**PRAKTIKUM DATA WAREHOUSING DAN DATA MINING**

**MODUL 13**

**PRINCIPAL COMPONENT ANALYSIS**

****

**Disusun oleh:**

**Adinda Aulia Hapsari**

**L200220037**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS KOMUNIKASI DAN INFORMATIKA**

**UNIVERSITAS MUHAMMADIYAH SURAKARTA**

**TAHUN 2024**

|  |  |
| --- | --- |
| Setelah kegiatan selesai, lembar kerja ini dicetak (di-print) dan dikumpulkan ke asisten.  NIM : L200220037  Nama : Adinda Aulia Hapsari  Nama Asisten : Diva Halimah  Tanggal Praktikum : 27 Desember 2024 | (Diisi oleh Asisten)  Nilai Praktek :  Tanda Tangan : |

**KEGIATAN PRAKTIKUM**

**Contoh Kasus:**

Terdapat sebuah dataset tentang review dari sebuah platform travel TripAdvisor, yang berjumlah 980 reviews. Kemudian dataset tersebut terdapat 11 features sebagai berikut:

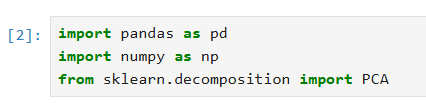
|  |  |
| --- | --- |
| Atribut 1: Unique user id | Merupakan atribut yang berisikan id user |
| Atribut 2: Average user feedback on art galleries | Merupakan atribut yang berisikan nilai rata-rata feedback dari art gallery |
| Atribut 3: Average user feedback on dance clubs | Merupakan atribut yang berisikan nilai rata-rata feedback dari dance club |
| Atribut 4: Average user feedback on juice bars | Merupakan atribut yang berisikan nilai rata-rata feedback dari juice bars |
| Atribut 5: Average user feedback on restaurants | Merupakan atribut yang berisikan nilai rata-rata feedback dari restoran |
| Atribut 6: Average user feedback on museums | Merupakan atribut yang berisikan nilai rata-rata feedback dari museum |
| Atribut 7: Average user feedback on resorts | Merupakan atribut yang berisikan nilai rata-rata feedback dari resorts |
| Atribut 8: Average user feedback on parks/picnic spots | Merupakan atribut yang berisikan nilai rata-rata feedback dari taman dan tempat wisata |
| Atribut 9: Average user feedback on beaches | Merupakan atribut yang berisikan nilai rata-rata feedback dari Pantai |
| Atribut 10: Average user feedback on theaters | Merupakan atribut yang berisikan nilai rata-rata feedback dari teater |
| Atribut 11: Average user feedback on religious institutions | Merupakan atribut yang berisikan nilai rata-rata feedback dari lembaga keagamaan |

**Hipotesis:**

Bagaimana mendapatkan dataset dengan dimensi yang lebih rendah dengan menggunakan algoritma PCA berdasarkan features yang telah diketahui pada dataset?

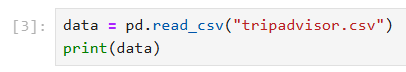
**13.4.1 Mengimport Library**

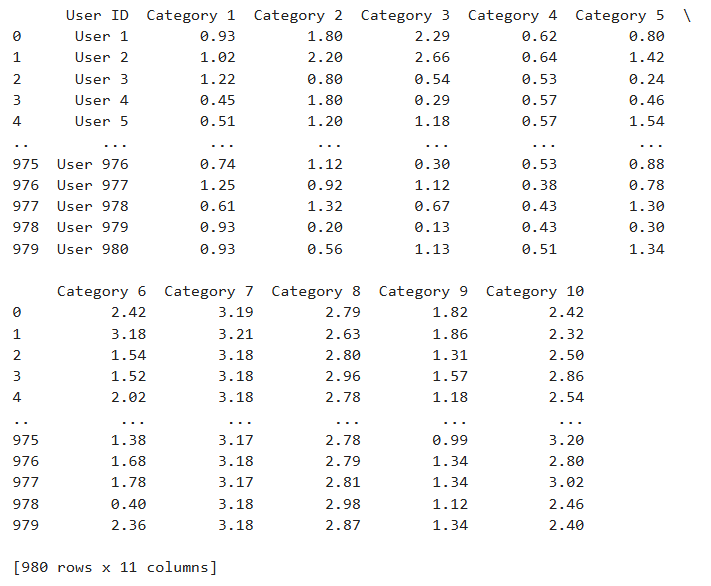
Meng-import library yang diperlukan, yaitu library pandas, numpy, dan PCA. Untuk meng-import library yang akan digunakan, kita meng import library pandas, numpy, dan PCA.



**13.4.2 Membaca Dataset**

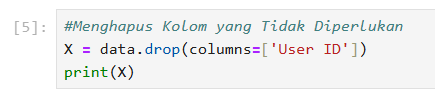
Membaca dataset dari sebuah directory. Disini, kita mengambil dataset tripadvisor.csv yang diambil dari platform gitea. Dataset tersebut akan tersimpan di variable data sebagai sebuah dataframe.

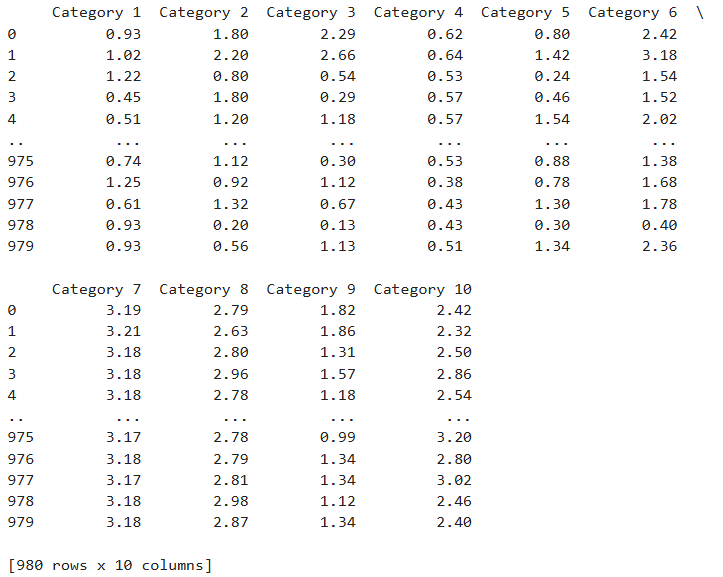




**13.4.3 Menghapus Kolom yang Tidak Diperlukan**

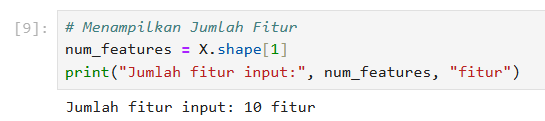
Disini kita akan mengeliminasi feature yang tidak diperlukan, dalam hal ini adalah User ID, karena User ID tidak akan berpengaruh pada hasil clustering. Pada kode di bawah ini, kita mengaplikasikan fungsi drop pada variable data yang menyimpan data frame. Fungsi drop disini bertujuan untuk menghapus kolom User ID. Setelah User ID dihapus, dataframe disimpan ke dalam variable X. Kemudian variable X dicetak.





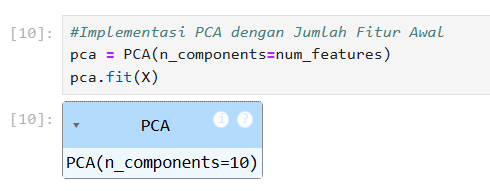
**13.4.4 Menampilkan Jumlah Fitur**

Kode di bawah ini akan menampilkan jumlah fitur yang ada pada dataset setalah User ID dieliminasi. Fungsi shape menampilkan bentuk atau ordo dari sebuah dataset. Untuk mengetahui jumlah baris pada dataset X, kita menggunakan X.shape[0] sedangkan apabila kita ingin mengetahui jumlah kolom pada dataset X, kita menggunakan X.shape[1] pada kode Python tersebut. Kemudian jumlah kolom kita simpan pada variable num\_features, yang akan kita cetak. Dalam kasus ini, kita memiliki 10 kolom/fitur.



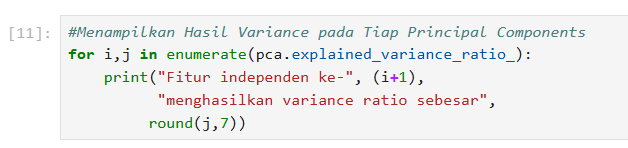
**13.4.5 Implementasi PCA dengan Jumlah Fitur Awal**

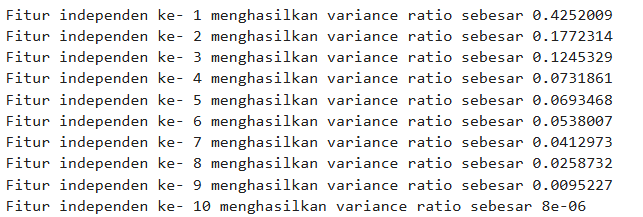
Selanjutnya, kita jalankan kode untuk PCA dengan cara memanggil fungsi PCA. Adapun fungsi PCA memiliki parameter wajib yaitu: n\_ components yang berarti jumlah fitur yang ada pada dataset. Disini, dikarenakan jumlah fitur pada dataset setelah User ID dieliminasi adalah sejumlah 10 fitur, maka kita menggunakan n\_components = 10 yang berarti kita akan membuat 10 principal components menggunakan teknik PCA.



**13.4.6 Menampilkan Hasil Variance pada Tiap Principal Components**

Kemudian, kita menampilkan hasil variance pada tiap principal components. Pada kode di bawah ini kita menerapkan fungsi enumerate untuk mendapatkan indeks dan data dari setiap elemen di dalam sebuah list. Elemen pada list pca.explained\_variance\_ratio\_ indeksnya akan tersimpan dalam variable i sedangkan data akan tersimpan dalam variable j. Setelah itu, pada setiap iterasi, program akan mencetak nilai variance untuk setiap principal components. Hasilnya, semakin rendah urutan principal components, nilai variance semakin tinggi.

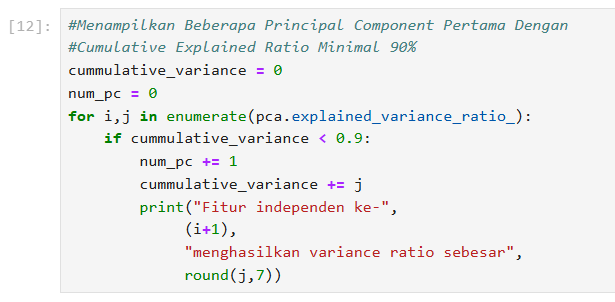


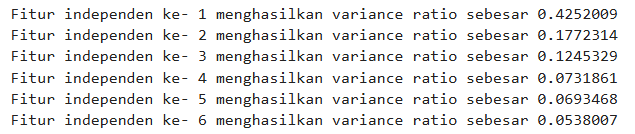


**13.4.7 Menampilkan Beberapa Principal Component Pertama Dengan Cumulative Explained Ratio Minimal 90%**

Pada kode di bawah ini kita menambahkan kondisi dimana kita akan mengambil nilai variance kumulatif dari principal components jika sudah mencapai angka minimal 0.9. Disini, kita menambahkan variable cumulative\_variance untuk menyimpan nilai variance kumulatif untuk principal components, juga variable num\_pc untuk menyimpan jumlah principal components saat iterasi dijalankan.

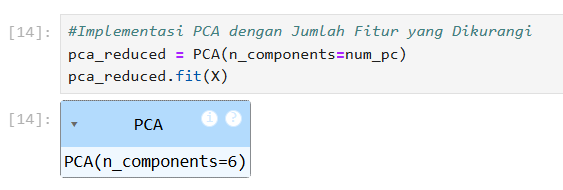
Pada setiap iterasi, nilai num\_pc bertambah satu dan nilai cumulative\_variance bertambah sesuai dengan nilai variance pada setiap principal components; program akan mencetak nilai variance untuk setiap principal components untuk cumulative variance maksimal 0.9.





**13.4.8 Implementasi PCA dengan Jumlah Fitur yang Dikurangi**

Selanjutnya, kita jalankan kode untuk PCA untuk mengurangi dimensi pada dataset. Disini kita akan menggunakan fungsi PCA dengan n\_components sejumlah 6 dikarenakan jumlah yang dapat mewakili 90% dari dataset adalah sebanyak 6 fitur, maka kita menggunakan n\_ components = 6 yang berarti kita akan membuat 6 principal components menggunakan teknik PCA.



**TUGAS**

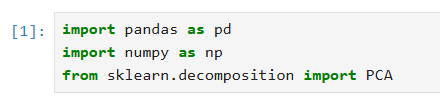
Terdapat dataset pada GPS Trajectories yang dapat diunduh pada halaman GPS+Trajectories berikut: http://archive.ics.uci.edu/ml/datasets/

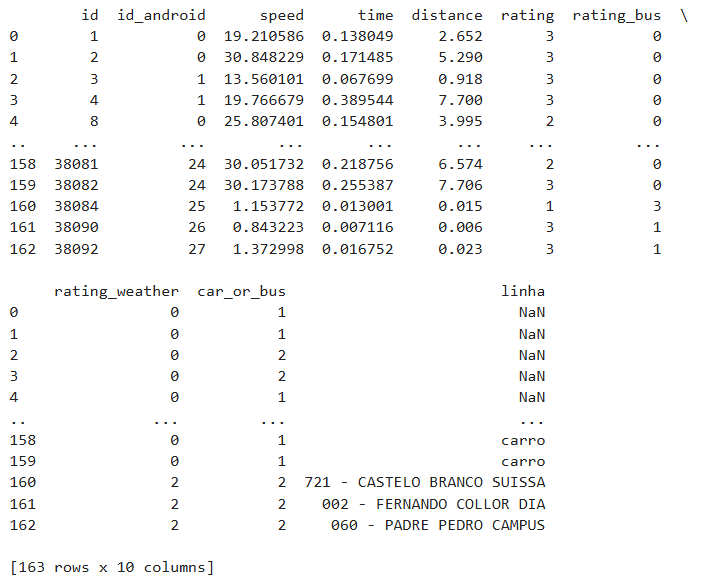
Setelah itu, bukalah dataset go\_track\_tracks.csv.

Pada dataset tersebut terdapat 163 data dengan atribut sebagai berikut:

1. id: id dari objek
2. id\_android: perangkat yang digunakan untuk membaca objek
3. speed: kecepatan rata-rata (km/h)
4. time: waktu tempuh perjalanan (h)
5. distance: jarak total (km)
6. rating: rating lalu lintas perjalanan. (3- baik, 2- normal, 1-buruk).
7. rating\_bus: rating bus (1 – Penumpang bus sedikit, 2 – Penumpang Bus cukup banyak, 3- Penumpang Bus banyak.
8. rating\_weather: rating cuaca (1- hujan, 2- cerah,).
9. mobil\_atau\_bus: (1 - mobil, 2 bus) 10.
10. linha: informasi tentang bus yang melakukan jalur tersebut

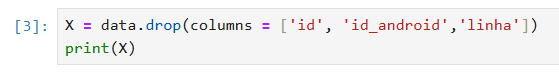
Kemudian kerjakanlah soal-soal berikut ini:

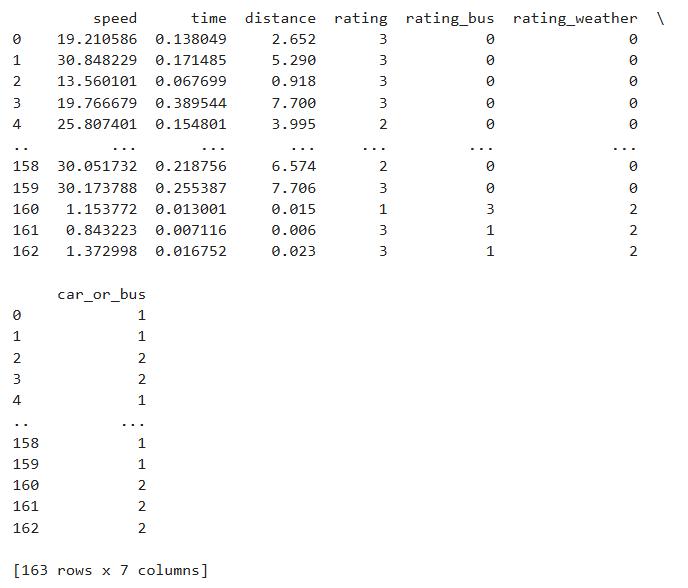


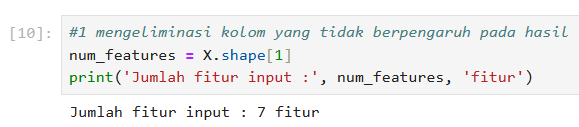


1. Tentukan berapa jumlah fitur input yang digunakan untuk PCA. Bagaimana cara anda mendapatkan nilai tersebut?

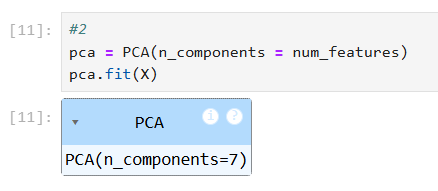
Menghapus fitur id, id android dan linha karena tidak akan berpengaruh pada hasil clustering. Sisa fitur adalah 7.



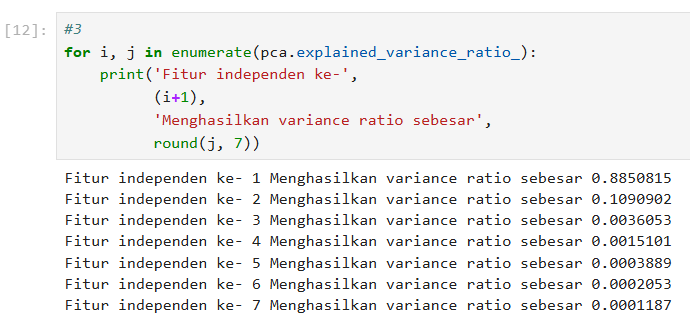


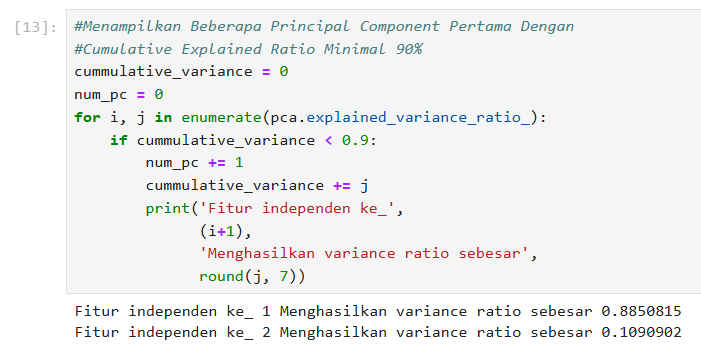


1. Tuliskan algoritma PCA dengan n\_components sebesar jumlah fitur input.



1. Setelah itu, tampilkan nilai variance ratio untuk setiap principal components. Kemudian, tentukan ada berapa fitur independen yang dapat memenuhi 90% cumulative variance ratio.





1. Cetaklah Data pada beberapa principal components pertama dengan Cumulative Explained Ratio Minimal 90%.

