Nama : Eva Fiorina Siahaan
NIM : 1103210101
Tugas Minggu 5 (Membuat Catatan PCA)

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

PCA (Principal Component Analysis) adalah metode statistik yang digunakan untuk mereduksi dimensi data dengan mempertahankan sebanyak mungkin informasi yang relevan. Proses ini dapat membantu kita memahami struktur dalam data yang kompleks dengan mengidentifikasi pola-pola yang tersembunyi di dalamnya. PCA dapat dijelaskan secara konseptual dengan menguraikan langkah-langkahnya menggunakan dekomposisi nilai singular (SVD). SVD adalah teknik matematis yang mendasari PCA untuk memecah data menjadi komponen-komponen utama. Tujuan PCA adalah mencari garis yang meminimalkan jarak tersebut, dengan kata lain garis yang memaksimalkan jarak dari titik proyeksi ke titik asal. Untuk memahami secara intuitif, jika garis lebih cocok dengan data, maka jarak dari titik garis berkurang, sedangkan jarak dari titik proyeksi ke titik asal bertambah. Secara matematis, PCA menggunakan teorema Pythagoras untuk mengukur jarak antara titik data dan garis proyeksi. Ketika titik data diproyeksikan ke dalam garis lurus, maka terbentuk sudut siku-siku antara garis proyeksi dan garis yang ditarik dari titik asal ke titik data. Ia dapat membagi jarak menjadi dua bagian yang bergerak berlawanan arah seiring bertambahnya jarak.

Untuk menggambar plot PCA, yaitu dengan cara memutar PC1 hingga menjadi horizontal, kemudian plot sampel pada plot tersebut menggunakan titik proyeksi. Plot penghalusan adalah representasi grafis dari persentase variasi untuk setiap komponen utama dan membantu memahami kontribusi setiap komponen terhadap variasi total. Untuk PCA tiga variabel (misalnya tiga gen), prosesnya hampir sama dengan PCA dua variabel. Jika datanya terpusat, maka garis yang paling sesuai adalah PC1, disusul PC2 yang tegak lurus PC1, dan seterusnya untuk PC3. Dalam contoh ini, gen dengan kontribusi terbesar pada masing-masing komponen utama menjadi fokus utama pembentukan resep komponen utama tersebut. Jadi, dengan menambahkan lebih banyak variabel (seperti gen), proses PCA menemukan lebih banyak komponen utama dengan menambahkan garis diagonal dan memutarnya agar lebih mencerminkan struktur data dengan lebih baik.

Kualitas komponen utama dapat dievaluasi dengan menggunakan nilai eigen. Nilai eigen mewakili jumlah jarak kuadrat antara titik proyeksi dan titik asal untuk masing-masing komponen utama, dan semakin besar nilai eigen, semakin banyak variasi yang dijelaskan oleh komponen utama. Dengan memahami konsep ini, kita dapat menggambar diagram PCA dua dimensi yang mewakili keberadaan sampel di ruang angkasa berdasarkan komponen utamanya (misalnya PC1 dan PC2).Ini membantu memvisualisasikan pola yang sulit dilihat pada data asli. Oleh karena itu, PCA adalah metode yang efisien untuk mereduksi dimensi data, memahami pola yang tersembunyi, dan mengidentifikasi variabel terpenting dalam data yang dikelompokkan.