A11.2021.13656

ILHAM SAFITRO

DATA MINING

A11.4607

Baik, berikut adalah contoh penerapan tahapan preprocessing data pada dataset "Iris" menggunakan bahasa pemrograman Python:

python

import pandas as pd

from sklearn.preprocessing import StandardScaler

# Load dataset

df = pd.read\_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data',

names=['sepal length', 'sepal width', 'petal length', 'petal width', 'class'])

# Drop missing values

df.dropna(inplace=True)

# Split features and labels

X = df.iloc[:, :-1]

y = df.iloc[:, -1]

# Standardize features

scaler = StandardScaler()

X\_std = scaler.fit\_transform(X)

Penjelasan dari masing-masing tahapan preprocessing data di atas adalah sebagai berikut:

Load dataset

Dataset "Iris" diambil dari UCI Machine Learning Repository.

Dataset ini terdiri dari 150 sampel data bunga iris dengan 4 fitur (panjang dan lebar kelopak dan mahkota) dan 1 target kelas (jenis bunga iris).

Dataset ini dibaca dan dimuat sebagai dataframe menggunakan library Pandas.

Drop missing values

Jika ada data yang kosong atau hilang, kita dapat menggunakan metode .dropna() untuk menghapus baris atau kolom yang mengandung data kosong.

Split features and labels

Setelah dataset dimuat, kita perlu memisahkan fitur (features) dari label (target) untuk proses selanjutnya.

Standardize features

Kita dapat menggunakan teknik standardisasi untuk membuat semua fitur memiliki skala yang sama dan mengurangi varians yang tidak diinginkan pada model machine learning.

Teknik standardisasi ini dilakukan dengan mengurangi nilai rata-rata dan membagi dengan standar deviasi setiap fitur.

Library Scikit-learn menyediakan class StandardScaler yang dapat digunakan untuk melakukan standardisasi pada dataset.