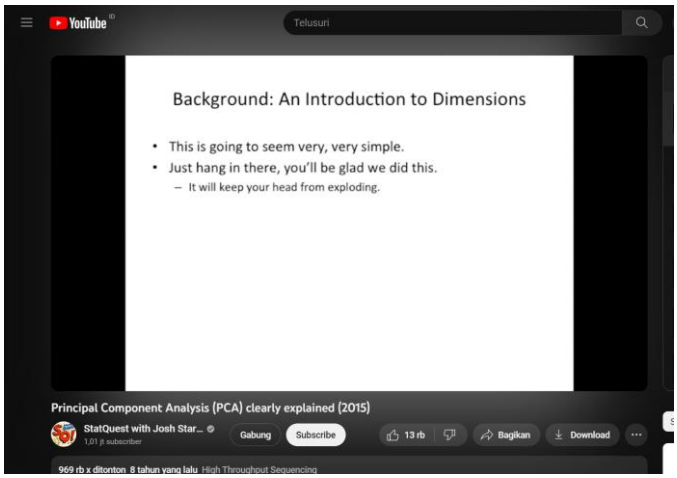
Nama : Ilham Muhamad Firdaus

NIM : 1103202001

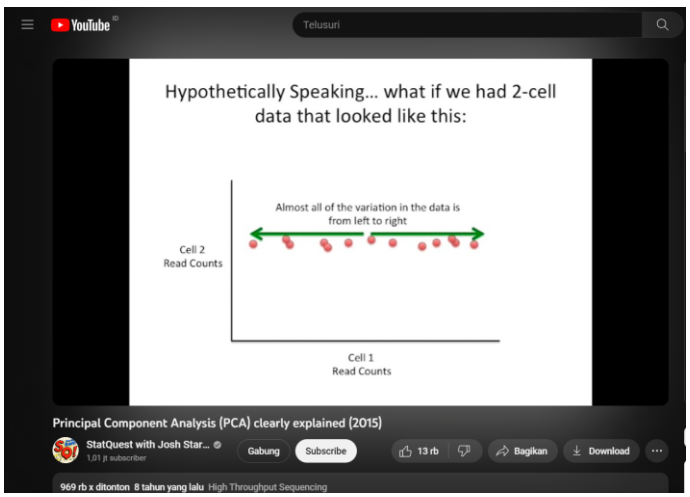
Kelas : Machine Learning TK44G4

Understanding 3 Link StatQuest

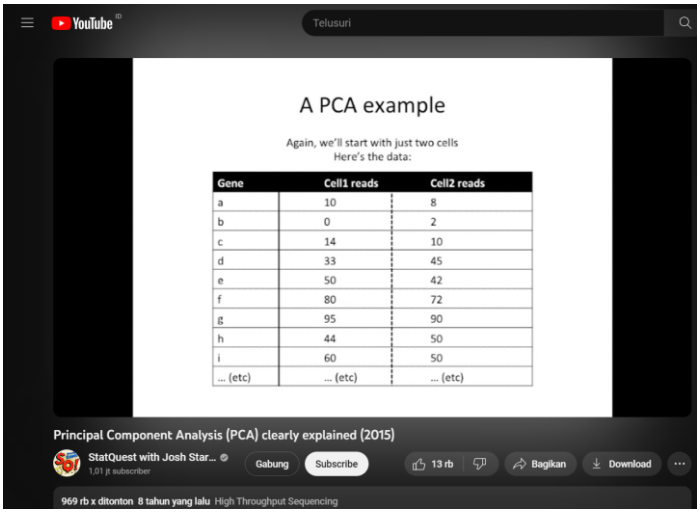
1. Principal Component Analysis (PCA)



Principal Component Analysis (PCA) adalah teknik statistik yang digunakan untuk mengurangi dimensi set data yang kompleks dan bervolume besar dengan mengekstraksi komponen utama yang mengandung informasi paling banyak. PCA adalah teknik yang populer untuk menganalisis dataset besar yang berisi sejumlah besar dimensi/fitur per observasi. Tujuan dari PCA adalah untuk mengurangi jumlah variabel dari sebuah set data dengan tetap mempertahankan informasi sebanyak mungkin.



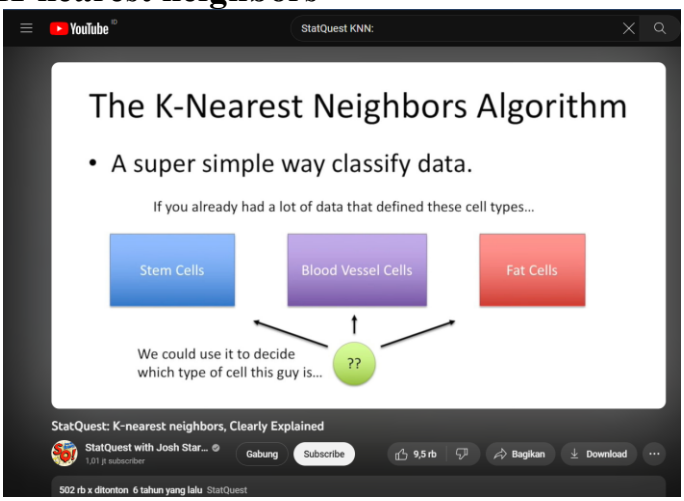
Tujuan dari PCA adalah untuk mengurangi jumlah variabel dari sebuah set data dengan tetap mempertahankan informasi sebanyak mungkin.



Berikut ini adalah contoh langkah demi langkah cara melakukan Analisis Komponen Utama (PCA):

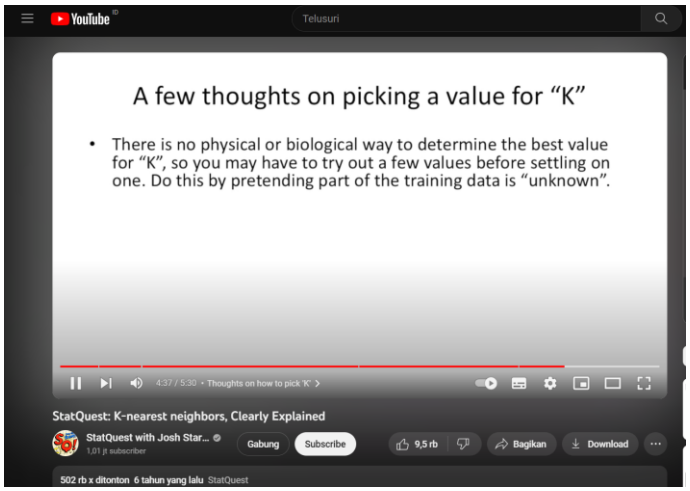
1. Standarisasi dataset dengan menghitung rata-rata dan deviasi standar setiap variabel.
2. Hitung matriks kovarians dari set data yang telah distandarisasi.
3. Hitung nilai eigen dan vektor eigen dari matriks kovarians.
4. Urutkan vektor eigen berdasarkan nilai eigen yang sesuai dalam urutan menurun.
5. Pilih k vektor eigen teratas yang menjelaskan sebagian besar varians dalam data, di mana k adalah jumlah komponen utama yang diinginkan.
6. Transformasi data ke dalam ruang dimensi-k yang baru dengan mengalikan data asli dengan k vektor eigen.

2. StatQuest: K-nearest neighbors



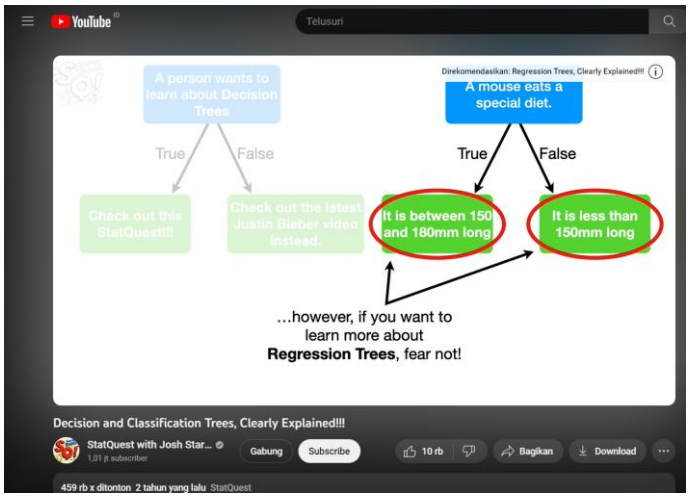
K-nearest neighbors (KNN) adalah algoritme pembelajaran non-parametrik dan terawasi yang digunakan untuk tugas klasifikasi dan regresi. Ini adalah algoritme sederhana yang menyimpan semua kasus yang tersedia dan mengklasifikasikan data baru berdasarkan ukuran kemiripan.

Secara umum, penerapan KNN pada heatmap melibatkan penggunaan algoritme KNN untuk memprediksi nilai untuk setiap sel pada heatmap dan kemudian memvisualisasikan nilai yang diprediksi pada heatmap menggunakan pustaka seperti Plotly atau Seaborn.



Ketika menggunakan algoritma K-nearest neighbors (KNN), memilih nilai K yang optimal sangat penting karena dapat mempengaruhi akurasi model secara signifikan. Ada beberapa cara untuk memilih K, termasuk memvisualisasikan plot tingkat kesalahan vs K, menggunakan validasi silang, memilih angka ganjil untuk K, mencoba nilai K yang berbeda, dan menggunakan akar kuadrat dari jumlah titik data.

3. Decision and Classification Trees



Decision tree adalah jenis algoritma supervised machine learning yang digunakan untuk mengkategorikan atau membuat prediksi berdasarkan bagaimana serangkaian pertanyaan sebelumnya dijawab. Model ini merupakan bentuk struktur pohon seperti flowchart di mana setiap simpul internal menunjukkan fitur, cabang menunjukkan aturan, dan simpul daun menunjukkan hasil algoritma.

Algoritma Decision Tree dengan Gini impurity melibatkan penghitungan Gini impurity dari dataset dan setiap fitur, memilih fitur yang menghasilkan Gini impurity terendah, membagi dataset pada fitur yang dipilih, dan mengulangi proses tersebut untuk setiap subset hingga semua subset murni atau kriteria penghentian terpenuhi.