Nama : Adifa Syahira

Nim : 1103202067

Kelas : TK 44G4

TUGAS UNDERSTANDING PRINCIPAL ANALYSIS COMPONENT (PCA)

Grafik yang Anda tunjukkan dihasilkan dari data single cell RNA-seq. Setiap sel memiliki sekitar 10.000 gen yang di-transkripsi. Setiap titik dalam grafik tersebut merepresentasikan satu sel dan profil transkripsinya. Pada grafik ini, kita dapat melihat bahwa sel darah membentuk satu kelompok yang berbeda dari sel pluripoten, yang juga berbeda dari sel saraf dan sel dermal atau epidermal. Pertanyaannya adalah, bagaimana mungkin 10.000 gen yang di-transkripsi dapat disusutkan menjadi satu titik pada grafik?

Di sini, jawabannya adalah melalui PCA (Principal Component Analysis), yang merupakan metode untuk mereduksi banyak data menjadi sesuatu yang menggambarkan inti dari data asli.

Untuk memahami konsep dimensi dalam analisis data:

- **Dimensi 1 (1-Dimensi)**: Bayangkan ini seperti garis bilangan. Dalam konteks data RNA-seq, ini bisa mewakili distribusi yang seragam atau tidak seragam dari ekspresi gen pada satu sel.
- **Dimensi 2 (2-Dimensi)**: Ini seperti grafik biasa dengan dua sumbu. Ini memungkinkan kita untuk melihat apakah ekspresi gen pada dua sel berkorelasi atau tidak. Jika ekspresi gen berkorelasi, artinya gen yang di-transkripsi tinggi dalam sel satu juga tinggi dalam sel kedua, dan sebaliknya. Jika ekspresi tidak berkorelasi, itu tidak memberi kita informasi tentang tingkat transkripsi pada sel kedua.
- **Dimensi 3 (3-Dimensi)**: Ini seperti grafik yang lebih kompleks dengan kedalaman, memiliki tiga sumbu terpisah. Ini berguna ketika kita memiliki data dari tiga sel.

PCA digunakan untuk menyederhanakan dataset dengan banyak dimensi menjadi 2 atau 3 dimensi agar lebih mudah divisualisasikan. Komponen utama pertama (PC1) menggambarkan arah variasi terbesar dalam ekspresi gen, sementara PC2 menggambarkan variasi kedua terbesar.

- Jika kita punya data dari 2 sel, PC1 akan menggambarkan arah variasi terbesar, dan PC2 menggambarkan variasi kedua terbesar.
- Dengan data dari 3 sel, PC1, PC2, dan PC3 menggambarkan variasi pertama, kedua, dan ketiga secara berturut-turut.
- Dengan data dari 4 sel, kita akan memiliki PC1, PC2, PC3, dan PC4, masing-masing menggambarkan variasi keempat terbesar.

Untuk menampilkan sel pada grafik, panjang dan arah PC1 sebagian besar ditentukan oleh gen-gen tertentu. Gen-gen yang memiliki pengaruh kecil pada PC1 mendapatkan nilai yang mendekati nol, sementara gen-gen yang memiliki pengaruh besar mendapatkan nilai yang jauh dari nol.

Untuk mengidentifikasi gen-gen kunci yang mempengaruhi posisi sel dermal di sebelah kiri dan sel neural di sebelah kanan, kita dapat melihat nilai pengaruh pada PC1. Dan jika kita ingin menemukan gen-gen yang membantu membedakan sel darah dari sel neural dan dermal, kita dapat melihat nilai pengaruh pada PC2. Selanjutnya , Langkah-langkah umum dalam PCA adalah sebagai berikut:

1. Standarisasi Data: Data harus dinormalisasi atau distandarisasi terlebih dahulu karena seringkali memiliki skala yang berbeda-beda. Hal ini dilakukan agar semua variabel memiliki dampak yang sama dalam analisis.

2.Menghitung Matriks Kovarians: PCA menghitung matriks kovarians dari data yang sudah distandarisasi. Matriks kovarians mengukur hubungan antara variabel dalam data.

3. Menghitung Vektor Eigen dan Nilai Eigen: Selanjutnya, PCA menghitung vektor eigen dan nilai eigen dari matriks kovarians. Nilai eigen mengukur jumlah variasi yang dapat dijelaskan oleh setiap komponen.

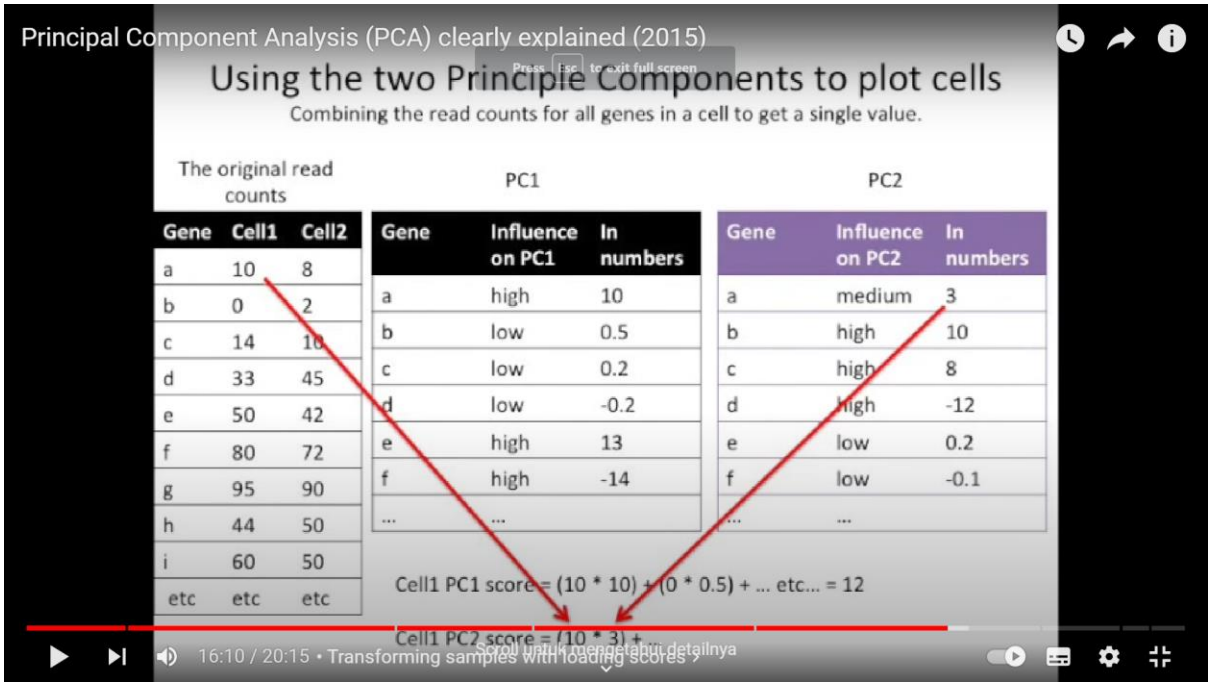
4. Pemilihan Komponen Utama: Komponen utama adalah vektor eigen yang sesuai dengan nilai eigen terbesar. Biasanya, komponen ini menjelaskan sebagian besar variasi dalam data.

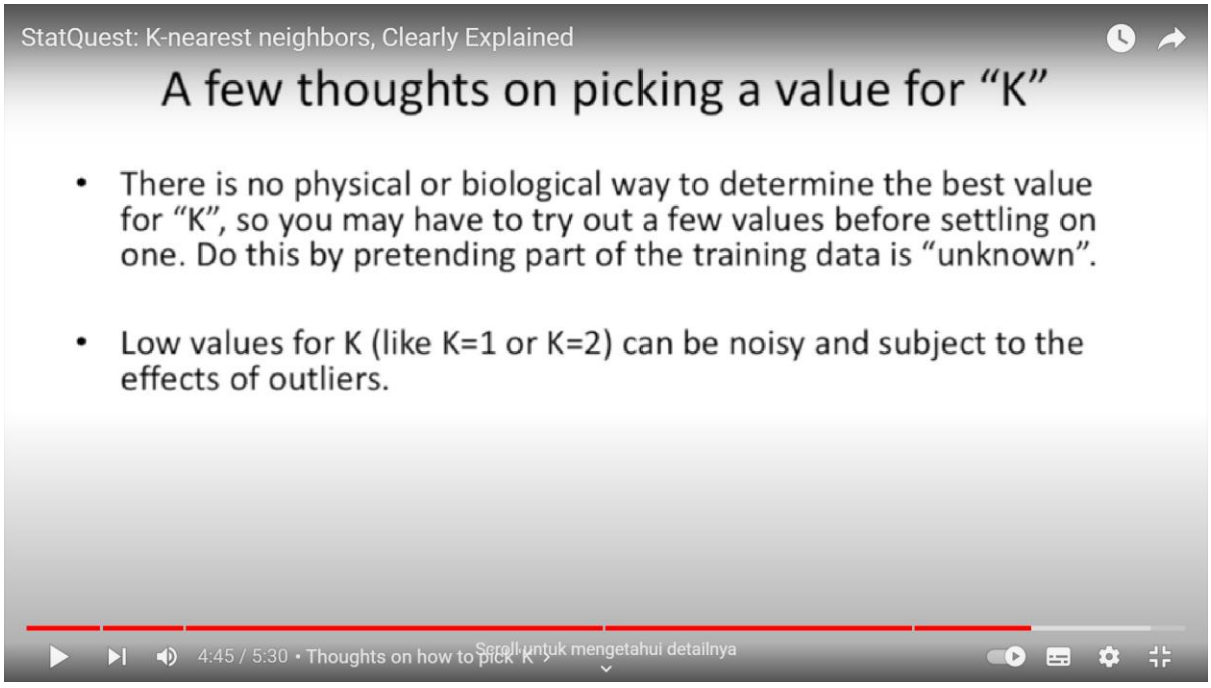
5. Transformasi Data: Data asli diubah (ditransformasikan) ke dalam sistem koordinat baru yang terdiri dari komponen-komponen utama. Ini menghasilkan representasi data dalam dimensi yang lebih rendah.

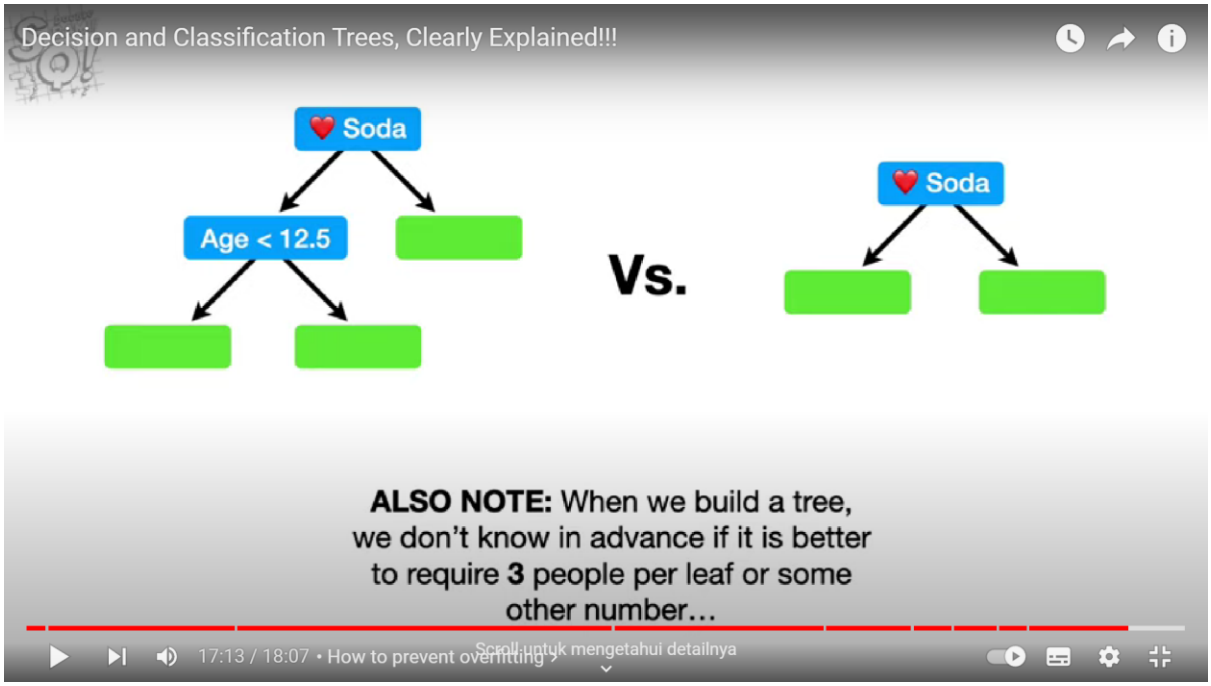
Analisis Hasil: Hasil PCA digunakan untuk memahami pola dalam data. Ini bisa berarti melihat bagaimana variabel asli berkontribusi pada komponen utama atau mengidentifikasi pola hubungan antara variabel. Demikianlah, PCA adalah alat penting dalam analisis data yang membantu kita memahami struktur data dan mengurangi kompleksitasnya.

K-nearest neighbors (KNN) adalah metode pembelajaran mesin yang digunakan untuk mengklasifikasikan data berdasarkan kemiripan dengan data lain. KNN mencari data yang paling mirip dengan data baru dan mengkategorikannya ke dalam kategori yang sama dengan data yang paling mirip. Untuk melakukan hal ini, KNN memerlukan data pelatihan yang mencakup berbagai kategori dan menghitung jarak antara data baru dengan data pelatihan. Jarak dapat dihitung dengan berbagai metode seperti jarak Euclidean, jarak Manhattan, atau jarak Chebyshev. Nilai K adalah jumlah data yang digunakan untuk menentukan kategori data baru. Nilai K yang lebih besar menghasilkan klasifikasi yang lebih halus, sementara nilai K yang lebih kecil menghasilkan klasifikasi yang lebih kasar.

Pohon Keputusan adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Pohon keputusan bekerja dengan membagi data menjadi subset yang lebih kecil dan lebih kecil hingga setiap subset hanya berisi data dari kategori yang sama. Pohon keputusan digunakan untuk klasifikasi, sementara pohon regresi digunakan untuk tugas regresi. Pohon keputusan menghasilkan kelas sebagai keluaran, sedangkan pohon regresi menghasilkan nilai numerik. Dalam terminologi pohon keputusan, terdapat simpul akar, simpul internal, simpul daun, fitur, ambang batas, dan pengotoran gini yang digunakan untuk mengukur ketidakmurnian suatu simpul.


Principal Component Analysis (PCA) clearly explained (2015)
Using the two Principle Components to plot cells
Combining the read counts for all genes in a cell to get a single value.
The original read counts
PC1
PC2
Gene Cell1 Cell2
a 10 8
b 0 2
c 14 10
d 33 45
e 50 42
f 80 72
g 95 90
h 44 50
i 60 50
etc etc etc
Gene Influence on PC1 In numbers
a high 10
b low 0.5
c low 0.2
d low -0.2
e high 13
f high -14
... ...
Gene Influence on PC2 In numbers
a medium 3
b high 10
c high 8
d high -12
e low 0.2
f low -0.1
... ...
Cell1 PC1 score = (10 * 10) + (0 * 0.5) + ... etc... = 12
Cell1 PC2 score = (10 * 3) +
16:10 / 20:15 • Transforming samples with loading scores
Scroll untuk mengetahui detailnya


StatQuest: K-nearest neighbors, Clearly Explained
A few thoughts on picking a value for "K"
There is no physical or biological way to determine the best value for "K", so you may have to try out a few values before settling on one. Do this by pretending part of the training data is "unknown".
Low values for K (like K=1 or K=2) can be noisy and subject to the effects of outliers.
4:45 / 5:30 • Thoughts on how to pick K
Scroll untuk mengetahui detailnya


Decision and Classification Trees, Clearly Explained!!!
Soda
Age < 12.5
Vs.
Soda
ALSO NOTE: When we build a tree, we don't know in advance if it is better to require 3 people per leaf or some other number...
17:13 / 18:07 • How to prevent overfitting
Scroll untuk mengetahui detailnya