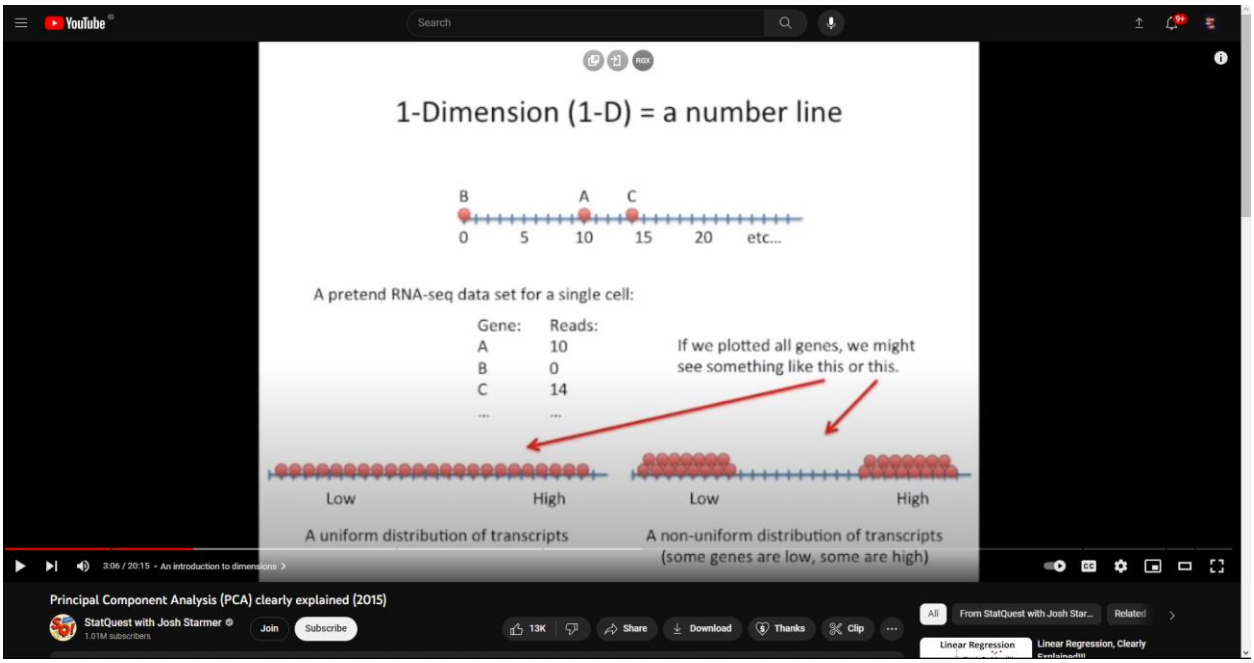
Raken Putra Athallah

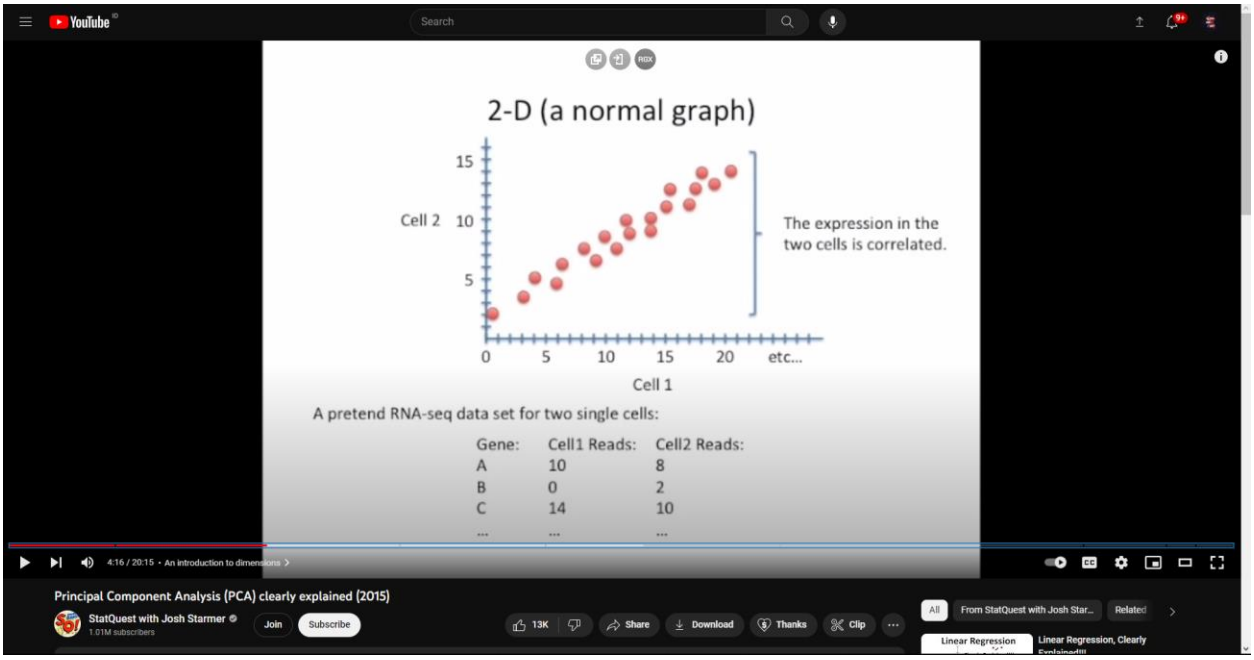
1103204186

Principal component analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk mengurangi jumlah variabel dalam data set tanpa kehilangan informasi yang signifikan.
PCA bekerja dengan memproyeksikan data ke dalam ruang baru dengan dimensi yang lebih sedikit. Dimensi-dimensi ini disebut komponen utama. Komponen utama adalah kombinasi linear dari variabel asli dalam data set.

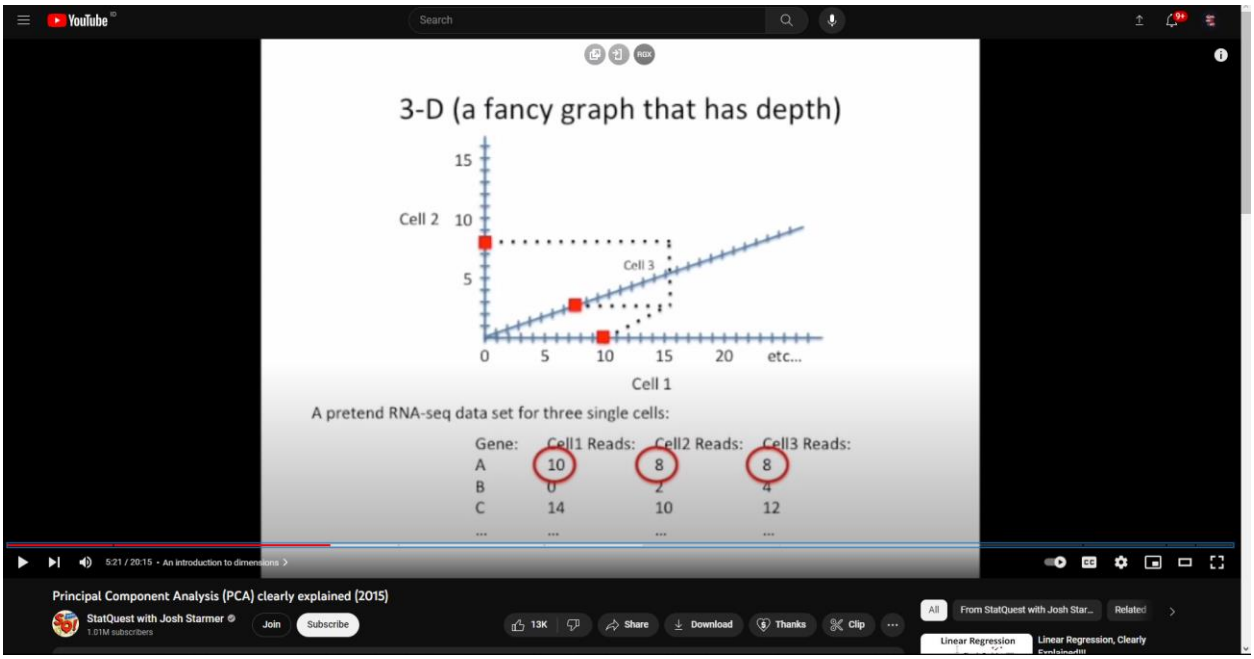
PCA 1 dimensi adalah representasi data yang paling sederhana. Ini hanya mewakili satu variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.



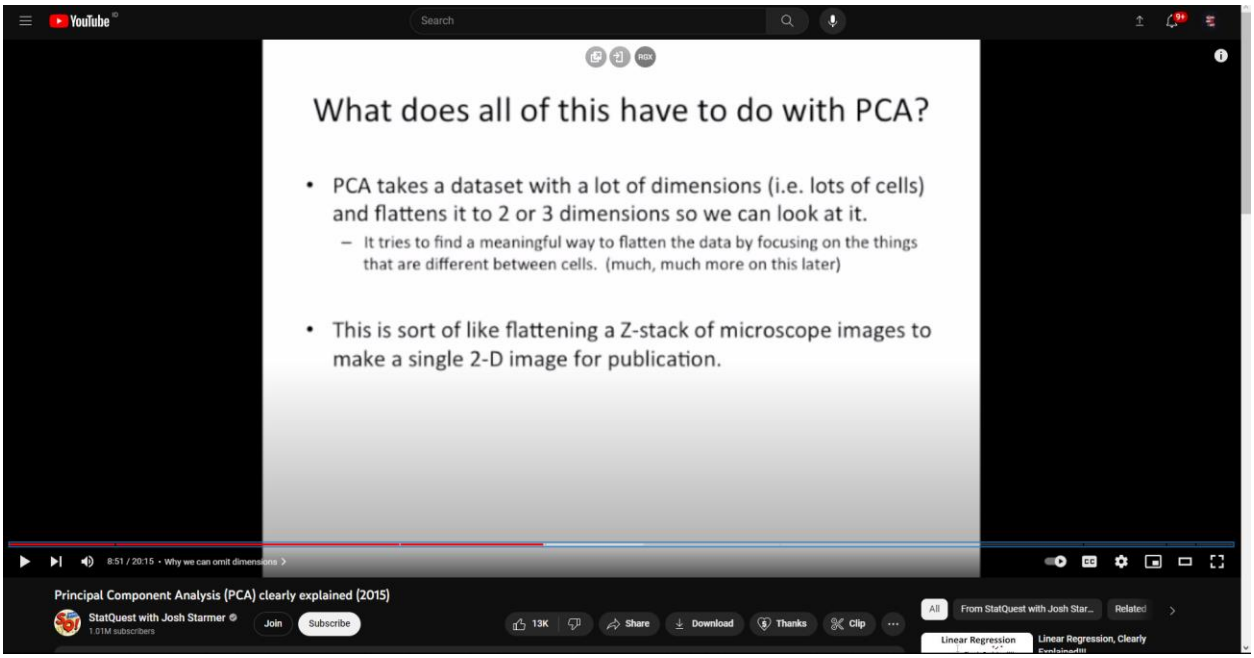
PCA 2 dimensi adalah representasi data yang lebih kompleks. Ini mewakili dua variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.



PCA 3 dimensi adalah representasi data yang paling kompleks. Ini mewakili tiga variabel, yang merupakan kombinasi linear dari semua variabel asli dalam data set.

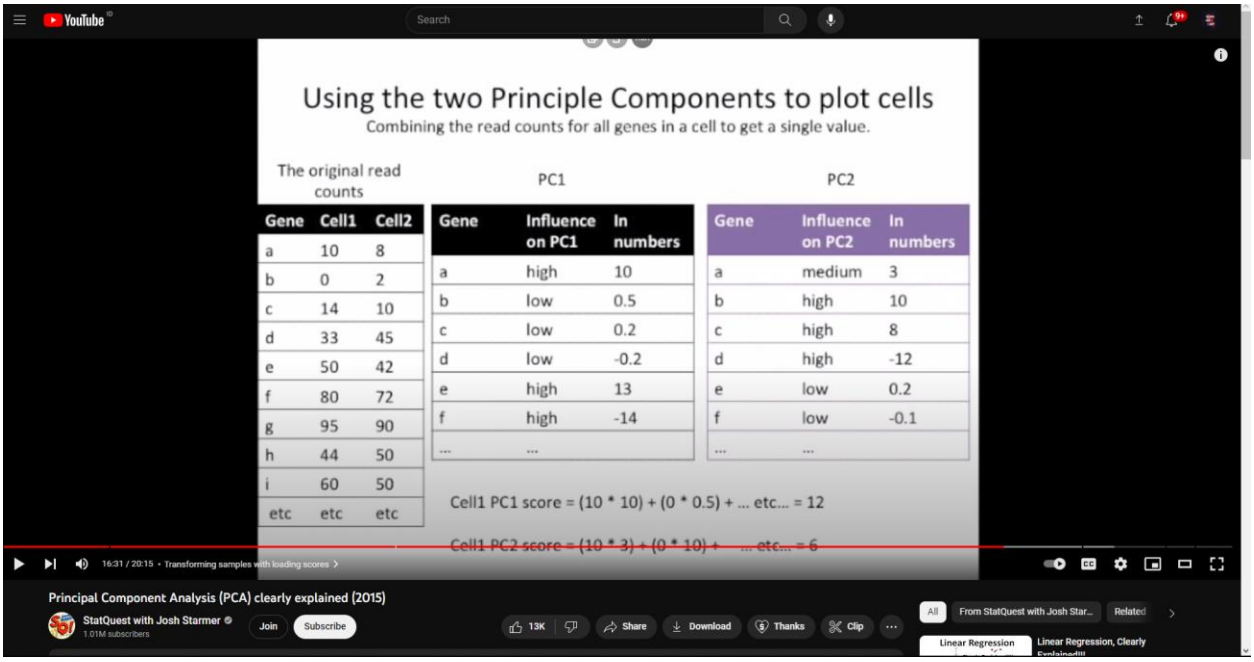


Why we can omit dimension

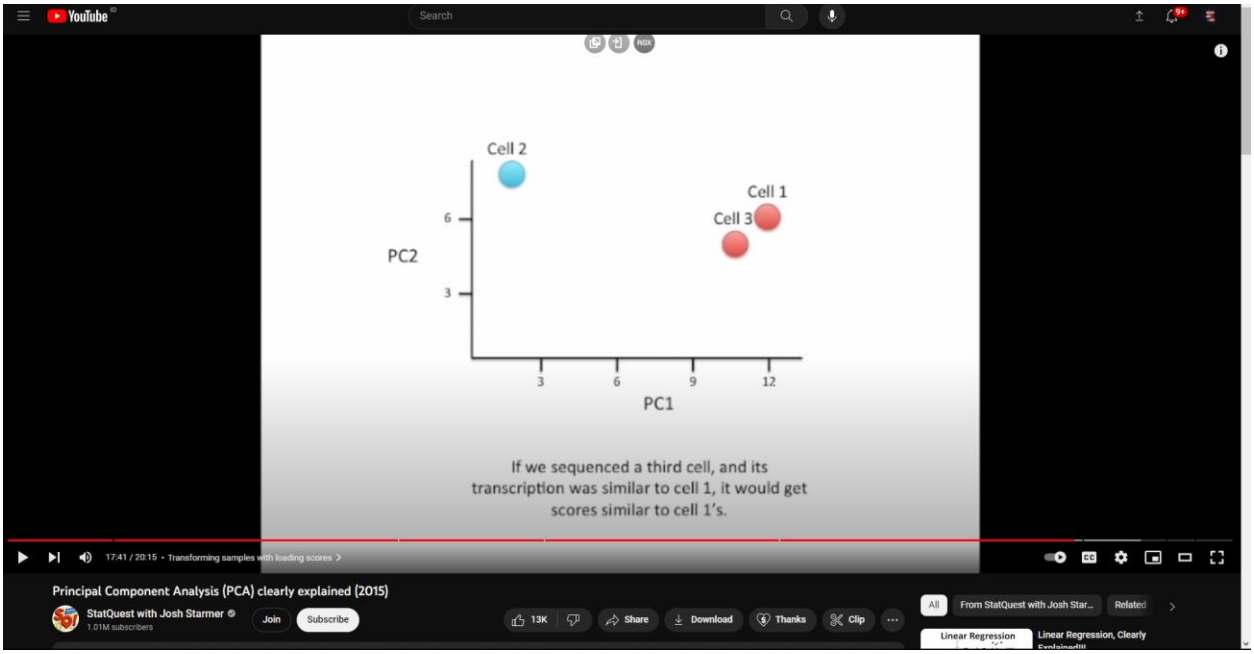


Kita bisa menghilangkan atau menurunkan dimensi untuk memudahkan kita untuk membaca data tanpa harus kehilangan informasi yang signifikan pada sebuah data

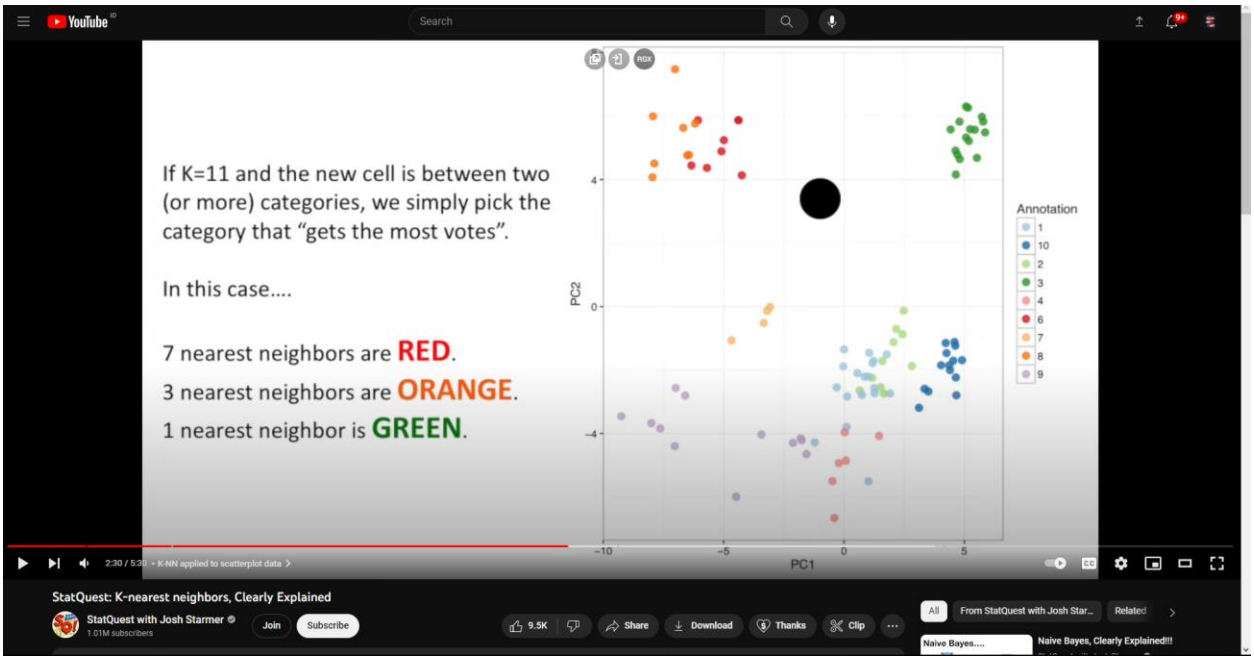
Tergantung pada banyak dimensi yang digunakan setiap dimensi akan mewakilkan 1 principal komponen berdasarkan pada axisnya masing2 masing,seperti contoh pada gambar pc1 mewakilkan most variation pada data dari kiri ke kana dan pc2 mewakilkan most 2nd variation dari atas sampai bawah



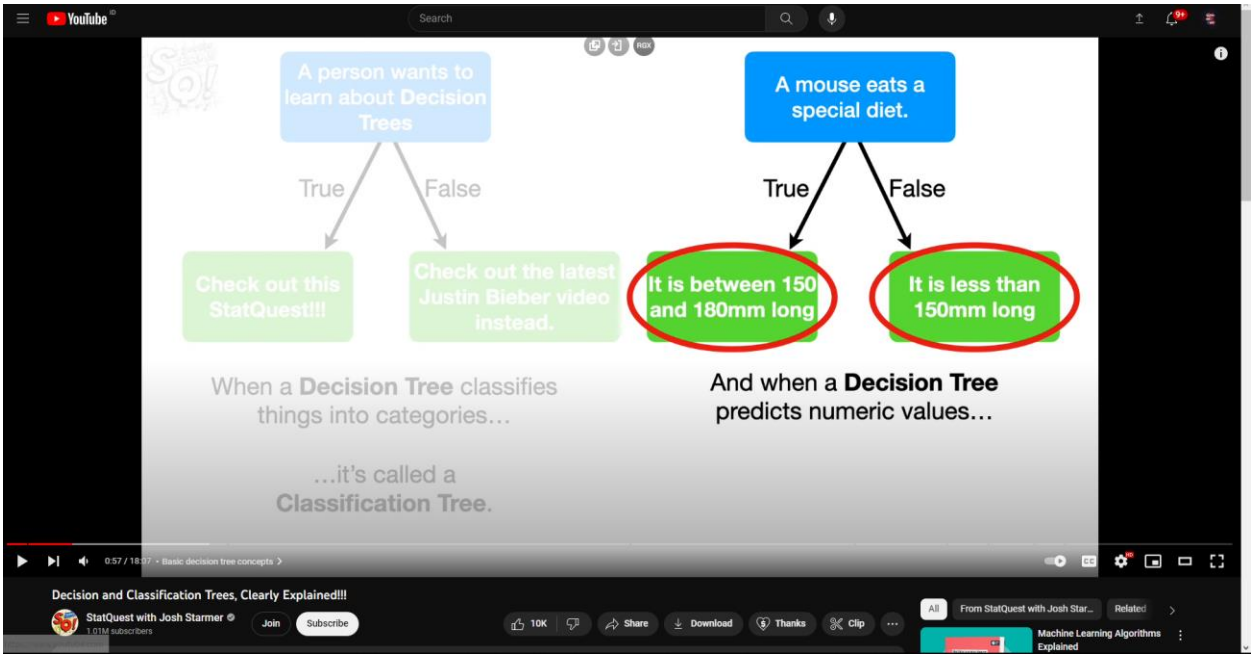
Untuk membuat polt cells dengan principal component dapat melakukannya dengan menentukan score atau seberapa mempengaruhi suatu gene pada suatu principal components(pc) pada contoh pc1 di dalam video genes yang paling berpengaruh adalah genes pada bagian paling kiri atau kanan yang disebut dengan extreme genes(?) .Setelah menentukan dan tingkat influence suatu genes dan memberikannya nilai atau dalam bentuk angka kita dapat menghitung seluruh principal score untuk dibuat menjadi plot cells



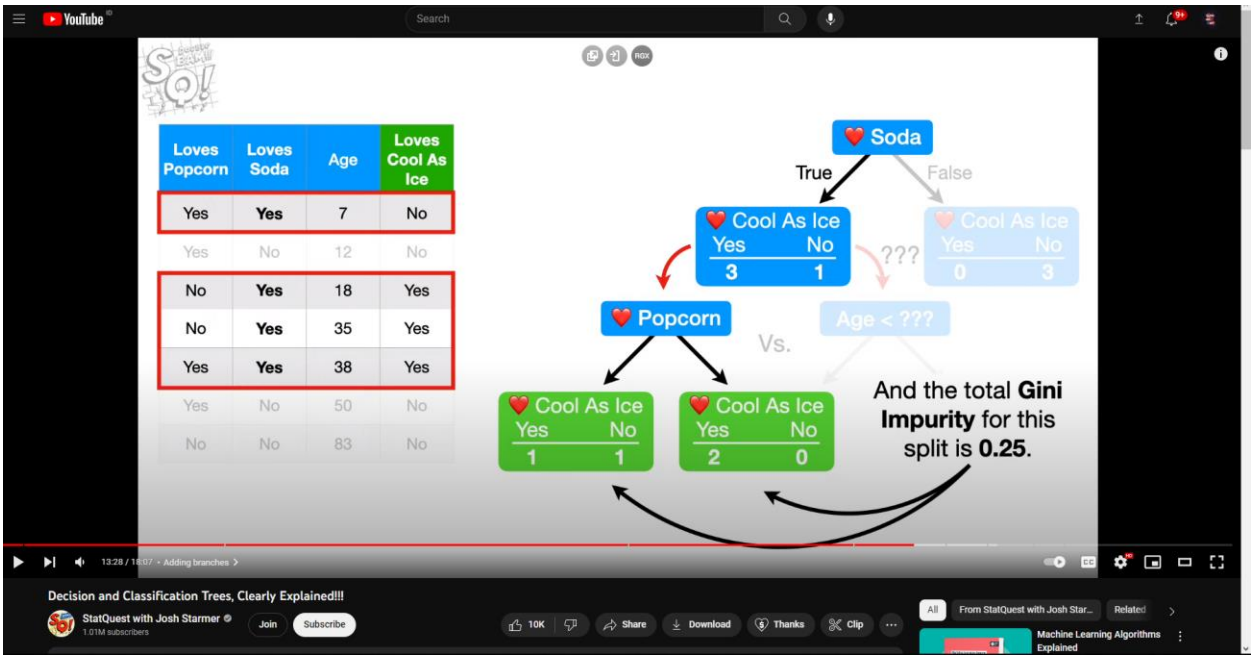
KNN



Pada contoh ini dia menjelaskan bahwa kita dapat menentukan suatu data jika kita sudah mempunyai banyak data yang memenuhi aspek-aspek yang diinginkan.pada contoh ini dengan menggunakan scatterplot sudah ada cells yang sudah diketahui dan ada 1 cell yang belum diketahui,dia menjelaskan kita dapat mengklasifikasikan cell tersebut dengan cell terdekat yang sudah terklasifikasi jika termasuk lebih dari 1 cell yang terklasifikasi kita dapat memilih dengan mencari cell terdekat paling banyak.Prinsip yang sama juga dapat diterapkan jika kita menggunakan heatmap.



Pada dasarnya Decision tree adalah seperti memilihi pilihan berdasarkan suatu statement berdasarkan pada jawabannya benar atau salah.Jika statement atau decision decision tree menggunakan kategori maka termasuk classification tree sedangkan jika menggunakan suatu prediksi maka termasuk regression tree



Untuk membuat decision tree kita haru menentukan apa yang menjadi statement pertama dengan mengambil statement yang memiliki gini Impurity paling sedikit.Impure sendiri adalah ketika suatu statement mendapatkan hasil yang cukup tidak konsisten atau mendapatkan jawaban yang cukup

acak.Gini impurity juga dapat digunakan untuk mebuat branch dengan metode yang hampir mirip dengan metode untuk menentukan statement pertama,kita dapat melakukannya hingga mendapatkan prediksi dari jawaban yang kita inginkan