Nama : Imron Bagas Sajiwo

Npm : 5230311037

Prodi : Sistem Informasi A

1. Apa perbedaan data cleaning & data reduction berikan pejelasan dan tuliskan contohnya.

Jawab:

A. Data Cleaning (Pembersihan Data)

Data cleaning adalah proses mengidentifikasi dan memperbaiki kesalahan atau inkonsistensi dalam data sehingga dapat meningkatkan kualitas data untuk analisis lebih lanjut. Kesalahan yang sering diperbaiki mencakup data yang hilang, data duplikat, data tidak relevan, atau data yang tidak sesuai format.

Contoh:

ID Pelanggan	Nama	Usia	Email	Kota
101	Andi	25	andi@email.com	Jakarta
102	Budi	30	NULL	Bandung
103	Siti	-5	siti@email.com	Surabaya
104	Budi	30	budi@email.com	Bandung

- Data "NULL" pada email Budi diperbaiki dengan mengisi data yang valid.
- Usia -5 adalah nilai tidak valid dan perlu diperbaiki atau dihapus.
- Data duplikat untuk Budi perlu dihapus atau dikoreksi.

B. Data Reduction (Reduksi Data)

Data reduction adalah proses menyederhanakan atau mengurangi ukuran dataset tanpa kehilangan informasi yang signifikan. Ini dilakukan untuk meningkatkan efisiensi penyimpanan dan kecepatan analisis data.

Contoh:

Dataset awal memiliki 10 fitur (kolom), tetapi setelah analisis, hanya 5 fitur yang paling relevan dipilih.

Sebelum reduksi:

ID	Nama	Usia	Email	Kota	Penghasilan	Status	Pendidikan	Hobi	Tanggal Daftar
1	Andi	25	andi@email.com	Jakarta	10 Juta	Menikah	S1	Sepak bola	6/12/2021

Setelah reduksi (hanya fitur penting yang dipilih):

ID	Nama	Usia	Kota	Penghasilan
1	Andi	25	Jakarta	10 Juta

2. Carilah data dari web yang sesuai dengan program studi saudara, jelaskan isinya, dan apa kontribusi data tersebut terhadap program studi saudara, lakukan pengolahan awal agar data tersebut siap diolah, apa saja yang saudara lakukan. berikan penjelasan caranya.

Jawab:

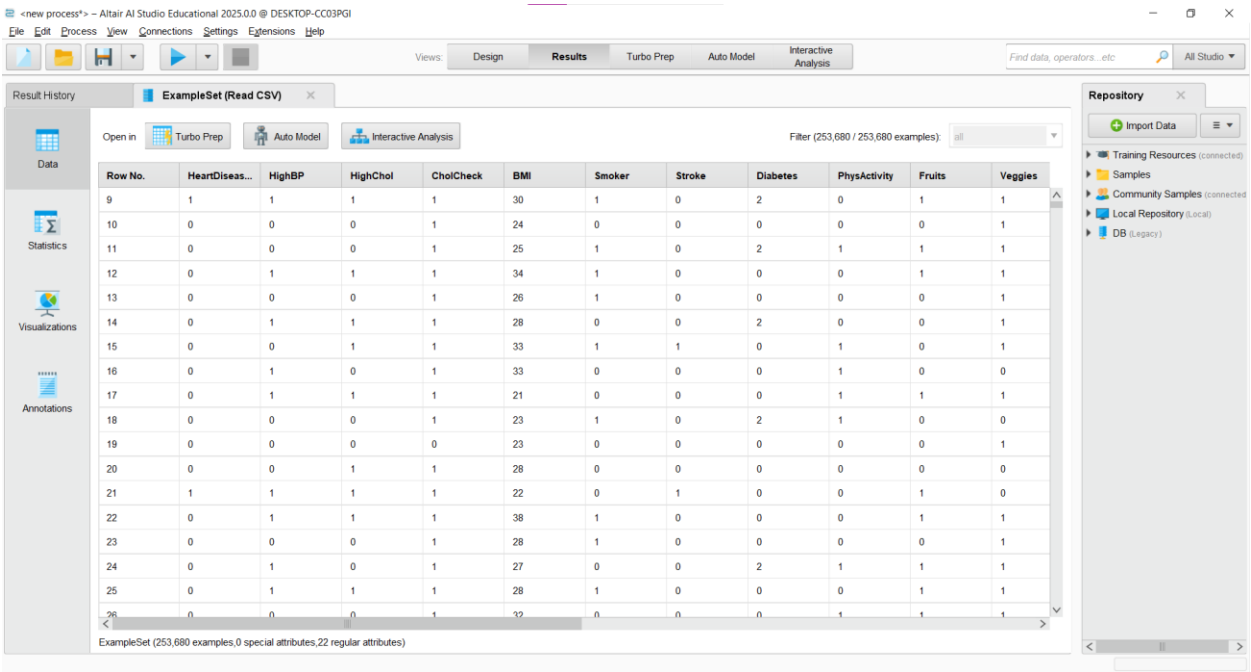
A. Isi Dataset

Dataset ini berisi **253.680 entri** dengan **22 kolom** yang mencatat indikator kesehatan terkait penyakit jantung dari survei BRFSS (Behavioral Risk Factor Surveillance System) tahun 2015.

Beberapa Kolom Utama dalam Dataset:

- HeartDiseaseorAttack – Indikator apakah seseorang pernah mengalami serangan jantung atau penyakit jantung.
- HighBP (Tekanan Darah Tinggi) – Menunjukkan apakah seseorang memiliki tekanan darah tinggi.
- HighChol (Kolesterol Tinggi) – Menunjukkan apakah seseorang memiliki kadar kolesterol tinggi.
- BMI (Indeks Massa Tubuh) – Mengukur apakah seseorang memiliki berat badan ideal, kurang, atau berlebih.
- Smoker (Perokok) – Indikator apakah seseorang pernah merokok minimal 100 batang dalam hidupnya.
- Stroke – Riwayat apakah seseorang pernah mengalami stroke.
- Diabetes – Indikator apakah seseorang memiliki diabetes.
- PhysActivity (Aktivitas Fisik) – Menunjukkan apakah seseorang melakukan aktivitas fisik secara rutin.
- Fruits & Veggies (Konsumsi Buah dan Sayur) – Menunjukkan kebiasaan makan sehat seseorang.

- HvyAlcoholConsump (Konsumsi Alkohol Berlebihan**)** – Menunjukkan apakah seseorang mengonsumsi alkohol dalam jumlah berlebihan.
- AnyHealthcare (Akses Layanan Kesehatan) – Menunjukkan apakah seseorang memiliki akses ke layanan kesehatan.
- NoDocbcCost (Kesulitan Bertemu Dokter karena Biaya) – Indikator seseorang tidak bisa menemui dokter karena kendala biaya.
- GenHlth, MentHlth, PhysHlth – Penilaian terhadap kondisi kesehatan umum, kesehatan mental, dan kesehatan fisik seseorang.
- Demografi (Sex, Age, Education, Income) – Informasi tentang jenis kelamin, usia, tingkat pendidikan, dan pendapatan seseorang.



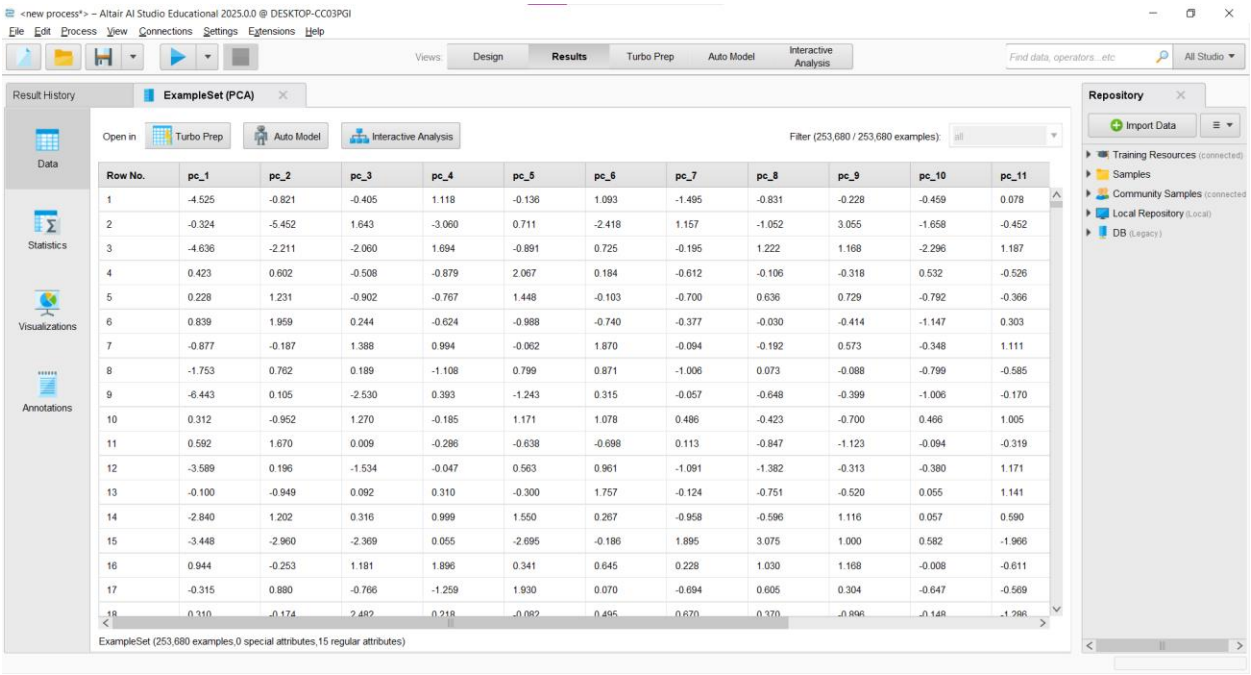
Penjelasan:

penjelasan diolah datanya saya pertama – tama memukan Rapid miner kemudian saya memberikan operator read CSV kemudian saya import data tersebut kemudian saya menyambungkan datanya lalu saya running untuk pengolahan awal.

3. Setelah pengolahan awal saudara lakukan, terapkan Operator PCA pada data saudara gunakan beberapa pilihan jumlah atributnya. Jelaskan langkah-langkahya dan tuliskan contoh hasilnya.

Jawab:

Contoh Hasil sesudah di PCA



Penjelasan:

Pertama tama Saya Import Data dari Read CSV kemudian saya memberikan Select Atributies Untuk Memilih Data yang akan dianalisis tetapi saya lupa dalam memilih data nya jadinya saya melakukan all Data jadinya data nya dianalisi semua, kemudian saya memberikan noralmalize untuk menormalkan data kemudian saya memberikan PCA tersebut agar data dapat dianalisis, setelah itu saya merunning data tersebut.