

Pembelajaran Mesin

Data Preparation dari

Sumber Open Source

Oleh:

Kelompok 2

Alvia Zuhra
Al-Mahfuzh Fadhlur Rohman
Nurul Uzratun Nashriyyah
Ganang Setyo Hadi
Azimah Al-Huda

(220810701003)
(220810701006)
(2208107010030)
(2208107010052)
(2208107010069)

Data Description

Nama Dataset

Heart Failure Prediction Dataset

Format Dataset

CSV

Informasi Fitur

11 fitur yaitu (1) Age; (2) Sex; (3) ChestPainType; (4) RestingBP; (5) Cholesterol; (6) FastingBS; (7) RestingECG; (8) MaxHR; (9) ExerciseAngina; (10) Oldpeak; (11) ST_Slope

Deskripsi

Link Dataset

Heart Failure Prediction Dataset adalah kumpulan data terbesar untuk penelitian penyakit jantung, dibuat dengan menggabungkan lima dataset sebelumnya. Dataset ini bertujuan membantu pengembangan model pembelajaran mesin untuk deteksi dan penanganan dini penyakit kardiovaskular, khususnya gagal jantung sebagai penyebab utama kematian global.

Label

HeartDisease dengan kelas output:
[1: heart disease, 0: Normal]

Data Loading

Langkah Awal Analisis

Memuat dataset Heart Failure Prediction ke dalam Python menggunakan Pandas untuk analisis data.

Membaca File

```
df = pd.read_csv('dataset/heart.csv')
```

Kesimpulan

Dataset Heart Failure Prediction berhasil dimuat ke dalam DataFrame Pandas, menyelesaikan langkah awal dalam analisis data. Kini, dataset siap untuk eksplorasi, pemahaman, dan preprocessing sebelum membangun model prediktif penyakit jantung.

Kelebihan Pandas dalam Data Loading

- ✓ Efisiensi
- ✓ Fleksibilitas
- ✓ Penanganan Tipe Data Otomatis
- ✓ Penanganan Nilai Kosong
- ✓ Integrasi dengan Ekosistem Data Science

Data Understanding

Mengecek ukuran dataset

```
df.shape
```

Menampilkan kolom

```
df.columns
```

Menampilkan informasi ringkasan dataset

```
df.info()  
  
<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    int64    
 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), int64(6), object(5)  
 memory usage: 86.2+ KB
```

Menampilkan statistik deskriptif dataset

```
df.describe()  
  
Age      RestingBP  Cholesterol  FastingBS  MaxHR  Oldpeak  HeartDisease  
count  918.000000  918.000000  918.000000  918.000000  918.000000  918.000000  918.000000  
mean   53.510893  132.396514  198.799564  0.233115  136.809368  0.887364  0.553377  
std    9.432617   18.514154  109.384145  0.423046  25.460334  1.066570  0.497414  
min    28.000000  0.000000   0.000000  0.000000  60.000000 -2.600000  0.000000  
25%   47.000000  120.000000  173.250000  0.000000  120.000000  0.000000  0.000000  
50%   54.000000  130.000000  223.000000  0.000000  138.000000  0.600000  1.000000  
75%   60.000000  140.000000  267.000000  0.000000  156.000000  1.500000  1.000000  
max   77.000000  200.000000  603.000000  1.000000  202.000000  6.200000  1.000000
```

Menampilkan distribusi variabel target

```
# Count of target variable (heart disease presence)  
print("\nDistribution of target variable (HeartDisease):")  
display(df['HeartDisease'].value_counts())  
print(f"Percentage of patients with heart disease: {df['HeartDisease'].mean()*100:.2f}%")  
  
Distribution of target variable (HeartDisease):  
  
HeartDisease  
1    508  
0    410  
Name: count, dtype: int64  
  
Percentage of patients with heart disease: 55.34%
```

Visualisasi Data yang digunakan

01

Menampilkan distribusi variabel kategorikal menggunakan Histogram

02

Menampilkan distribusi fitur numerik menggunakan Histogram

03

Analisis korelasi dataset penyakit jantung menggunakan Heatmap

04

Menampilkan perbandingan beberapa fitur menggunakan Boxplot

Data Preparation

01

Menangani Nilai Abnormal

Nilai 0 pada kolesterol diatasi dengan imputasi median terstratifikasi untuk menghindari bias. Outlier ditangani dengan Capping IQR Termodifikasi ($2.5 \times \text{IQR}$) agar data tetap akurat tanpa distorsi.

02

Encoding Fitur Kategorikal

Kami menggunakan encoding optimal untuk fitur kategorikal guna menjaga esensi medis, meningkatkan efisiensi, dan mendukung analisis prediktif.

03

Feature Scaling/ Normalization

Feature scaling menyeimbangkan skala fitur agar model tidak bias. Kami menyesuaikan metode scaling dengan distribusi tiap fitur untuk akurasi optimal.

04

Mengekspor Dataset

```
[ ] 1 scaled_df.to_csv('dataset/processed/df_ready.csv', index=False)
```

Kesimpulan

- 01 Dataset Heart Failure Prediction dari Kaggle dengan 918 sampel dan 12 fitur, digunakan untuk prediksi penyakit jantung
- 02 Dataset dimuat dalam format CSV menggunakan pandas, dengan bantuan NumPy, Matplotlib, dan Seaborn untuk analisis
- 03 Analisis dilakukan menggunakan df.shape, df.info, dan df.describe untuk melihat ukuran, tipe data, dan statistik dasar. Distribusi data dan hubungan antar fitur divisualisasikan menggunakan boxplot, histogram, dan heatmap korelasi.
- 04 Gunakan imputasi median terstratifikasi untuk nilai nol, capping IQR termodifikasi untuk outlier, One-Hot Encoding dan Label Encoding untuk fitur kategorikal, serta Min-Max Scaling atau Standardization untuk normalisasi data

THANKYOU

By: Kelompok 2