

Nama: Khairi Hibatullah Ridho

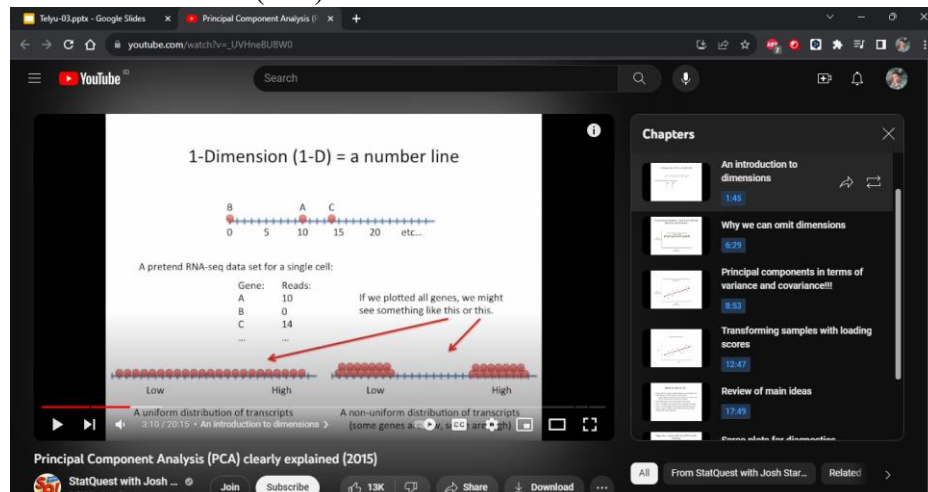
NIM: 1103228240

Kelas: Machine Learning TK44GAB4

Understanding 3 link StatQuest (Youtube: Josh Starmer)

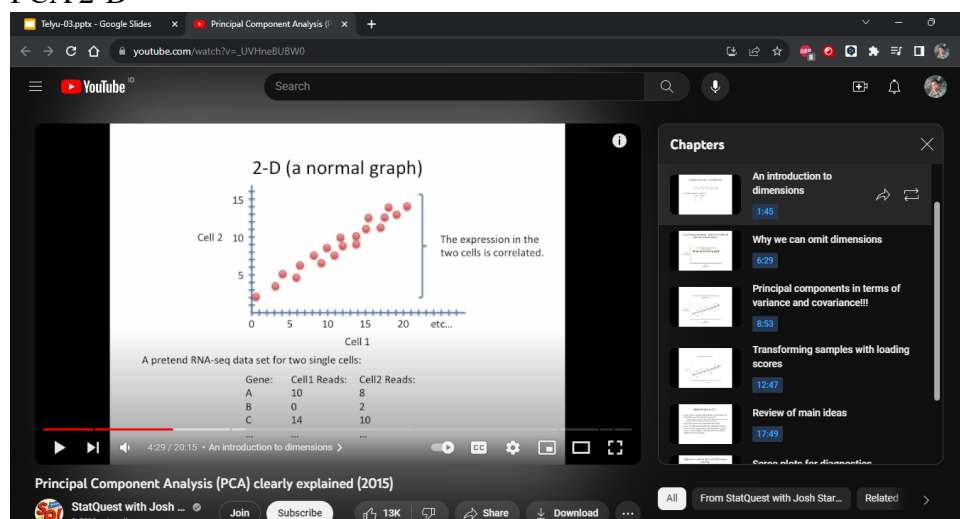
1. Principal Component Analysis (PCA) clearly explained (2015)

- PCA 1-Dimension (1-D) = a number line



PCA adalah teknik statistik yang digunakan untuk mengurangi dimensi data. Data yang memiliki banyak dimensi dapat sulit untuk dianalisis dan dipahami. PCA dapat digunakan untuk mengurangi dimensi data tanpa kehilangan terlalu banyak informasi. Kemudian, dijelaskan di video bahwa PCA 1-Dimension adalah teknik yang mengurangi dimensi data dengan mencari arah perubahan yang paling besar dalam data dengan menggunakan contoh data RNA-seq untuk single cell dengan melakukan plotting genes nya.

- PCA 2-D

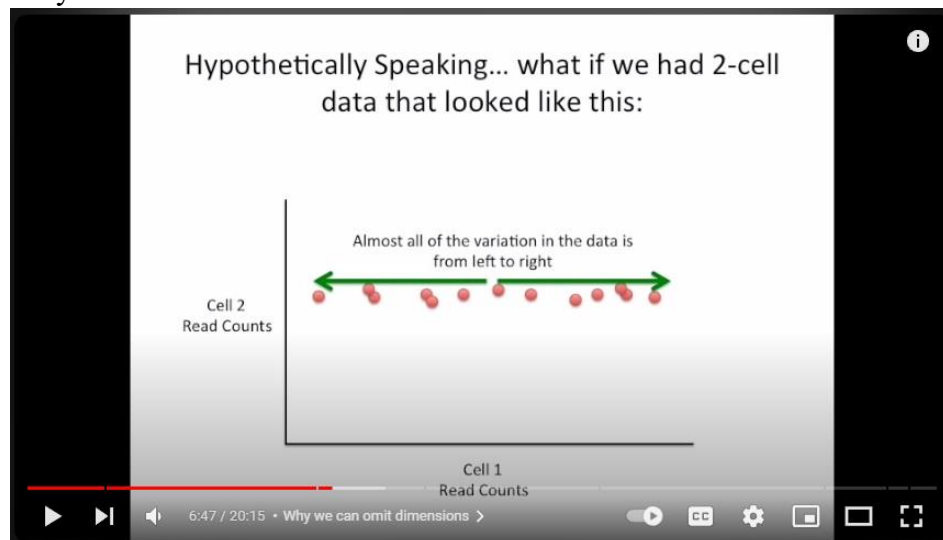


Disini juga menggunakan dataset yang sama dengan sebelumnya, tapi disini dapat dilihat berbagai data yang memiliki korelasi dan data data yang tidak memiliki korelasi pada gene dan cell reads nya.

- Dimensi yang diketahui

Dijelaskan bahwa tiap cell akan menambah dimensi dari graph nya.

- Why we can omit dimensions.

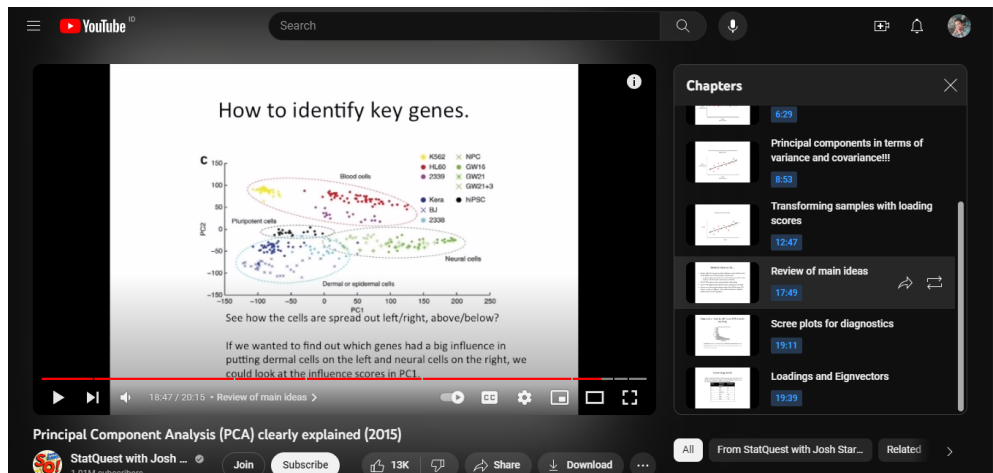


Disini dijelaskan untuk 2 cell data yang terlihat pada grafik ini akan muncul gen yang ditranskripsi rendah dan gen yang ditranskripsi tinggi

- What does all of this have to do with PCA?

PCA menggunakan dataset yang banyak dimensinya dan bisa diperkecil menjadi 2 atau 3 dimensi.

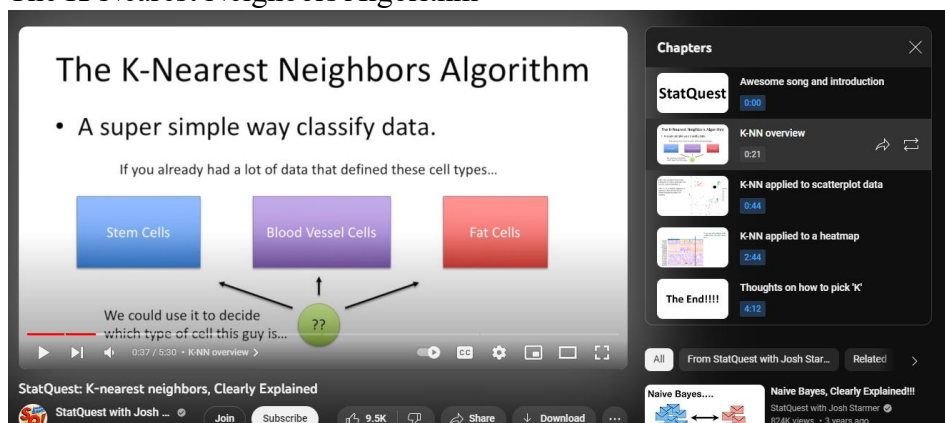
- How to identify key genes



Dilakukan dengan cara melihat gen mana yang mempunyai pengaruh besar dalam menempatkan sel kulit di sebelah kiri dan sel saraf di sebelah kanan, kita dapat melihat skor pengaruhnya di PC1 dan jika kita ingin mengetahui gen mana yang membantu membedakan sel darah dari sel saraf dan sel kulit, kita akan melihat skor pengaruh di PC2.

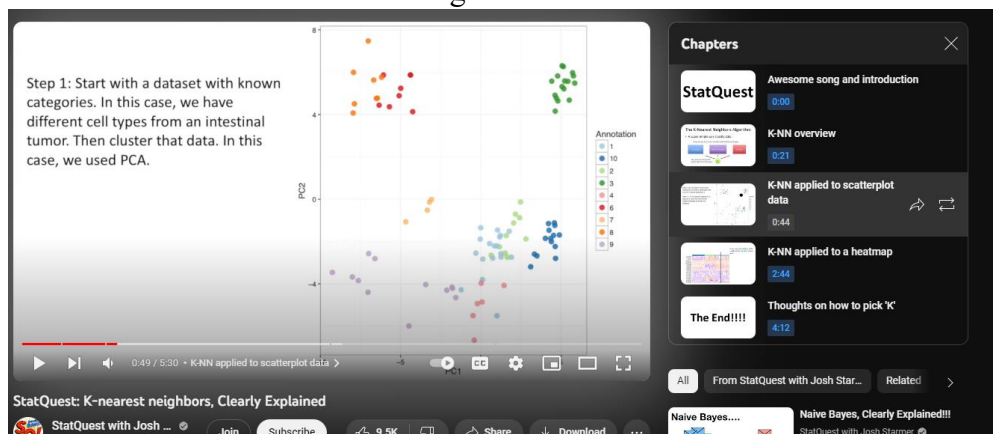
2. StatQuest: K-nearest neighbors, Clearly Explained

- The K-Nearest Neighbors Algorithm



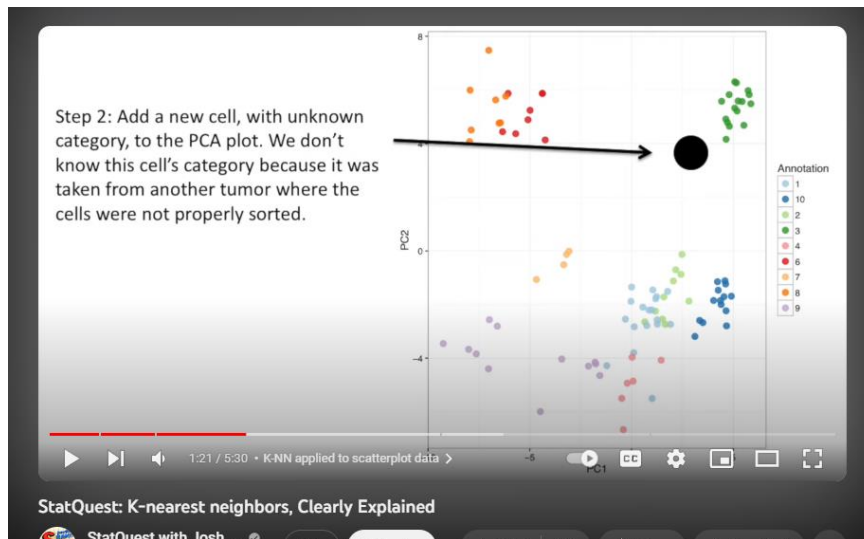
K-Nearest Neighbors yang dijelaskan oleh statquest digunakan untuk mengklasifikasikan data yang bentuknya seperti stem cells, blood vessel cell, dan fat cell untuk mengetahui cell mana yang masuk ke kategori mana.

- Start with dataset with known categories.



Pada case ini ada beberapa tipe cell yang berbeda, data nya diclusterkan menggunakan PCA

- Tambahkan cell baru



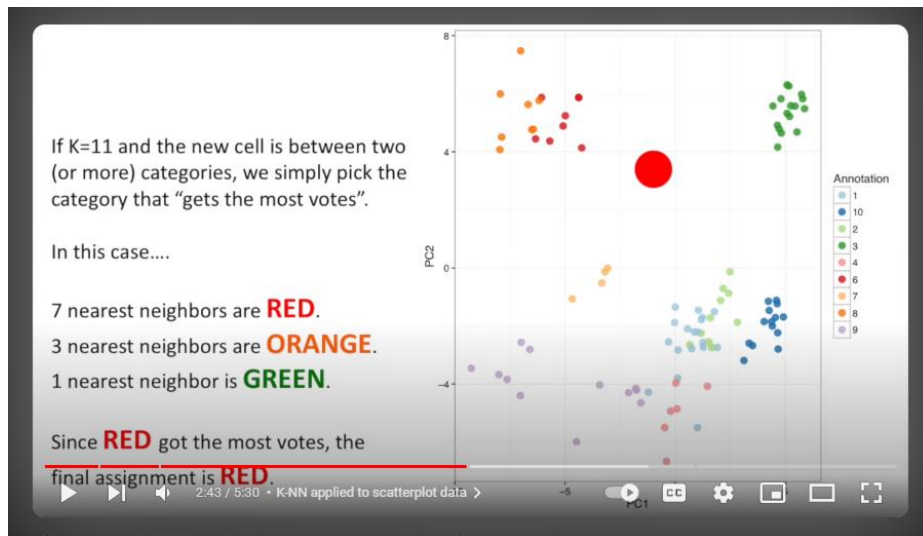
Cell ini belum diketahui karena diambil dari tumor yang belum dilakukan sorting

- Klasifikasi cell baru itu dengan cell terdekatnya (neighbors/tetangga)



Pada kasus ini tetangga terdekatnya itu warna hijau, karena menggunakan K=11.

- Apabila cell ada di Tengah-tengah antara dua cell apa yang harus dilakukan?



Akan dilakukan voting dimana tetangga paling dekatnya ada berapa banyak dan apabila sudah ditemukan tetangga paling dekatnya paling banyak berapa maka itu yang dipilih.

3. Decision and Classification Trees, Clearly Explained!!!

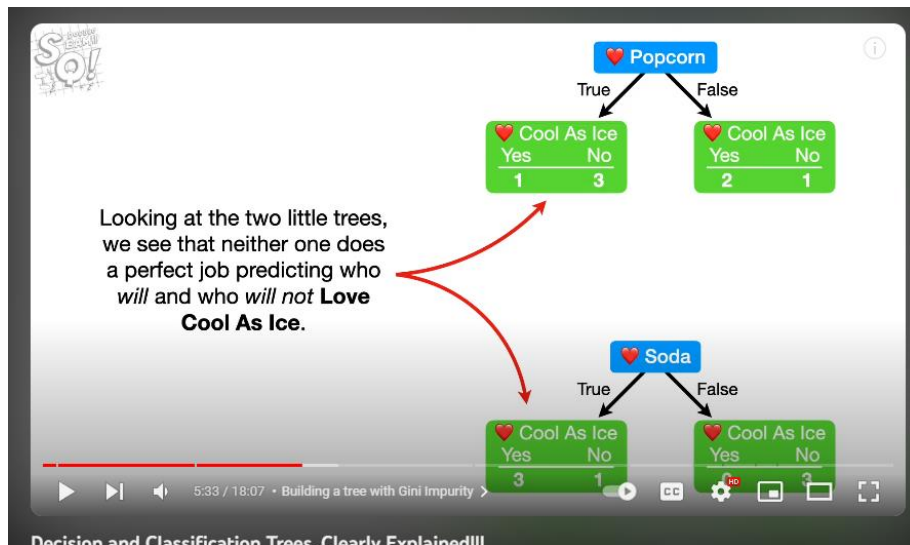
- Menggunakan data mentah

So we will use this data...

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

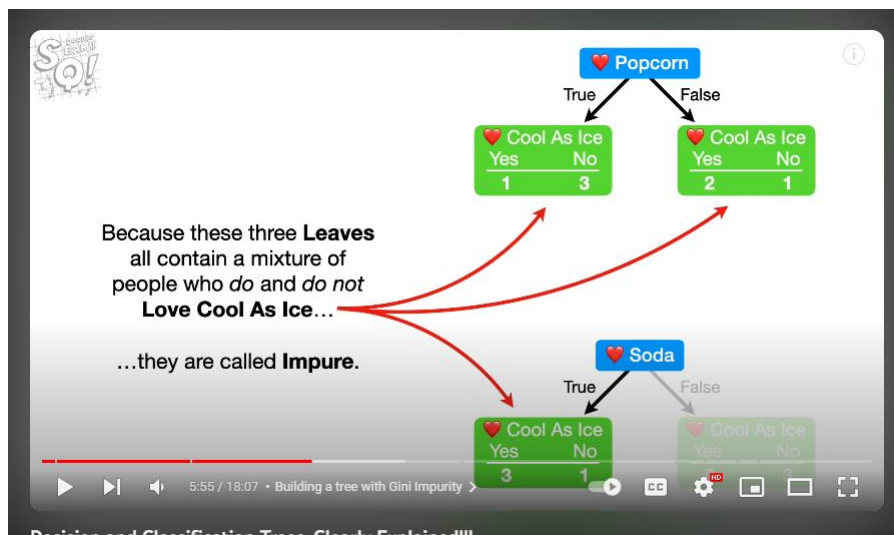
Pada data tersebut dapat dilihat ada 5 kolom data dimana ada "Loves Popcorn, Loves Soda, Age, dan Loves Cool as Ice".

- Dilakukan percobaan



Decision and Classification Trees, Clearly Explained!!!

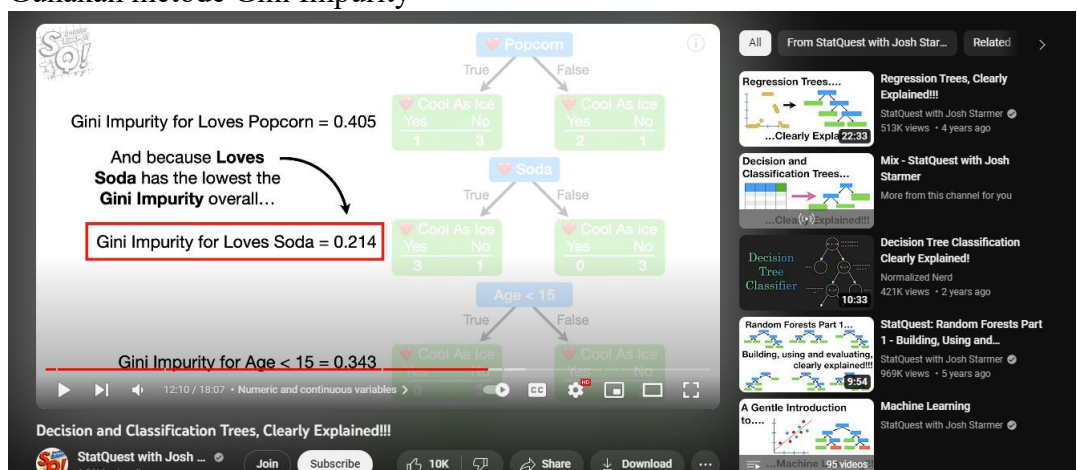
Pertama, dilakukan percobaan dimana orang menyukai popcorn dimana yang ada di data tersebut menyukainya cool as ice atau tidak menyukainya. Kemudian, dilakukan juga hal yang sama untuk orang yang suka soda.



Decision and Classification Trees, Clearly Explained!!!

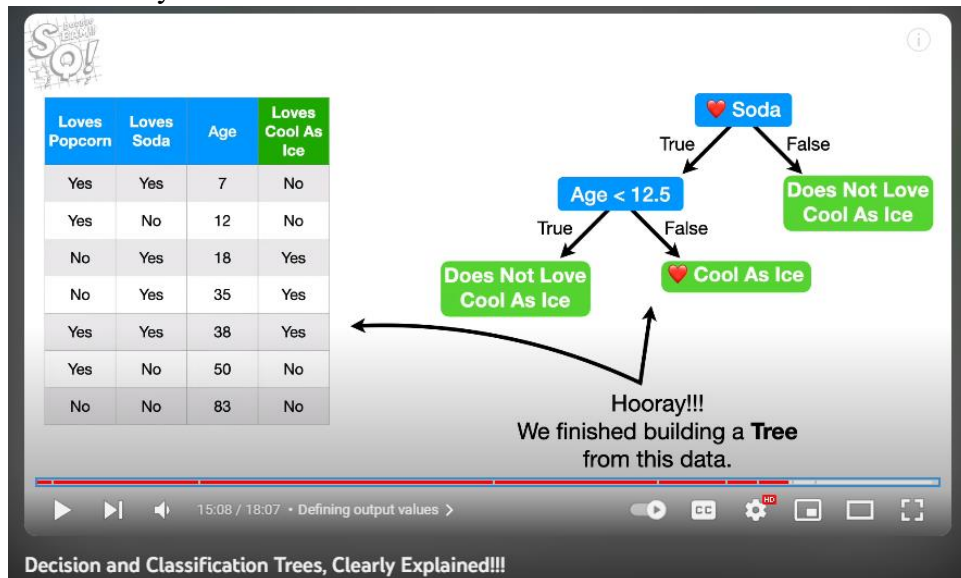
Karena ada 3 daun yang memiliki campuran antara suka dan tidak suka love cool as ice maka bisa bilang datanya tidak murni.

- Gunakan metode Gini Impurity



Metode Gini Impurity digunakan dalam algoritma decision tree untuk mengukur impurity atau ketidakmurnian dari suatu node dalam pohon keputusan. Impurity yang lebih tinggi menunjukkan bahwa node tersebut lebih beragam dan tidak dapat dengan mudah diklasifikasikan. Karena Loves Soda mempunyai gini impurity paling rendah maka Love soda dipasang di bagian paling atas dari tree tersebut.

- Buat Tree nya dari Loves soda tersebut



Karena pada bagian false loves soda semuanya data nya does not love cool as ice maka diambil kesimpulan untuk leaf tersebut “Does not Love Cool As Ice” dan juga pada leaf umur <12.5 tahun yang true. Kemudian, untuk umur dibawah 12.5tahun yang false banyak hasil voting dari data di leaf nya “Loves cool as ice” maka diambil kesimpulan seperti itu.