LAPORAN PRAKTIKUM PREPROCESSING DATA



DISUSUN OLEH:

Deasyana Prakasi (09011282328050)

DOSEN PENGAMPU:

Anggun Islami M.Kom

PROGRAM STUDI SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2024

1. Pendahuluan

Pra-pemrosesan data merupakan salah satu langkah yang sangat krusial dalam proses penambangan data. Langkah ini mencakup serangkaian kegiatan seperti mempersiapkan dan mentransformasi data agar sesuai dengan kebutuhan proses penambangan. Tujuan utama dari pra-pemrosesan data adalah untuk merampingkan ukuran data, mengidentifikasi hubungan antar data, melakukan normalisasi, menghilangkan outlier, serta mengekstraksi fitur-fitur penting dari data yang ada.

Dalam proses ini, berbagai teknik digunakan untuk memastikan data dalam kondisi optimal sebelum dianalisis lebih lanjut. Beberapa teknik yang umum dipakai dalam pra-pemrosesan data meliputi pembersihan data, integrasi data, transformasi, serta pengurangan data. Pembersihan data berfungsi untuk menangani data yang tidak lengkap atau tidak konsisten, sementara normalisasi dan pengurangan data bertujuan untuk memastikan bahwa data lebih mudah diolah dan dianalisis (Alasadi, S. A., & Bhaya, W. S., 2017).

2. Tujuan

- Melakukan eksplorasi awal terhadap dataset.
- Meningkatkan kualitas data dengan membersihkan data dari kesalahan, seperti missing data atau outlier Mengubah variabel kategori menjadi variabel numerik.
- Melakukan normalisasi atau standardisasi pada data numerik.
- Membuat data lebih mudah dibaca.

3. Alat dan Bahan

- Dataset yang digunakan: Heart Disease UCI dari Kaggle
- Jupyter notebook

4. Langkah-Langkah Data Preprocessing

4.1. Import dan setup

Pada bagian ini, kita memasukkan library yang akan digunakan

- a. Pandas: Digunakan untuk mengolah dan menganalisis data dalam format DataFrame.
- b. numpy: Menyediakan fungsi matematika dan memungkinkan kita bekerja dengan array multidimensi.
- c. Seaborn untuk membuat grafik statistik yang informatif dan estetis
- d. Matplotlib.pyplot digunakan untuk berbagai keperluan visualisasi data seperti pembuatan grafik, subplot.
- e. from sklearn.preprocessing import StandardScaler: Digunakan untuk menstandarisasi fitur data sehingga memiliki rata-rata 0 dan deviasi standar 1.

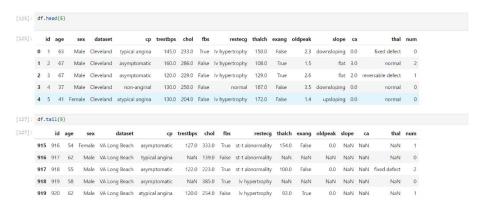
4.2. Membaca data

- a. file path untuk menyimpan lokasi data
- b. df untuk menyimpan pd.read csv untuk membaca file csv

```
[123]: file_path = r'C:\Users\User\Downloads\archive\heart_disease_uci.csv'
df = pd.read_csv(file_path)
```

4.3. Menampilkan informasi dataset

a. 5 baris atas dan bawah

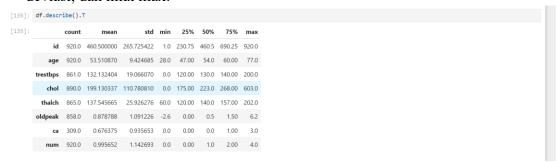


b. Menampilkan jumlah baris dan kolom



c. Menampilkan informasi tipe data, jumlah non-null, dan ukuran memori yang digunakan di setiap kolom.

d. Menampil gambaran umum tentang distribusi data, seperti nilai rata-rata, standar deviasi, dan nilai max.



4.4. Memeriksa baris duplikat

```
[112]: df.duplicated().sum()
[112]: 0
```

4.5. Menghapus kolom

a. Menampilkan semua nama kolom pada dataset

```
b. Menghapus kolom id
 df = df.drop(columns=['id'], errors='ignore')
 df.columns
 Index(['age', 'sex', 'dataset', 'cp', 'trestbps', 'chol', 'fbs', 'restecg',
       'thalch', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num'],
      dtype='object')
```

4.6. Missing Value

exang oldpeak slope thal num 6

a. Menampilkan/Memeriksa nilai missing value

```
[909]: df.isnull().sum()
[909]: age
         sex
dataset
         cp
trestbps
chol
          fbs
                          90
         restecg
thalch
                          55
55
62
         exang
oldpeak
          slope
                         309
         ca
thal
```

b. Mengisi nilai missing value dengan mean dan mode dan menampilkan hasilnya

```
[929] # Mengisi nilai missing value dengan mean

(f'trestbps'] = df['trestbps'].fillna(df['trestbps'].mean())

df['chol'] = df['thol'].fillna(df['chol'].mean())

df['thol'] = df['thol'].fillna(df['thol'].mean())

df['thol'] = df['thol'].fillna(df['thol'].mean())

df['tholh'] = df['thol'].fillna(df['tholh'].mean())

df['ca'] = df['ca'].fillna(df['ca'].mean())

# Mengisi nilai missing value dengan nilai mode

df['restecg'] = df['exang'].fillna(df['restecg'].mode()[0])

df['slope'] = df['exang'].fillna(df['slope'].mode()[0])

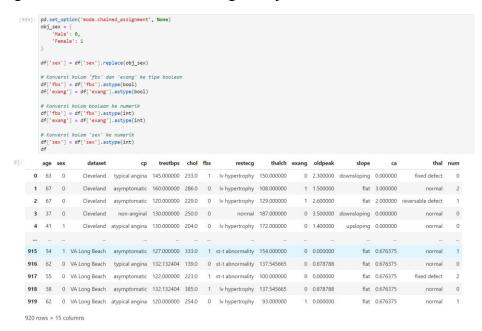
df['slope'] = df['slope'].fillna(df['thol'].mode()[0])

df['thol'] = df['thol'].fillna(df['thol'].mode()[0])

df.'snull().suu()
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         ★ ① ↑ ↓ 占 ♀ î
                                         df.isnull().sum()
                                         dataset
                                           cp
trestbps
                                           chol
fbs
                                         restecg
thalch
```

4.7. Mengubah kolom menjadi numerik

Mengubah Kolom "sex", "fbs", "exang" Menjadi Numerik



4.8. Normalisasi Data



4.9. Outliers

a. Cek Outliers

b. Mengatasi Outliers [1839]: def fill_outliers(df, columns):

4.10. Visuliasi

dtype: int64

a. Menampilkan Visualisasi Box Plot

```
[1043]: # Visualizasi Box Plot
df.plot(kind*box', subplots*True, layout=(len(df.columns)//2, 2), figsire=(10, 15))
plx.tight_layout()
plx.supritle('\n' \n Visualizasi "BOX PLOT" \n \n', fontsize=12)
plx.tight_layout()
plt.show()
```

b. Menampilkan Histogram

```
[1044]: # Visualisasi Histogram

of.hist(figsize(8, 12))

plt.tight_layout()

plt.suptile('\n' \n Visualisasi "HISTOGRAM" \n \n', fontsize=12)

plt.tight_layout()

plt.show()
```

5. Hasil dan Analisis

- a. Tidak ada baris yang duplikat.
- b. Menghapus kolom id.
- c. Beberapa kolom yang memiliki missing value yaitu trestbps, chol, fbs, restecg, thalch, exang, oldpeak, slope, ca thal. Nilai missing Value telah diisi dengan mean dan mode.
- d. Mengubah Kolom sex, fbs, dan exang menjadi numerik. Kolom fbs dan exang diubah terlebih dahulu menjadi boolean kemudian baru dijadikan numerik.
- e. Kolom yang outliers yaitu trestbps, chol, fbs, restecg, thalch, exang, oldpeak, slope, ca thal.
- f. Analisis pada Box Plot
 - Variabel Numerix yaitu sex, fbs, exang, dan ca tetap menunjukkan distribusi terbatas setelah normalisasi karena hanya memiliki dua nilai, yaitu 0 dan 1. Pada boxplot, variabel ini terlihat memiliki sedikit atau bahkan tanpa variasi.
 - Variabel kontinu seperti age, trestbps, chol, thalach, dan oldpeak menunjukkan variasi yang signifikan dalam data. Normalisasi yang diterapkan sudah bekerja dengan baik untuk mengatur variasi data, tetapi ada beberapa outlier yang muncul.
 - Variabel ca menunjukkan nilai yang hampir semuanya mendekati nol. Ini menunjukkan bahwa nilai ca tidak tersebar secara normal.

g. Analisis Pada Histogram

Age

Data diambil pada kelompok usia tua dengan mayoritas berusia 50-60.

Sex

menunjukkan data di dominasi oleh pria.

Trestbps

Ada beberapa yang memiliki trestbps melebihi rata rata tetapi lebih dominan trestbps di rata-rata normal.

Chol

Kebanyakan data chol berada di angka rendah, jadi sebagian besar orang punya kadar chol normal.

• Fbs

Sebagian besar bernilai 0 yang berarti mayoritas memiliki fbs yang normal.

• Thalach

Kebanyakan orang detak jantung maksimal yang normal, tapi ada juga yang detak jantungnya lebih tinggi dari biasanya.

Exang

Lebih banyak orang yang tidak mengalami exang dan hanya sebagian kecil yang mengalami exang.

Oldpeak

Sebagian besar orang memiliki nilai oldpeak rendah, yang menandakan depresi ST ringan. Namun, ada beberapa yang memiliki nilai lebih tinggi, kemungkinan menunjukkan masalah jantung lebih serius.

• Ca

Kebanyakan orang dalam dataset tidak memiliki pembuluh darah yang terdeteksi oleh fluoroskopi, menunjukkan data terpusat pada satu nilai. Ini bisa berarti mayoritas tidak mengalami penyumbatan pembuluh darah yang serius.

Num

Mayoritas data berada pada kategori 0 (tidak ada penyakit jantung), tetapi ada beberapa yang memiliki penyakit dengan tingkat keparahan berbeda/serius.

6. Kesimpulan

data preprocessing merupakan langkah penting dalam analisis data untuk membersihkan dan menyiapkan data sebelum digunakan dalam analisis lebih lanjut atau model machine learning. Melalui serangkaian proses, seperti mengecek baris duplikat, menangani missing values dengan mengisi nilai yang hilang, mengubah beberapa variabel menjadi numerik, dan menghapus outlier walaupun pada praktikum kali ini pelaksanaab outlier belum baik, dan melakukan normalisasi atau standardisasi pada data.

7. Refrensi

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.