

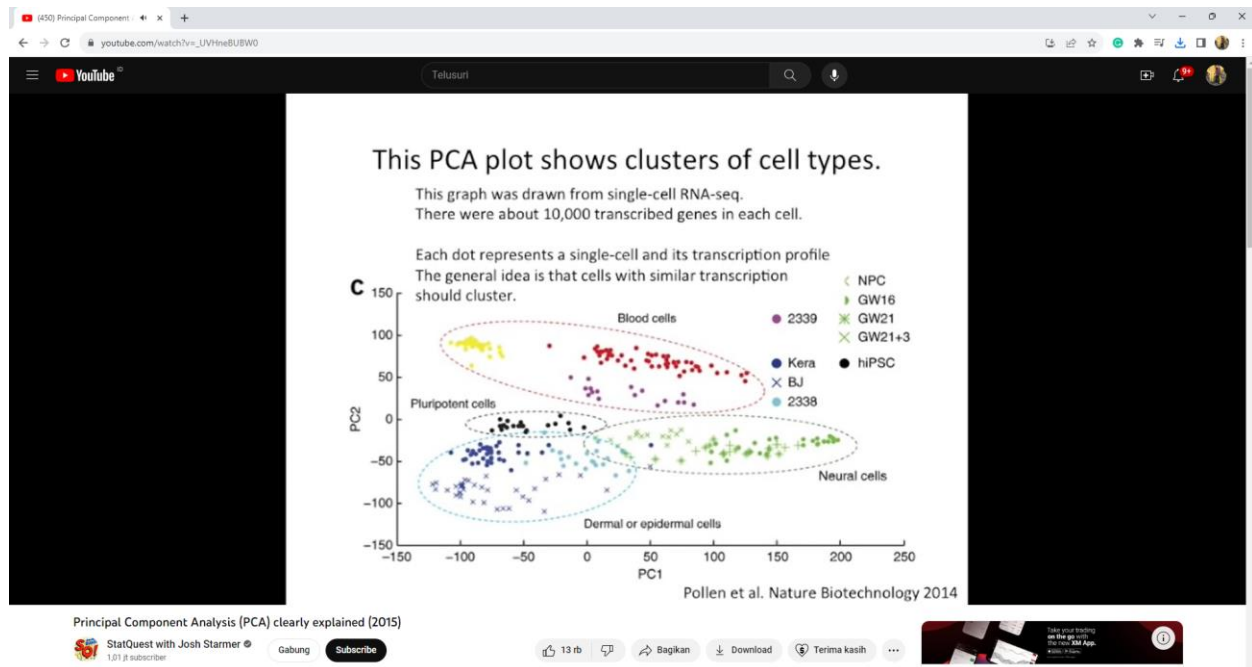
Nama : Ivan Fernanda Prayoga

NIM : 1103204035

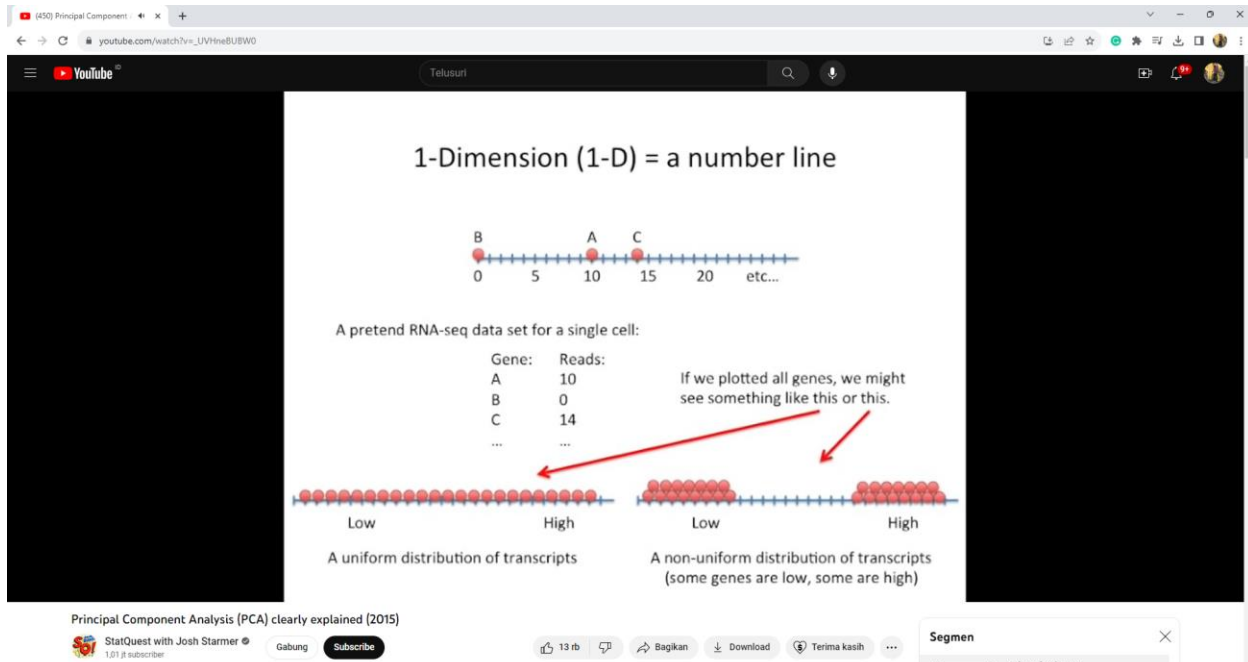
Understanding 3 link StatQuest

Principal Component Analysis (PCA) clearly explained (2015)

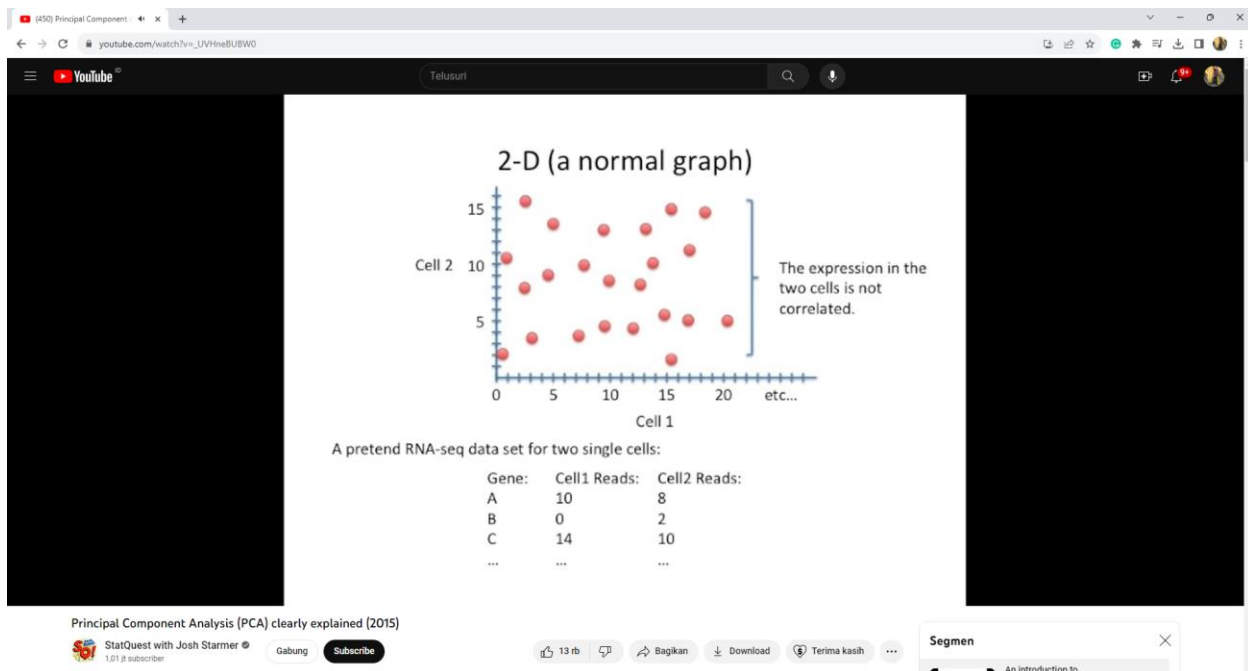
Principal Component Analysis (PCA) adalah teknik statistik yang digunakan dalam analisis data untuk mengidentifikasi pola dalam data dan mengekspresikan data tersebut dengan cara yang meminimalkan kerumitan. Tujuan utamanya adalah untuk mengurangi dimensi dari set data berdimensi besar menjadi set data berdimensi lebih rendah, sambil mempertahankan sebanyak mungkin informasi yang relevan.



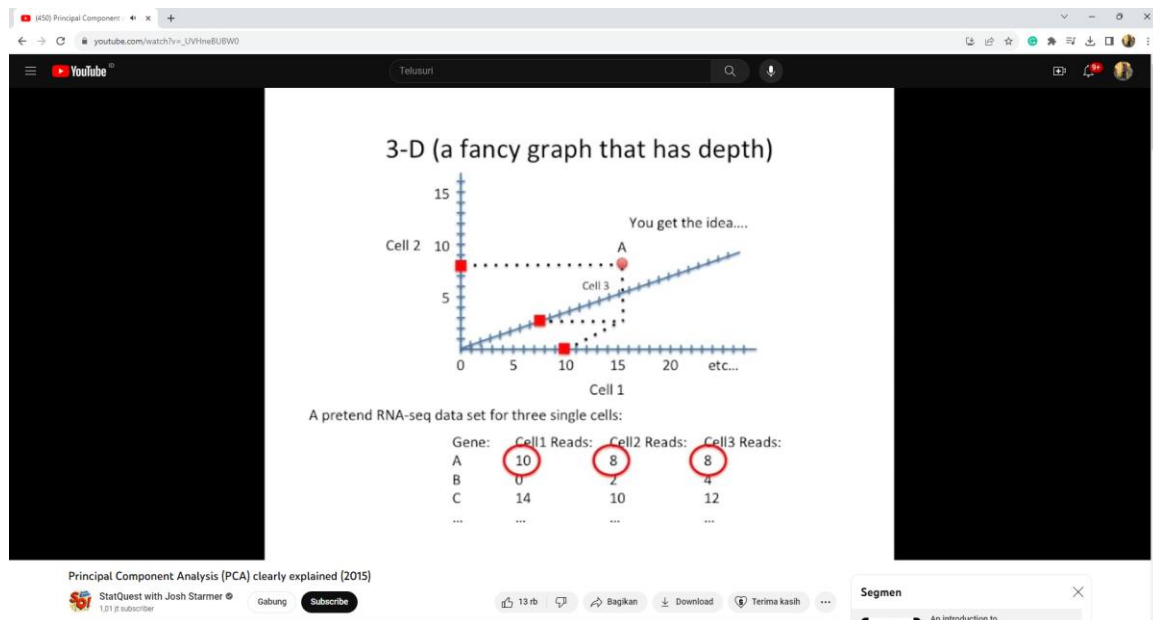
Pada gambar diatas, diagram berasal dari RNA sel tunggal. Sekitar 10.000 gen ditranskripsi dalam setiap sel. Setiap poin mewakili satu sel dan profil transkripsinya. Konsep dasarnya adalah bahwa sel-sel dengan pola transkripsi yang mirip seharusnya mengelompok bersama. Teknik PCA digunakan untuk meredam jumlah data menjadi sesuatu yang mencakup inti dari data asli.



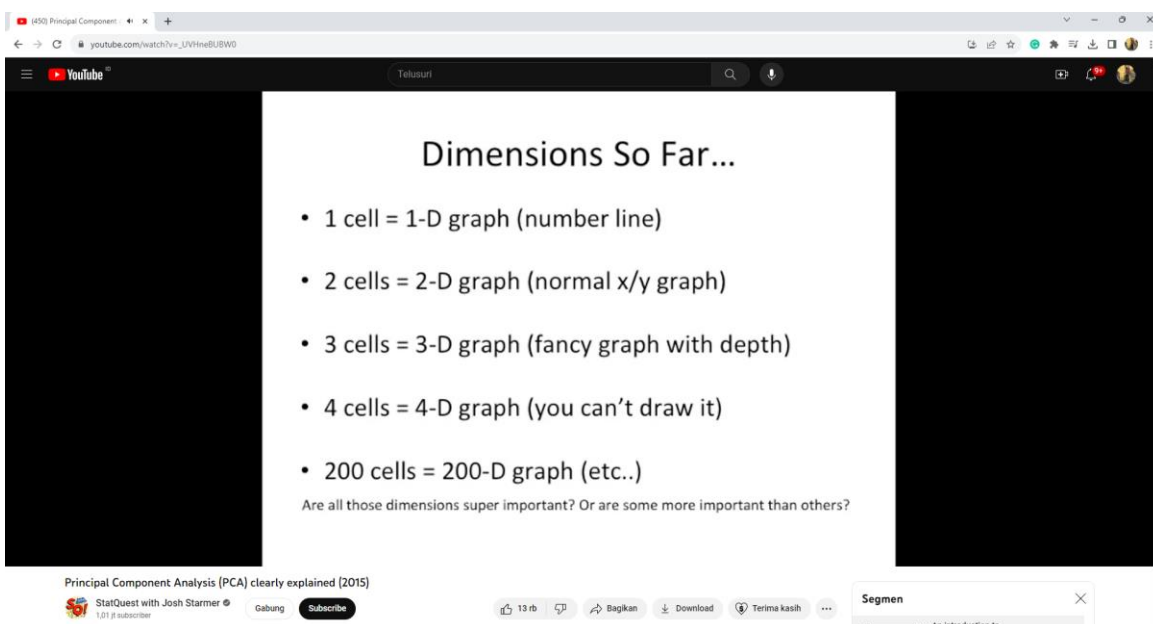
Dari gambar diatas menunjukan bagaimana memplot data transkrip gen dalam satu dimensi menggunakan garis. Misalnya, jika kita memiliki data sekuen RNA dari satu sel dengan gen A, B, dan C yang memiliki hitungan bacaan 10, 0, dan 14 masing-masing, kita dapat menempatkan titik pada angka tersebut di garis bilangan. Dengan cara ini, meskipun garis bilangan adalah grafik yang sangat sederhana, kita bisa mendapatkan informasi tentang distribusi transkripsi gen dalam sel tersebut.



Pada gambar diatas menjelaskan cara memplot data sekuen RNA dari dua sel berbeda dalam grafik dua dimensi, di mana gen dapat berkorelasi (gen yang ditranskripsi tinggi/rendah di satu sel cenderung sama di sel lainnya) atau tidak berkorelasi (tingkat transkripsi gen di satu sel tidak memprediksi tingkat transkripsi di sel lainnya).



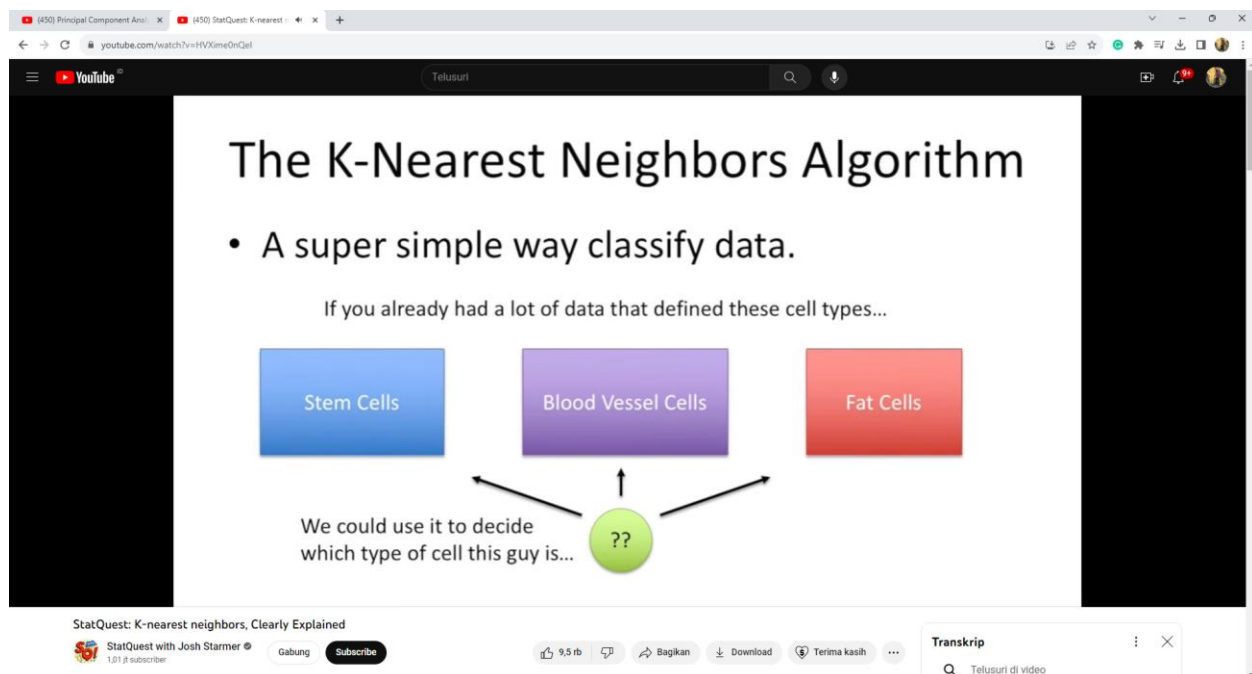
Pada gambar diatas menjelaskan tentang cara memplot data sekuen RNA dari tiga sel berbeda dalam grafik tiga dimensi. Dalam hal ini, kita memiliki sumbu ketiga yang menambah kedalaman ke grafik, memungkinkan kita untuk memvisualisasikan data dari tiga sel sekaligus. Misalnya, untuk Gen A, kita pergi ke 10 pada sumbu x (Sel 1), naik ke 8 pada sumbu y (Sel 2), dan mundur ke 8 pada sumbu z (Sel 3). Kemudian kita menggambar garis tegak lurus ke setiap sumbu untuk mengetahui di mana mereka bertemu dan menempatkan titik di sana.



Pada gambar diatas menggambarkan bahwa kompleksitas visualisasi data dalam grafik berkorelasi dengan jumlah sel yang datanya kita analisis. Untuk satu sel, sebuah garis bilangan (grafik satu dimensi) sudah memadai. Untuk dua sel, kita membutuhkan grafik XY (dua dimensi). Dan tiga sel memerlukan representasi dalam tiga dimensi. Namun, masalah akan muncul ketika kita memiliki data dari lebih dari tiga sel, seperti empat atau bahkan 200 sel individual - di mana maka diperlukan grafik empat-dimensi atau 200-dimensi.

StatQuest: K-nearest neighbors, Clearly Explained

StatQuest: K-Nearest Neighbors (K-NN) membahas tentang algoritma K-NN, sebuah metode sederhana untuk mengklasifikasikan data.



Pada gambar diatas kita memiliki tiga jenis sel yang berfungsi sebagai kategori klasifikasi: stem cells, blood vessel cells, dan fat cells.

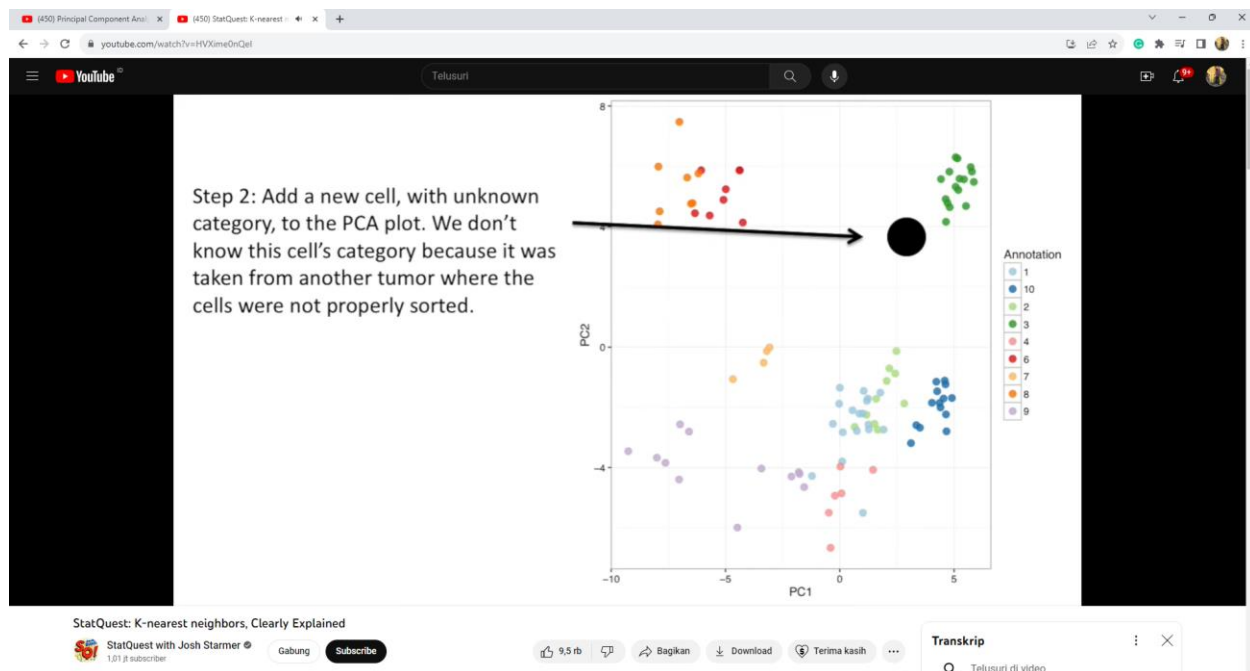
Stem cells merupakan satu dari beberapa kemungkinan kelas dalam masalah klasifikasi. Sampel atau data dapat ditempatkan dalam kelompok "stem cells" jika mereka menunjukkan karakteristik atau fitur tertentu yang sesuai dengan definisi ini.

Blood vessel cells mewakili kelas lainnya dalam skenario klasifikasi. Sampel atau data yang menunjukkan atribut-atribut spesifik akan dikategorikan sebagai "blood vessel".

Fat cells adalah jenis sel ketiga dan terakhir yang digunakan sebagai sebuah kelas dalam konteks ini. Data atau sampel dengan fitur-fitur tertentu akan dikelompokkan ke dalam "fat cells".



Pada step ke-1 kita mulai dengan dataset yang kategori-kategorinya sudah diketahui, kemudian mengelompokkan data tersebut (misalnya menggunakan metode Principal Component Analysis), dan menambahkan data baru yang belum dikategorikan ke dalam plot atau grafik.



Pada step ke-2 tambahkan sel baru yang kategorinya masih tidak diketahui, yang berasal dari jenis tumor lain dan belum terklasifikasi, ke dalam dataset kita.

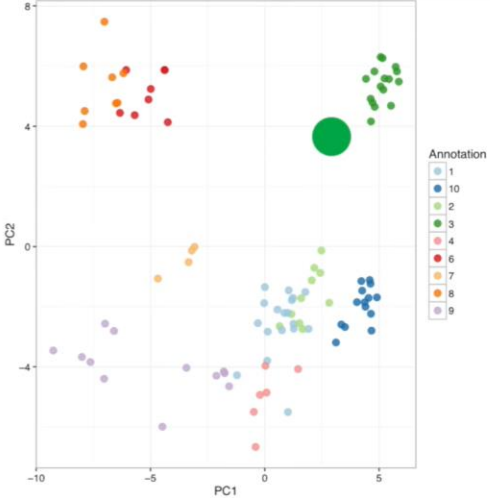
Step 3: We classify the new cell by looking at the nearest annotated cells. (i.e. the "nearest neighbors").

If the "K" in "K-nearest neighbors" is equal to 1, then we only use the nearest neighbor to define the category.

In this case, the category is **GREEN**.

If K=11, we would use the 11 nearest neighbors.

In this case, the category is still **GREEN**.



Annotation

- 1
- 10
- 2
- 3
- 4
- 6
- 7
- 8
- 9

StatQuest: K-nearest neighbors, Clearly Explained



StatQuest with Josh Starmer

1,011 subscribers

Gabung

Subscribe

9,5 rb



Bagikan



Download



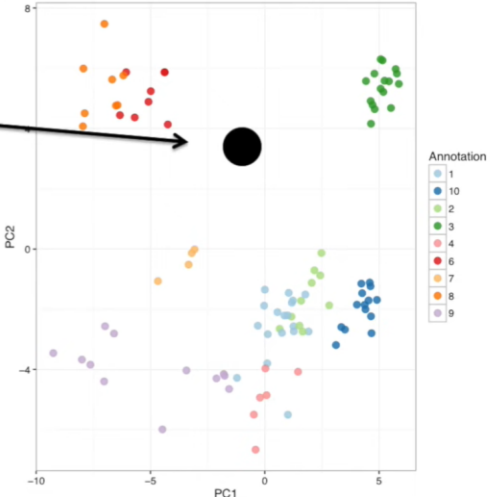
Terima kasih



Transkrip

Telusuri di video

Now the new cell is somewhere more interesting...



Annotation

- 1
- 10
- 2
- 3
- 4
- 6
- 7
- 8
- 9

StatQuest: K-nearest neighbors, Clearly Explained



StatQuest with Josh Starmer

1,011 subscribers

Gabung

Subscribe

9,5 rb



Bagikan



Download



Terima kasih



Transkrip

Telusuri di video

If $K=11$ and the new cell is between two (or more) categories, we simply pick the category that "gets the most votes".

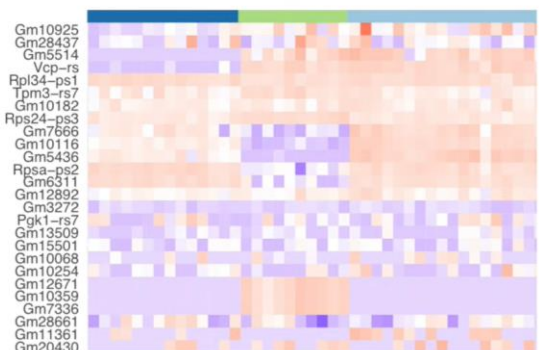
In this case....

7 nearest neighbors are **RED**.
 3 nearest neighbors are **ORANGE**.
 1 nearest neighbor is **GREEN**.

Since **RED** got the most votes, the final assignment is **RED**.



The same principle applies to heatmaps... This heatmap was drawn with the same data and clustered using hierarchical clustering.

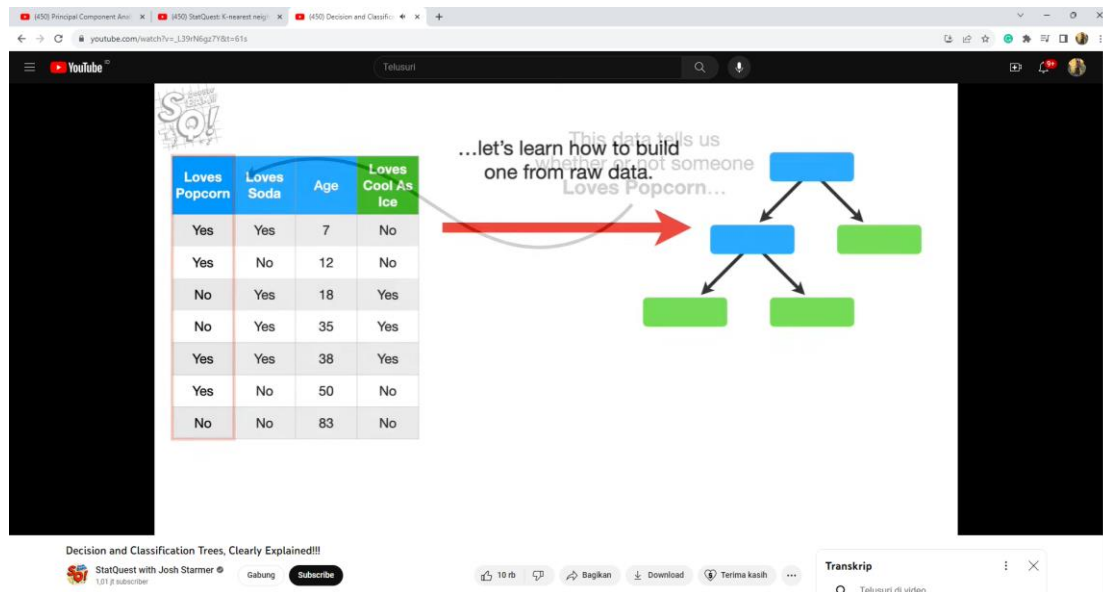


Pada step terakhir sel baru diklasifikasikan berdasarkan sel-sel terdekat yang sudah dianotasi atau neighbors terdekat. Jika parameter K adalah 1, maka hanya neighbors terdekat yang digunakan untuk menentukan kategori sel baru. Misalnya, jika neighbors terdekat adalah tipe sel hijau, maka sel baru juga akan dikategorikan sebagai hijau. Namun, jika K adalah 11, kita menggunakan 11 neighbors terdekat untuk menentukan kategori. Jika posisi sel baru berada di antara dua atau lebih kategori, kita memilih kategori dengan jumlah suara mayoritas dari neighborsnya; contohnya jika tujuh dari sebelas tetangga terdekat adalah merah dan sisanya oranye dan hijau,

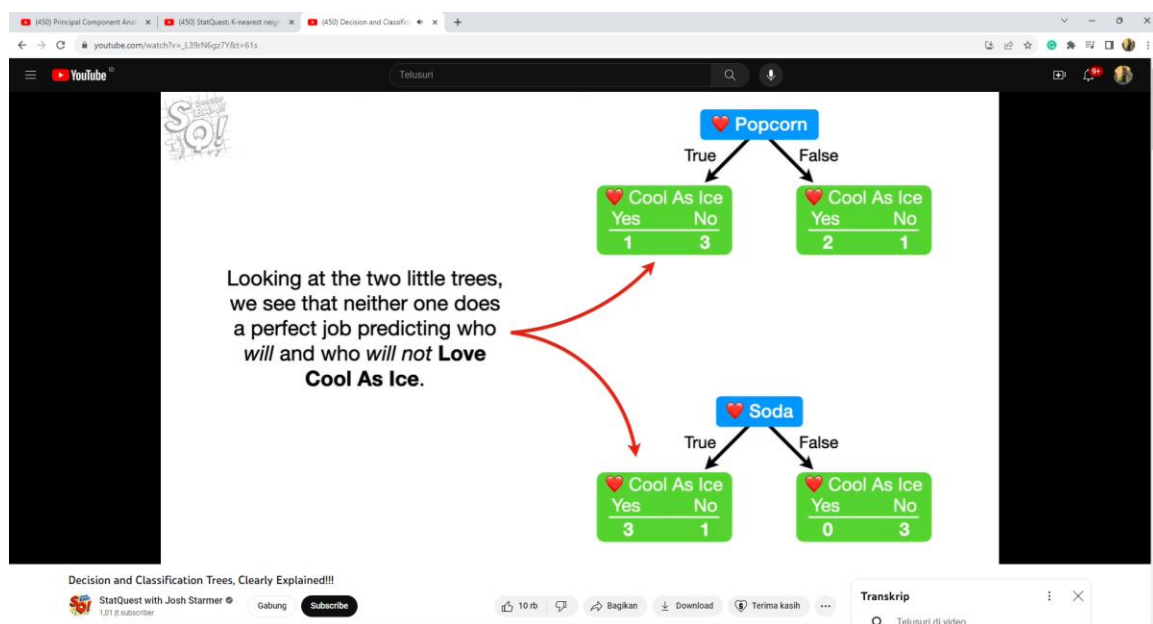
maka karena merah mendapatkan suara paling banyak, sel baru akan dikategorikan sebagai merah. Prinsip ini juga dapat diterapkan pada heatmap.

Decision and Classification Trees, Clearly Explained

Decision and Classification Trees adalah metode statistik yang digunakan untuk memprediksi hasil berdasarkan serangkaian pertanyaan biner.



Pada gambar diatas menjelaskan cara membangun Decision and Classification Trees dari data mentah untuk memprediksi apakah seseorang menyukai film "Cool as Ice" berdasarkan preferensi terhadap popcorn, soda, dan age.



Then we multiply that weight by its associated **Gini Impurity, 0.375.**

Gini Impurity = 0.375

Total Gini Impurity = weighted average of Gini Impurities for the Leaves

$$= \left(\frac{4}{4 + 3} \right) 0.375$$

Decision and Classification Trees, Clearly Explained!!!

StatQuest with Josh Starmer 1,018 subscribers

10 likes, 1 share, 1 download, 1 like

Transkrip

Telusuri di video

Pada gambar diatas menjelaskan tentang konsep 'impurity' dalam classification tree dan bagaimana mengukur impurity dengan metode 'Gini Impurity', yang melibatkan perhitungan probabilitas kuadrat dari masing-masing hasil dan penggunaan rata-rata tertimbang untuk mendapatkan total Gini Impurity.

Decision and Classification Trees, Clearly Explained!!!

StatQuest with Josh Starmer 1,018 subscribers

10 likes, 1 share, 1 download, 1 like

Transkrip

Telusuri di video

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Double BAM!!!

Untuk bagian 'false loves soda', semua data menunjukkan 'does not love cool as ice', sehingga kita menyimpulkan bahwa daun tersebut adalah "Does not Love Cool As Ice". Sama halnya dengan

daun untuk umur <12.5 tahun yang true. Selanjutnya, untuk umur di bawah 12.5 tahun yang false, mayoritas hasil dari data di daun menunjukkan "Loves cool as ice".