

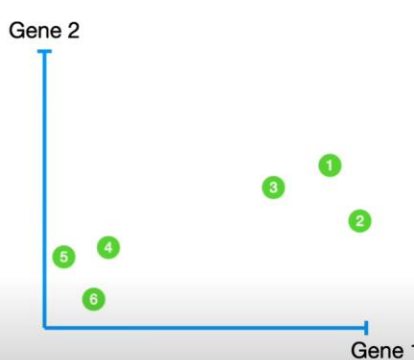
Nama : Rheza Ilham F
NIM : 1103204033
Kelas : TK44GAB4

Understanding 3 Link StatQuest (Youtube : Josh Starmer)

1. Principal Component Analysis (PCA) Step-by-step

In this StatQuest, we're going to go through Principal Component Analysis (PCA) one step at a time using Singular Value Decomposition (SVD). You'll learn what PCA does, how it does it, and how to use it to get deeper insight into your data.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



StatQuest: Principal Component Analysis (PCA), Step-by-Step

StatQuest with Josh Starmer

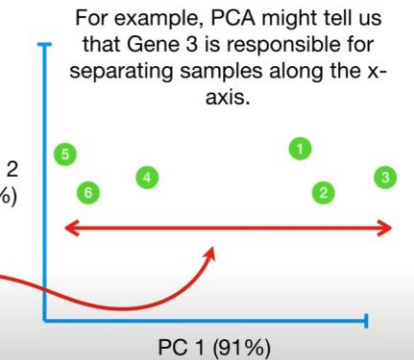
Page 1 of 1 27 words English (Indonesian) Test Predictions: On Accessibility: Good to go

PCA (Principal Component Analysis) adalah suatu teknik statistik yang berguna dalam mengurangi dimensi data sambil menjaga sebagian besar variabilitas data asli. Metode ini dapat diterapkan dalam berbagai situasi, termasuk analisis data genetik seperti "gene 1" dan "gene 2."

- Penerapan PCA pada sumbu x

For example, PCA might tell us that Gene 3 is responsible for separating samples along the x-axis.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7



PC 2 (4%)

PC 1 (91%)

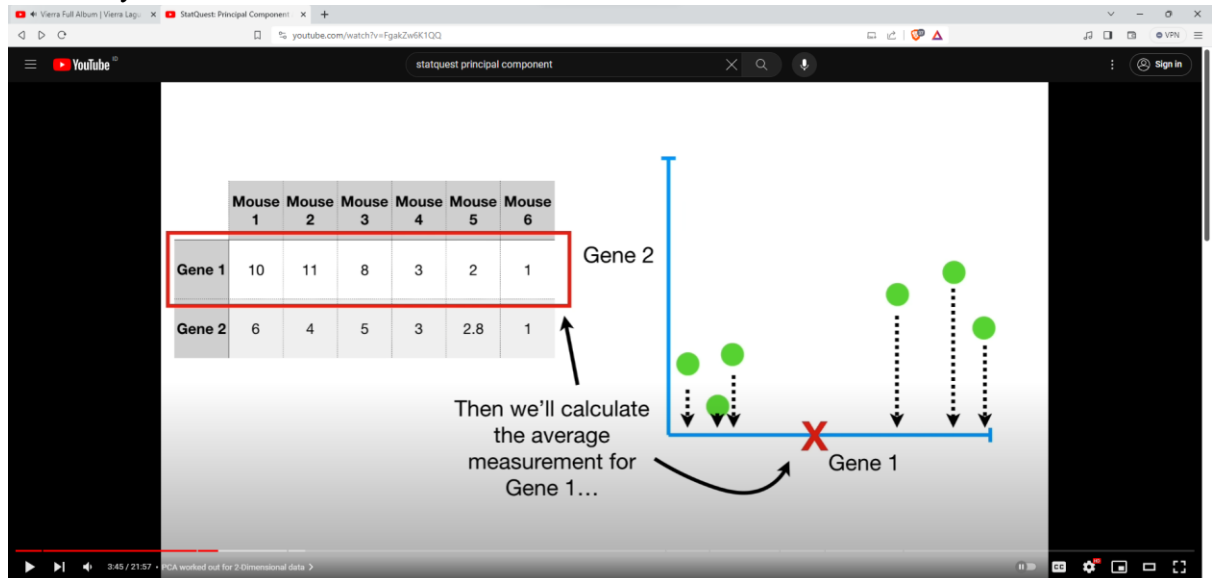
StatQuest: Principal Component Analysis (PCA), Step-by-Step

StatQuest with Josh Starmer

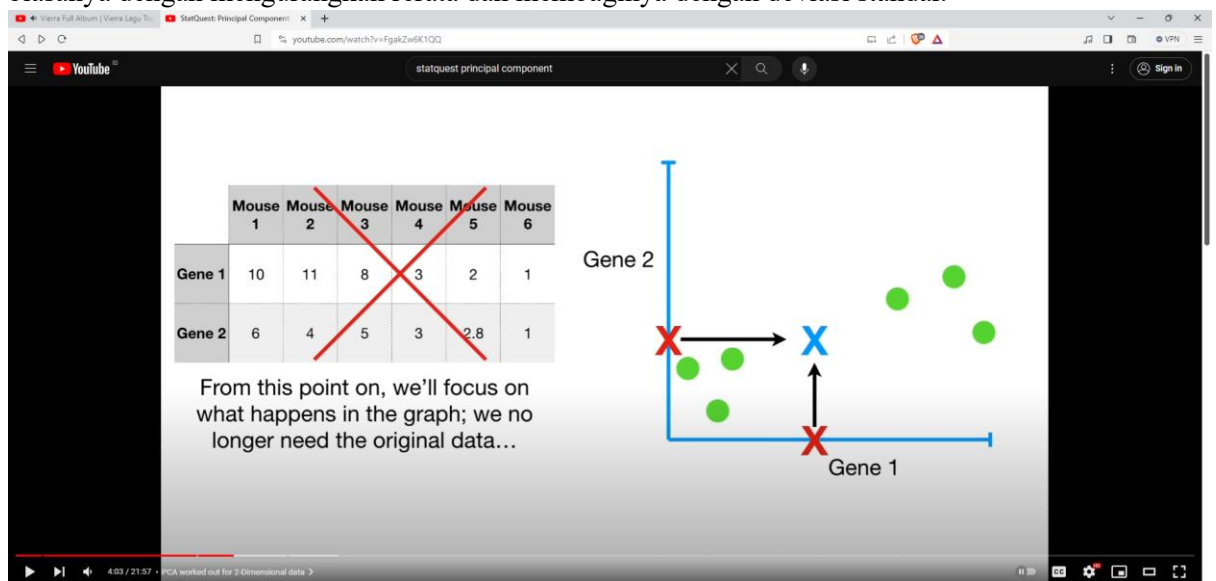
Page 1 of 1 65 words English (Indonesian) Test Predictions: On Accessibility: Investigate

PC1 akan menjadi arah dengan variabilitas tertinggi, dan PC2 akan menjadi arah dengan variabilitas kedua tertinggi. Dengan demikian, PCA akan mengubah data dua dimensi menjadi data dua dimensi yang lebih kompak, tetapi mempertahankan sebagian besar variabilitas asli.

- PCA analytic for 2 dimensional



Standarisasi data untuk memastikan kedua variabel (x dan y) memiliki skala yang sama, biasanya dengan mengurangi rerata dan membaginya dengan deviasi standar.



Mempertahankan komponen yang cukup untuk menangkap persentase tinggi (misalnya, 95%) dari varians total

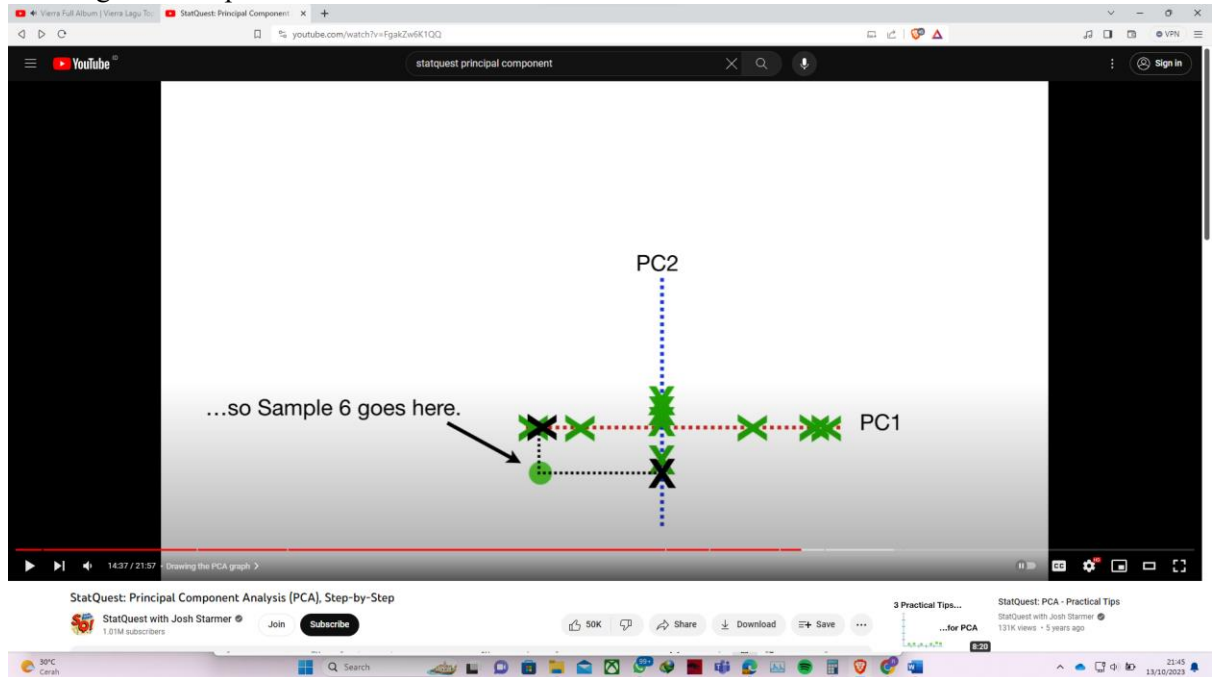
- Find PC1

Ekstrak komponen utama pertama (PC1) dari dataset dengan mengambil produk dot data standar yang asli dengan vektor eigen yang sesuai dengan nilai eigen terbesar. PC1 mewakili kombinasi linear dari fitur asli yang menjelaskan variansi terbesar dalam data.

- Find PC2

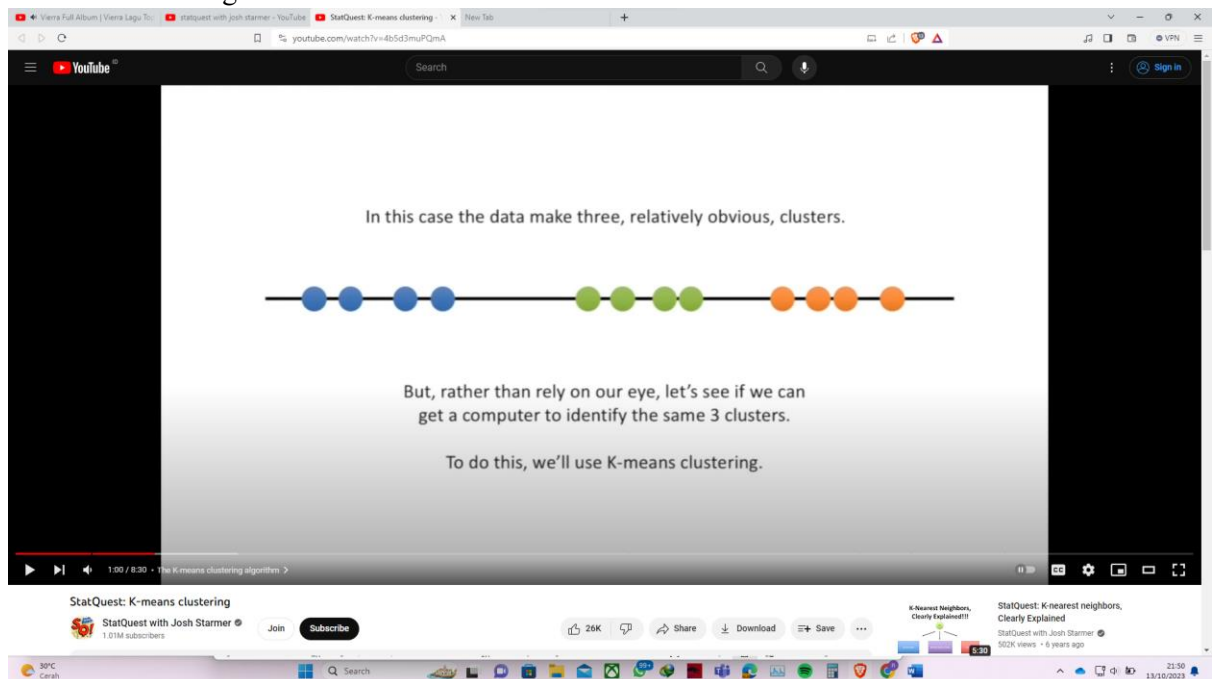
Menghitung komponen utama kedua (PC2) dengan menggunakan data yang telah "dipotong" pada PC1

- Drawing PCA Graph



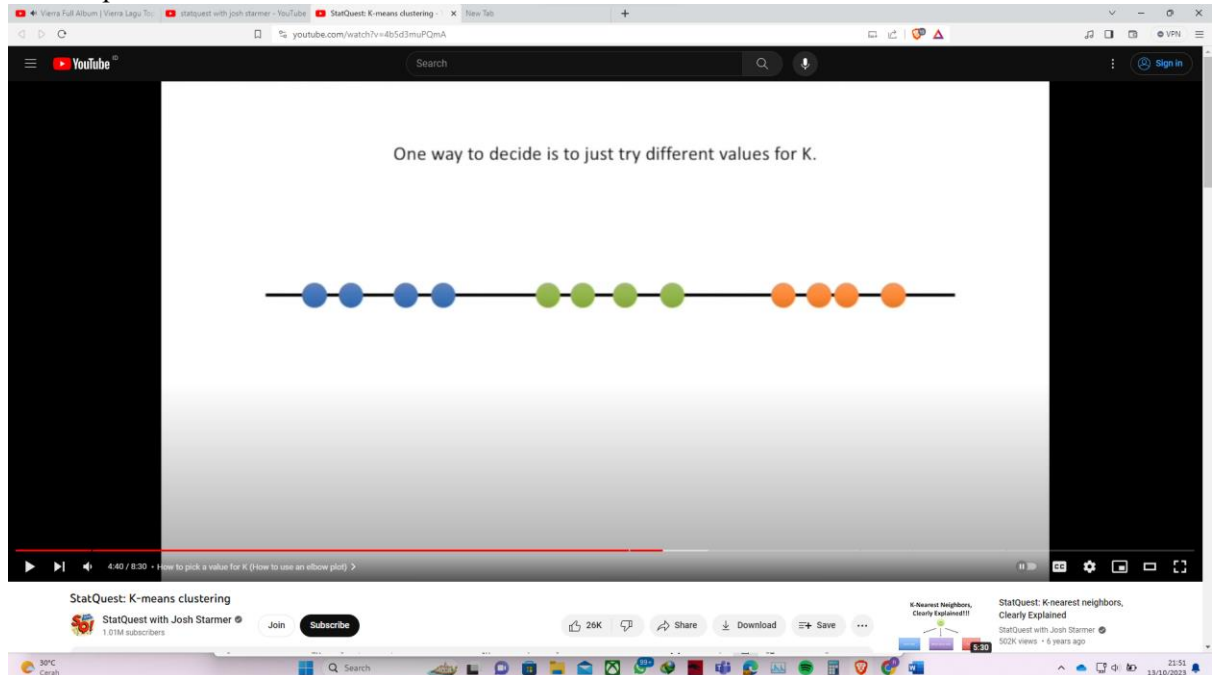
Grafik PCA adalah cara yang baik untuk memahami sejauh mana variasi data dapat dijelaskan oleh komponen utama yang ada. Semakin terpisah poin-poin data dalam grafik PCA, semakin baik komponen utama dapat menjelaskan variasi dalam data.

2. K-means clustering



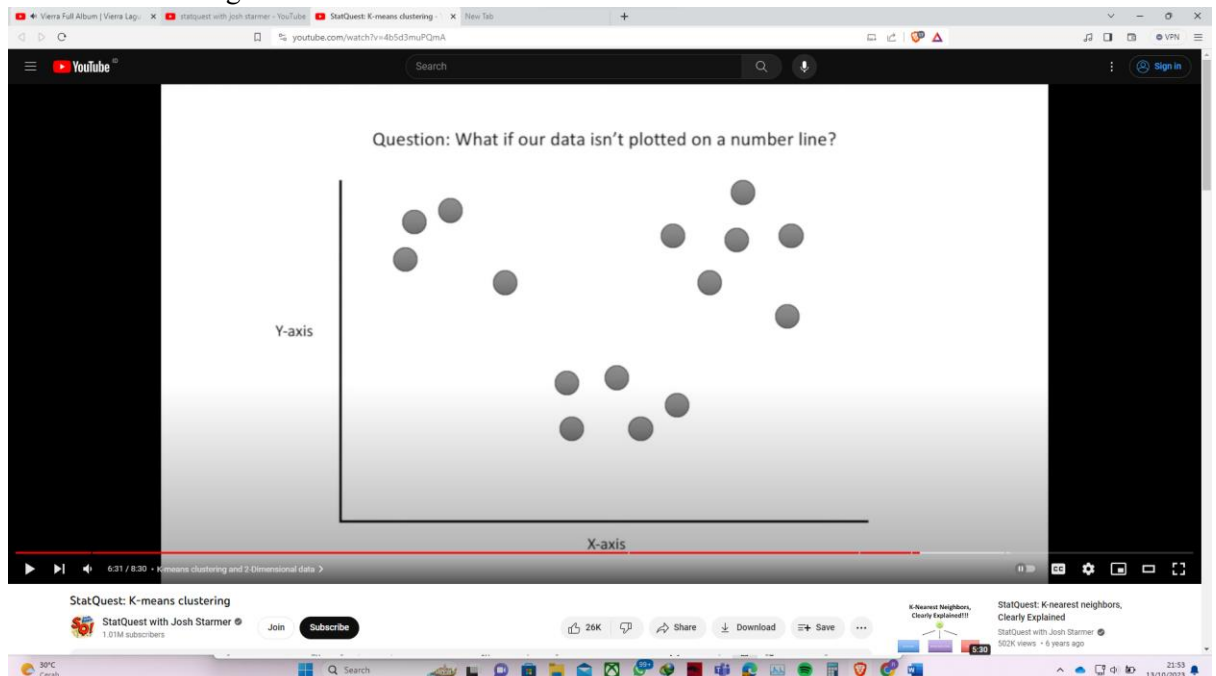
K-Means Clustering adalah salah satu algoritma dalam data mining dan analisis data yang digunakan untuk mengelompokkan data berdasarkan kemiripan fitur atau atribut mereka. Algoritma ini mencoba untuk mengelompokkan data ke dalam sejumlah kelompok atau cluster, di mana setiap cluster terdiri dari data yang memiliki karakteristik yang serupa.

- How to pick a value K



Metode Elbow adalah salah satu metode yang paling umum digunakan. Ini melibatkan menjalankan algoritma K-Means untuk berbagai nilai K dan memplotkan jumlah jarak kuadrat antara titik data dan pusat kluster yang telah ditetapkan. Titik di mana jarak kuadrat tersebut mulai menurun pada tingkat yang lebih lambat menyerupai "siku" pada grafik. Titik "siku" ini adalah indikasi yang baik untuk nilai K yang optimal.

- K-Means clustering 2-dimensional data



Ketika berbicara tentang "2 dimensional," itu berarti sedang memproses data yang memiliki dua fitur atau dimensi. Contohnya adalah data yang dapat diplot pada grafik dua dimensi, seperti sumbu x dan y.

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

When a **Decision Tree** classifies things into categories...
...it's called a **Classification Tree**.

And when a **Decision Tree** predicts numeric values...

Pohon Klasifikasi adalah bentuk yang umum dari Pohon Keputusan yang digunakan dalam klasifikasi data. Dalam Pohon Klasifikasi, setiap simpul atau "node" pada pohon mewakili keputusan atau pengujian terhadap atribut data. Setiap cabang pada pohon menggambarkan hasil dari pengujian tersebut, dan daun atau "leaf" mewakili kelas atau hasil akhir dari pemilihan keputusan.

- Metode Gini

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

And because they do not Love Cool As Ice...

Metode Gini adalah salah satu metode yang umum digunakan dalam pembuatan Pohon Klasifikasi. Ini digunakan untuk mengukur sejauh mana ketidakmurnian atau "impurity" dalam sebuah node pada pohon. Node yang lebih murni akan memiliki nilai Gini yang lebih rendah.

- Numerik and continuous variables

Age < 9.5

True

False

Cool As Ice

Yes No

0 1

Cool As Ice

Yes No

3 3

Gini Impurity = 0

Impurity = 1 - (probability of "Yes")² - (probability of "No")²

= 1 - ($\frac{3}{3+3}$)²

Then we calculate the **Gini Impurity** for the **Leaf** on the right...

Variabel ini dapat memiliki nilai-nilai yang berbeda dan sering digunakan dalam analisis statistik. Contoh variabel numerik adalah usia seseorang, pendapatan tahunan, suhu, dan sebagainya.

- Branches

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

Soda

True

False

Cool As Ice

Yes No

3 1

Cool As Ice

Yes No

0 3

Popcorn

Cool As Ice

Yes No

1 1

Cool As Ice

Yes No

2 0

Age < ???

The remaining 2 people that **Love Soda**, but **do not Love Popcorn** end up on the right.

Penambahan cabang pada pohon keputusan adalah salah satu aspek penting dalam membangun model yang lebih kompleks dan informatif. Cabang-cabang ini juga disebut sebagai "node" atau "split" dalam pohon keputusan. Penambahan cabang dapat dilakukan dengan mempertimbangkan variabel yang paling relevan untuk memisahkan data ke dalam kelompok yang berbeda.