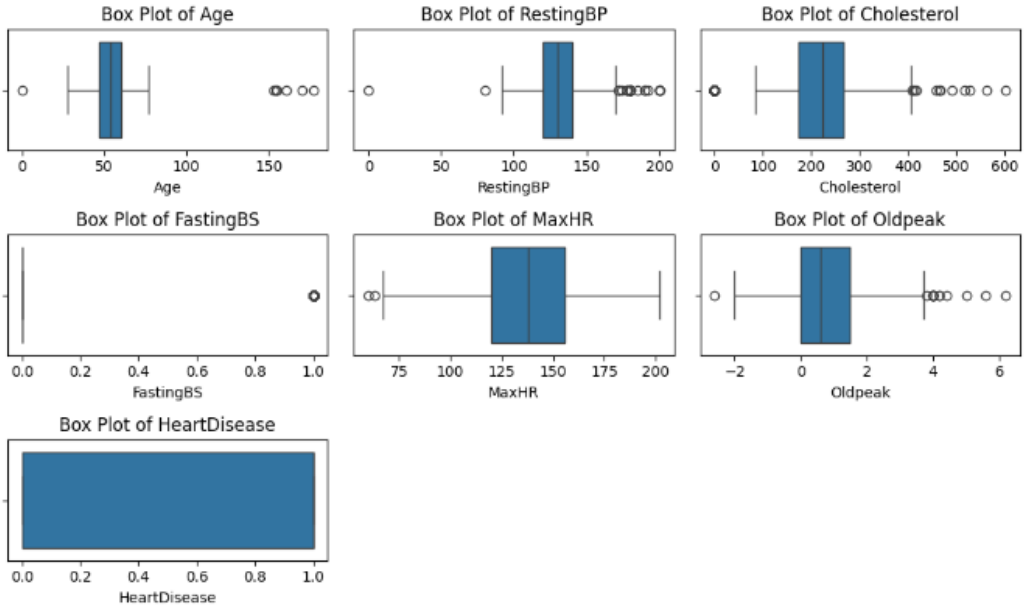
- Tentukan strategi pembersihan data berdasarkan hasil telaah data
- Koreksi data yang kotor berdasarkan strategi pembersihan data

```
# Daftar kolom numerik untuk pemeriksaan outlier
numeric_columns = ['Age', 'RestingBP', 'Cholesterol',
                   'FastingBS', 'MaxHR', 'Oldpeak', 'HeartDisease']

# Fungsi untuk mendeteksi outlier menggunakan IQR
def detect_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
    return outliers
```

✓ 0.0s

```python
# Deteksi dan cetak outlier untuk setiap kolom numerik
for column in numeric_columns:
    outliers = detect_outliers(df, column)
    #print(f"Outliers in {column}:")
    #print(outliers)
    #print("\n")

# Visualisasi outlier menggunakan box plot
plt.figure(figsize=(10,6))
for i, column in enumerate(numeric_columns, 1):
    plt.subplot(3, 3, i)
    sns.boxplot(x=df[column])
    plt.title(f'Box Plot of {column}')
plt.tight_layout()
plt.show()
```

✓ 1.7s



2. LAPORAN DAN REKOMENDASI HASIL PEMBERSIHAN DATA KOTOR

Instruksi Kerja:

- Deskripsikan masalah dan teknis koreksi data sesuai dengan kondisi data dan strategi pembersihan data
- Lakukan evaluasi berdasarkan analisis koreksi yang telah dilakukan
- Dokumentasikan evaluasi proses dan hasil pembersihan data kotor

```python
# fungsi untuk mengganti nilai outliers dengan median
def replace_outliers_with_median(data, col):
    Q1 = np.percentile(data[col].dropna(), 25)
    Q3 = np.percentile(data[col].dropna(), 75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    median = data[col].median()
    data[col] = np.where((data[col] < lower_bound) | (data[col] > upper_bound), median, data[col])
    return data
```

✓ 0.0s Python

```python
replaced_data = df.copy()
for col in numeric_columns:
    replaced_data = replace_outliers_with_median(replaced_data, col)
print("Data setelah di tangani:")
print(replaced_data.describe().T)
```

✓ 0.0s

```
Data setelah di tangani:
              count        mean        std   min    25%    50%    75%    max
Age           918.0   53.456427   9.357888  28.0   47.0   54.0   60.0   77.0
RestingBP     918.0  131.079521  15.597206  92.0  120.0  130.0  140.0  170.0
Cholesterol   918.0  237.442266  46.339984  85.0  214.0  223.0  264.0  407.0
FastingBS     918.0    0.000000   0.000000   0.0    0.0    0.0    0.0    0.0
MaxHR         918.0  136.976035  25.215656  67.0  120.0  138.0  156.0  202.0
Oldpeak       918.0    0.829412   0.958009  -2.0    0.0    0.6    1.5    3.7
HeartDisease  918.0    0.553377   0.497414   0.0    0.0    1.0    1.0    1.0
```

BUKTI 6-ADS

Kode Unit	:	J.62DMI00.009.1
Judul Unit	:	Mengkonstruksi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengkontruksi data untuk proyek data science.

Langkah Kerja:

1) Menganalisis teknik transformasi adta
2) Melakukan transformasi data
3) Membuat dokumentasi konstruksi data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Tools pengolah kata

1. ANALISIS TEKNIK TRANSFORMASI DATA

Instruksi Kerja:

- Lakukan analisis data untuk menentukan representasi fitur data awal
- Lakukan analisis representasi fitur data awal untuk menentukan teknik rekayasa fitur yang diperlukan untuk pembangunan model data science

menambahkan fitur baru HighCholesterol untuk mendeteksi pasien yang mempunyai value cholestrol diatas 200mg/dl

```
# Menambahkan fitur baru HighCholesterol
df['HighCholesterol'] = (df['Cholesterol'] > 200).astype(int)

#Menangani nilai yang hilang untuk fitur yang baru ditambahkan
df['HighCholesterol'].fillna(0, inplace=True)
```

✓ 0.0s

```python
# Mengubah fitur kategori menjadi numerik menggunakan LabelEncoder
categorical_features = ['Sex', 'ChestPainType', 'RestingECG',
                        'ExerciseAngina', 'ST_Slope']
label_encoders = {}

for feature in categorical_features:
    le = LabelEncoder()
    df[feature] = le.fit_transform(df[feature])
    label_encoders[feature] = le
df
```

✓ 0.1s Python

```python
numerical_features = ['Age', 'RestingBP', 'Cholesterol',
                      'MaxHR', 'Oldpeak', 'HighCholesterol']
scaler = StandardScaler()
df[numerical_features] = scaler.fit_transform(df[numerical_features])
```

✓ 0.0s

fitur scaling yang digunakan adalah standarisasi untuk memastikan bahwa semua fitur dalam dataset berada dalam skala yang konsisten

2. TRANSFORMASI DATA

Instruksi Kerja:

- Lakukan transformasi untuk mendapatkan fitur data awal
- Lakukan rekayasa fitur data untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease	HighCholesterol
54	1	1	140	289	0	1	172	0	0.0	2	0	1
49	0	2	160	180	0	1	156	0	1.0	1	1	0
37	1	1	130	283	0	2	98	0	0.0	2	0	1
48	0	0	138	214	0	1	108	1	1.5	1	1	1
54	1	2	150	195	0	1	122	0	0.0	2	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
45	1	3	110	264	0	1	132	0	1.2	1	1	1
68	1	0	144	193	1	1	141	0	3.4	1	1	0
57	1	0	130	131	0	1	115	1	1.2	1	1	0
57	0	1	130	236	0	0	174	0	0.0	1	1	1
38	1	2	138	175	0	1	173	0	0.0	2	0	0

3. DOKUMENTASI KONSTRUKSI DATA

Instruksi Kerja:

- Jabarkan teknis transformasi data dalam bentuk tertulis
- Tuangkan hasil transformasi data dan rekomendasi hasil transformasi dalam bentuk tertulis

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) mengalisis teknik transformasi data; dan (2) melakukan transformasi data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat dokumentasi konstruksi data; dapat diabaikan.

BUKTI 7-ADS

Kode Unit	:	J.62DMI00.010.1
Judul Unit	:	Menentukan Label Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menentukan label data untuk pembangunan model data science.

Langkah Kerja:

1) Melakukan pelabelan data
2) Membuat laporan hasil pelabelan data

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Aplikasi pengolah kata
  - Aplikasi pelabelan data

1. PELABELAN DATA

Instruksi Kerja:

- Uraikan kesesuaian antara analisis hasil pelabelan data sejenis yang sudah ada dengan Standard Operating Procedure (SOP) pelabelan
- Lakukan pelabelan data sesuai dengan SOP pelabelan

```
# Split fitur & target
X = df.drop('HeartDisease', axis=1)
y = df['HeartDisease']

# split data pelatihan dan test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

✓ 0.0s Python

2. LAPORAN HASIL PELABELAN DATA

Instruksi Kerja:

- Uraikan statistik hasil pelabelan pada laporan
- Uraikan evaluasi proses pelabelan pada laporan

x = Variabel ini menyimpan data fitur (independen) dengan menghapus kolom HeartDisease dari DataFrame . Semua kolom lain di df dianggap sebagai fitur yang akan digunakan untuk memprediksi target.

y = Variabel ini menyimpan target (dependen) yaitu kolom HeartDisease. Ini adalah variabel yang ingin kita prediksi menggunakan fitur-fitur di variabel x.

test_size=0.2: Menunjukkan bahwa 20% dari data akan digunakan sebagai data uji, sedangkan 80% akan digunakan sebagai data pelatihan.

random_state=42: Menentukan seed untuk generator angka acak. Ini memastikan bahwa hasil pembagian data selalu sama setiap kali kode dijalankan.

BUKTI 8-ADS

Kode Unit	:	J.62DMI00.013.1
Judul Unit	:	Membangun Model

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membangun model.

Langkah Kerja:

1) Menyiapkan parameter model
2) Menggunakan tools pemodelan

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer dan peralatannya
  - Perangkat lunak data science di antaranya: rapid miner, weka, atau development untuk bahasa pemrograman tertentu seperty Python atau R.
- Perlengkapan
  - Dokumen best practices kriteria dan evaluasi penilaian

1. PARAMETER MODEL

Instruksi Kerja:

- Identifikasi parameter-parameter yang sesuai dengan model
- Tetapkan nilai toleransi parameter evaluasi pengujian sesuai dengan tujuan teknis

Model yang saya buat yaitu decision tree & logistic regression.

Membuat model Decision Tree dengan parameter random_state diatur ke 42 untuk memastikan hasil yang sama setiap kali kode dijalankan.

Sedangkan pada model logistic regression saya menggunakan parameter max_iter=1000 untuk memastikan model dapat melakukan iterasi yang cukup hingga menemukan solusi yang cukup baik dan random_state=42 untuk memastikan hasil yang sama setiap kali kode dijalankan.

2. TOOLS PEMODELAN

Instruksi Kerja:

- Identifikasi tools untuk membuat model sesuai dengan tujuan teknis data science
- Bangun algoritma untuk teknik pemodelan yang ditentukan menggunakan tools yang dipilih
- Eksekusi algoritma pemodelan sesuai dengan skenario pengujian dan tools untuk membuat model yang telah ditetapkan

- Optimasi parameter model algoritma untuk menghasilkan nilai parameter evaluasi yang sesuai dengan skenario pengujian

```
# Membuat dan melatih model decision tree
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)

# Membuat dan melatih model logistic regression
lr_model = LogisticRegression(max_iter=1000, random_state=42)
lr_model.fit(X_train, y_train)

# memprediksi kedua model
y_pred_dt = dt_model.predict(X_test)
y_pred_lr = lr_model.predict(X_test)
```

BUKTI 9-ADS

Kode Unit	:	J.62DMI00.014.1
Judul Unit	:	Mengevaluasi Hasil Pemodelan

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengevaluasi hasil pemodelan.

Langkah Kerja:

1) Menggunakan model dengan data riil
2) Menilai hasil pemodelan

Peralatan dan Perlengkapan:

- Peralatan
  - Komputer
- Perlengkapan
  - Tools untuk mengeksekusi model
  - Tools untuk pengumpulan data riil

1. PENGGUNAAN MODEL DENGAN DATA RIIL

Instruksi Kerja:

- Kumpulkan data baru untuk evaluasi pemodelan sesuai kebutuhan yang mengacu kepada parameter evaluasi
- Uji model dengan menggunakan data riil yang telah dikumpulkan

```
# Menghitung metrik evaluasi untuk model decision tree
conf_matrix_dt = confusion_matrix(y_test, y_pred_dt)
accuracy_dt = accuracy_score(y_test, y_pred_dt)
precision_dt = precision_score(y_test, y_pred_dt)
recall_dt = recall_score(y_test, y_pred_dt)
f1_dt = f1_score(y_test, y_pred_dt)

print("Decision Tree Metrics:")
print("Confusion Matrix:\n", conf_matrix_dt)
print("Accuracy:", accuracy_dt)
print("Precision:", precision_dt)
print("Recall:", recall_dt)
print("F1 Score:", f1_dt)
```

✓ 0.0s

```python
# Menghitung metrik evaluasi untuk model logistic regression
conf_matrix_lr = confusion_matrix(y_test, y_pred_lr)
accuracy_lr = accuracy_score(y_test, y_pred_lr)
precision_lr = precision_score(y_test, y_pred_lr)
recall_lr = recall_score(y_test, y_pred_lr)
f1_lr = f1_score(y_test, y_pred_lr)

print("\nLogistic Regression Metrics:")
print("Confusion Matrix:\n", conf_matrix_lr)
print("Accuracy:", accuracy_lr)
print("Precision:", precision_lr)
print("Recall:", recall_lr)
print("F1 Score:", f1_lr)
```

✓ 0.0s

2. PENILAIAN HASIL PEMODELAN

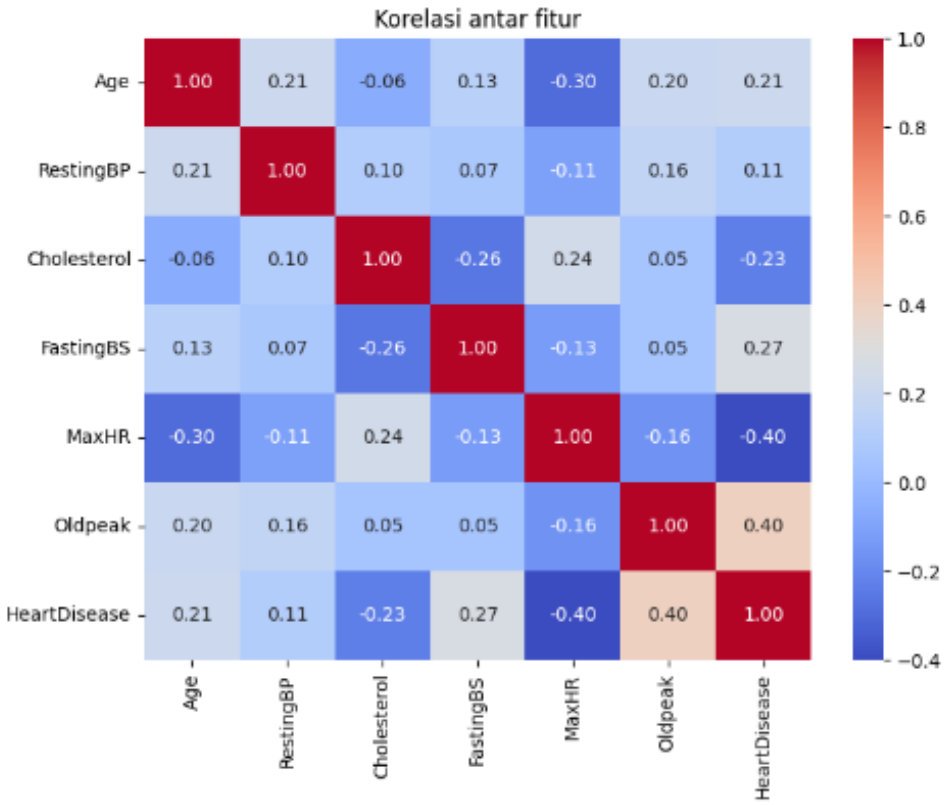
Instruksi Kerja:

- Nilai keluaran pengujian model berdasarkan metrik kesuksesan
- Dokumentasikan hasil penilaian sesuai standar yang berlaku

```
Decision Tree Metrics:
Confusion Matrix:
 [[64 13]
 [18 89]]
Accuracy: 0.8315217391304348
Precision: 0.8725490196078431
Recall: 0.8317757009345794
F1 Score: 0.8516746411483254


Logistic Regression Metrics:
Confusion Matrix:
 [[68  9]
 [20 87]]
Accuracy: 0.842391304347826
Precision: 0.90625
Recall: 0.8130841121495327
F1 Score: 0.8571428571428571
```

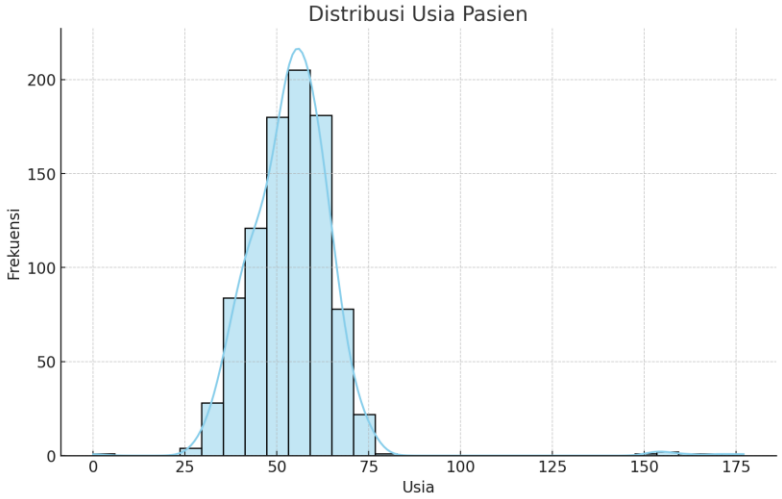
Kedua model memiliki performa yang baik untuk prediksi penyakit jantung. Namun Model decision tree mempunyai recall yang sedikit lebih tinggi dari model logistic regression.

INSIGHT DARI HASIL ANALISIS



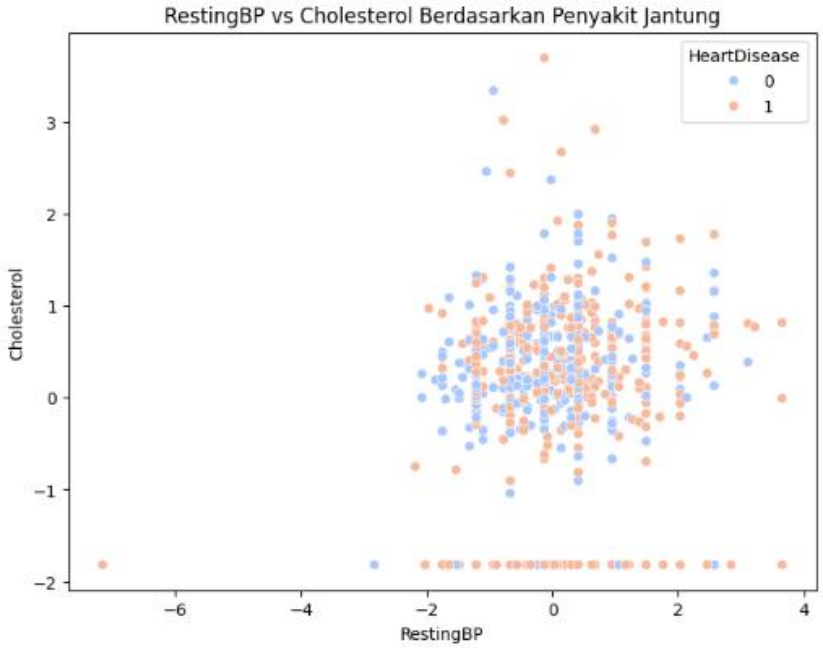
- **Oldpeak**: Korelasi positif yang kuat dengan penyakit jantung (0.40), menunjukkan bahwa depresi ST yang lebih tinggi berkaitan dengan risiko penyakit jantung yang lebih tinggi.
- **FastingBS**: Korelasi positif sedang (0.27), menunjukkan bahwa kadar gula darah puasa yang lebih tinggi mungkin berkaitan dengan risiko penyakit jantung yang lebih tinggi.
- **Age dan MaxHR**: Korelasi negatif (-0.30), menunjukkan bahwa detak jantung maksimum cenderung menurun seiring bertambahnya usia.
- **MaxHR**: Korelasi negatif yang kuat dengan penyakit jantung (-0.40), menunjukkan bahwa detak jantung maksimum yang lebih tinggi berkaitan dengan risiko penyakit jantung yang lebih rendah.

1. Berapa rata-rata usia pasien yang mempunyai penyakit jantung?



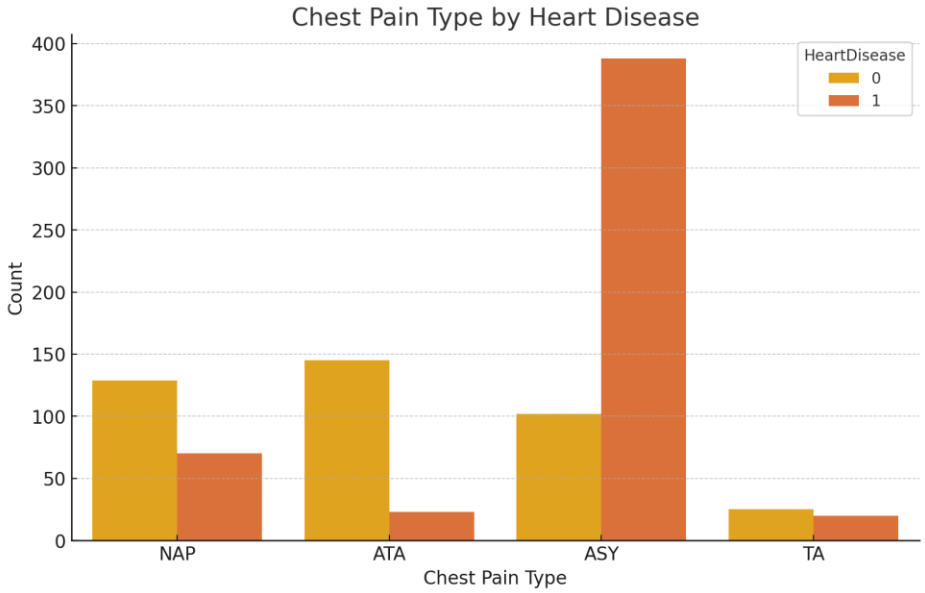
Berdasarkan diagram ini bisa di lihat bahwa usia pasien dengan penyakit jantung berkisar 28 – 75 tahun. Dengan puncak distribusi ada di usia 50 – 55 tahun.

2. Bagaimana hubungan antara kadar kolesterol (Cholesterol) dan tekanan darah istirahat (RestingBP) dengan status penyakit jantung?

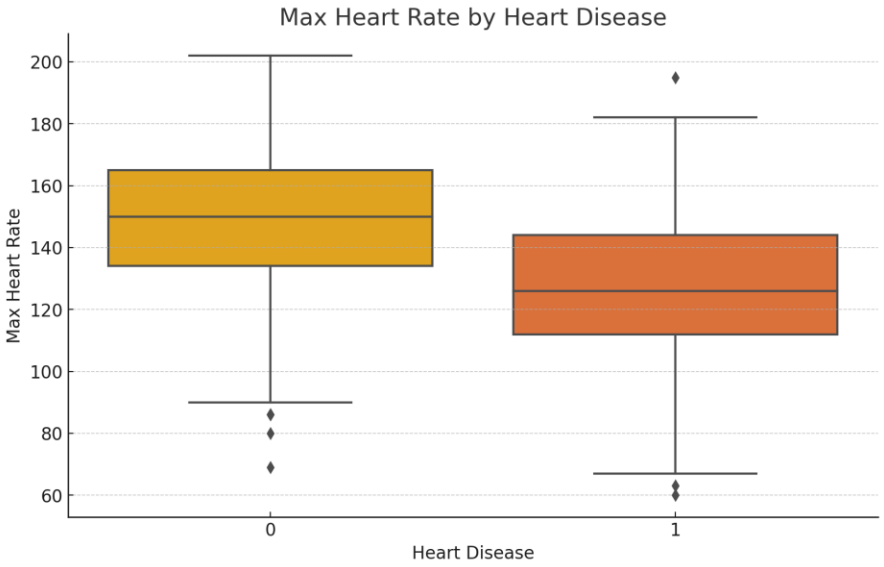


tidak ada pola yang jelas terlihat antara tekanan darah istirahat dan kadar kolesterol berdasarkan status penyakit jantung. Kedua fitur ini tidak menunjukkan korelasi yang signifikan dalam kaitannya dengan penyakit jantung dalam dataset ini.

3. Tipe nyeri dada yang mana yang biasanya banyak dialami oleh pasien dengan penyakit jantung?



nyeri dada 'ASY' (Asymptomatic) atau kondisi ketika seseorang telah positif menderita suatu penyakit namun tidak menunjukkan gejala klinis apapun lebih sering terjadi pada individu dengan penyakit jantung.



Orang tanpa penyakit jantung cenderung memiliki denyut jantung maksimal yang lebih tinggi, mengindikasikan bahwa kemampuan jantung untuk bekerja keras lebih baik pada individu sehat.