LAPORAN TUGAS MANDIRI MACHINE LEARNING

Nama : **Muhamad aditia**
Nim : **0110224213**
Rombel : **TI02**
Link Git : **https://github.com/muhammadaditia433/Machine-Leraning**

SEKOLAH TINGGI TEKNOLOGI TERPADU NURUL FIKRI

PROGRAM STUDI TEKNIK INFORMATIKA

DEPOK 2025

ABSTRAK

Dataset medis umumnya memiliki jumlah fitur yang besar sehingga dapat meningkatkan kompleksitas komputasi dan menyulitkan proses analisis data. Salah satu metode yang dapat digunakan untuk mengatasi permasalahan tersebut adalah Principal Component Analysis (PCA). Penelitian ini bertujuan untuk menerapkan PCA pada Breast Cancer Wisconsin (Diagnostic) Dataset guna memahami penerapan reduksi dimensi pada dataset medis berdimensi tinggi. Dataset ini terdiri dari 569 sampel dengan 30 fitur numerik yang merepresentasikan karakteristik sel tumor payudara, dengan dua kelas diagnosis yaitu benign dan malignant.

Tahapan penelitian meliputi proses standarisasi data, penerapan PCA untuk mereduksi jumlah fitur menjadi dua komponen utama, visualisasi hasil transformasi, serta evaluasi performa klasifikasi menggunakan algoritma Logistic Regression. Hasil penerapan PCA menunjukkan bahwa dua komponen utama mampu merepresentasikan sebagian besar variansi data dan memberikan visualisasi pemisahan kelas yang cukup jelas. Meskipun terjadi sedikit penurunan akurasi dibandingkan dengan penggunaan seluruh fitur asli, PCA terbukti efektif dalam mengurangi dimensi data dan menyederhanakan proses analisis tanpa kehilangan informasi yang signifikan.

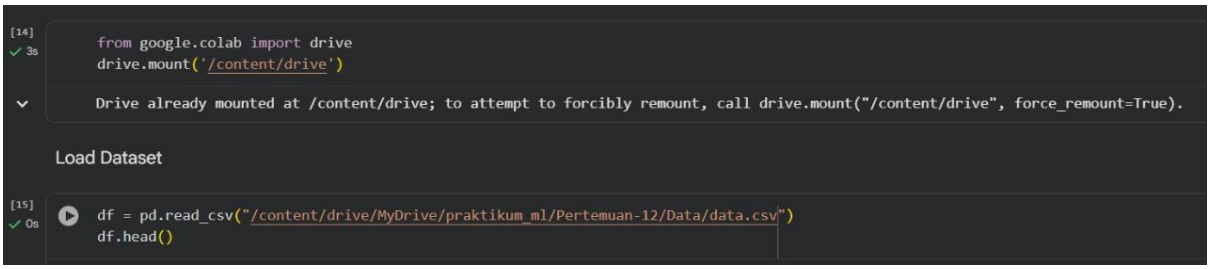
1.

Import Library

[13] ✓ 0s

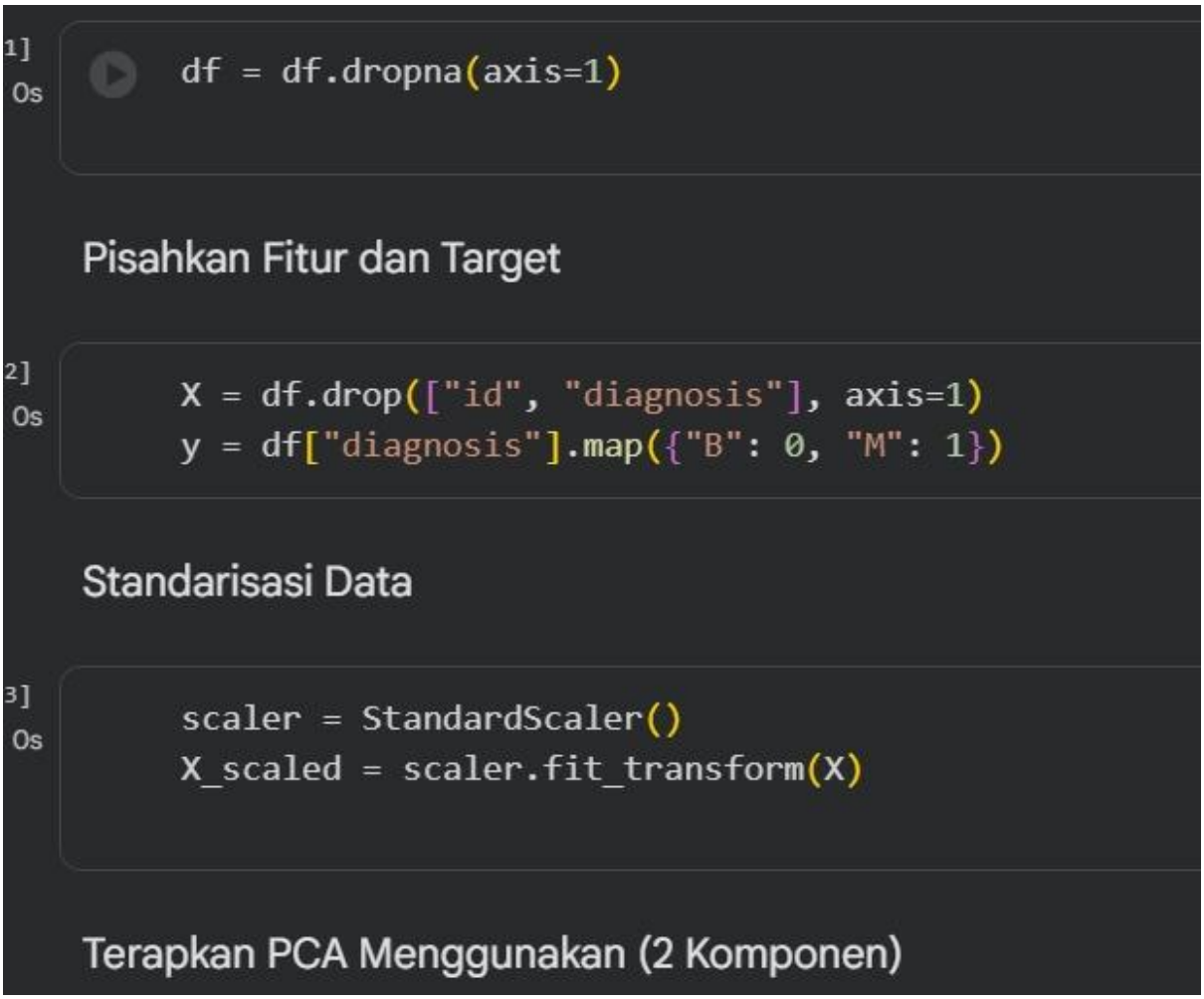
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

2.



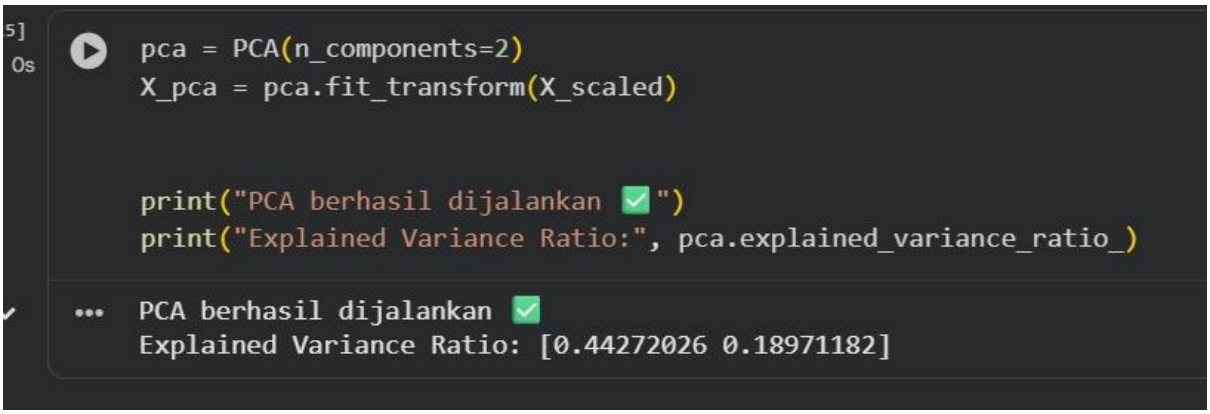
Kode tersebut digunakan untuk menghubungkan Google Drive ke Google Colab dan memuat dataset**.** Baris `from google.colab import drive` dan `drive.mount('/content/drive')` berfungsi untuk mengakses Google Drive di Colab, sedangkan pesan yang muncul menandakan Drive sudah terhubung. Selanjutnya, `pd.read_csv(...)` digunakan untuk membaca file CSV bernama `data.csv` dari folder Drive ke dalam variabel `df`, dan `df.head()` menampilkan 5 baris pertama data untuk memastikan dataset berhasil dimuat.

3.



Kode diatas melakukan pra-pemrosesan data untuk machine learning, yaitu menghapus kolom yang memiliki nilai kosong dengan `df.dropna(axis=1)`, memisahkan fitur (X) dengan menghapus kolom `id` dan `diagnosis`, serta menentukan **target (y)** dari kolom `diagnosis` yang diubah menjadi numerik (B = 0, M = 1). Selanjutnya, data fitur distandarisasi menggunakan `StandardScaler` agar semua fitur memiliki skala yang sama sebelum digunakan pada tahap analisis atau pemodelan lanjutan seperti PCA.

4.



Kode tersebut digunakan untuk menerapkan Principal Component Analysis (PCA) pada data yang telah distandarisasi dengan tujuan mereduksi dimensi data. PCA dibentuk dengan dua komponen utama sehingga data yang awalnya berdimensi tinggi diubah menjadi dua dimensi baru. Proses ini berhasil dijalankan, yang ditandai dengan munculnya pesan keberhasilan. Nilai explained variance ratio menunjukkan bahwa komponen utama pertama mampu menjelaskan sekitar 44,27% variasi data, sedangkan komponen kedua menjelaskan sekitar 18,97%, sehingga secara keseluruhan dua komponen utama tersebut mampu merepresentasikan sekitar 63% informasi dari dataset asli.

5.

```
plt.figure(figsize=(8,6))
plt.scatter(
    X_pca[:, 0],
    X_pca[:, 1],
    c=y,
    alpha=0.7
)
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.title("PCA pada Breast Cancer Dataset")
plt.show()
```

Kode tersebut digunakan untuk memvisualisasikan hasil PCA dalam bentuk grafik dua dimensi. Grafik dibuat dengan ukuran 8×6 dan menampilkan titik-titik data berdasarkan nilai Principal Component 1 pada sumbu X dan Principal Component 2 pada sumbu Y. Setiap titik merepresentasikan satu data pasien, sedangkan warna titik ditentukan oleh label kelas (benign atau malignant), dengan tingkat transparansi diatur agar titik yang saling tumpang tindih tetap terlihat. Pemberian label sumbu dan judul grafik bertujuan untuk memperjelas interpretasi visualisasi, sehingga pola distribusi data dan pemisahan antar kelas setelah penerapan PCA dapat diamati dengan lebih mudah.

KESIMPULAN

Berdasarkan hasil implementasi dan analisis PCA pada Breast Cancer Dataset, dapat disimpulkan bahwa metode PCA efektif dalam mereduksi dimensi dataset medis berdimensi tinggi dari 30 fitur menjadi 2 komponen utama. Proses standarisasi data sebelum penerapan PCA sangat berpengaruh terhadap hasil reduksi dimensi. Visualisasi hasil PCA menunjukkan adanya pola pemisahan antara kelas benign dan malignant, sehingga PCA dapat membantu dalam eksplorasi dan pemahaman struktur data.

Selain itu, penggunaan hasil PCA sebagai fitur input pada model klasifikasi mampu menghasilkan performa yang cukup baik dengan kompleksitas komputasi yang lebih rendah. Dengan demikian, PCA dapat digunakan sebagai teknik preprocessing yang bermanfaat untuk meningkatkan efisiensi analisis dan pemodelan pada dataset medis berdimensi tinggi.