

## PCA

[https://scikit-learn.org/stable/auto\\_examples/decomposition/plot\\_pca\\_iris.html](https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html)

contoh penggunaan Python dengan library seperti matplotlib, numpy, dan scikit-learn untuk menggambarkan dataset Iris dalam tiga dimensi (3D) dengan menggunakan analisis komponen utama (PCA) untuk mengurangi dimensi data. Setelah memuat dataset Iris, code ini membuat plot 3D yang menunjukkan cluster data berdasarkan tiga komponen utama yang dihasilkan dari PCA. Setiap kelas dalam dataset diberi label dengan teks yang ditempatkan di pusat cluster masing-masing. Warna titik-titik data digunakan untuk membedakan kelas dengan mengganti urutan label kelas. Label sumbu x, y, dan z dihapus untuk memberikan fokus pada visualisasi data. Hasilnya adalah visualisasi 3D yang membantu memahami pola dan hubungan antara berbagai kelas dalam dataset Iris dalam tiga dimensi.

## LDA

[https://scikit-learn.org/0.16/auto\\_examples/decomposition/plot\\_pca\\_vs\\_lda.html](https://scikit-learn.org/0.16/auto_examples/decomposition/plot_pca_vs_lda.html)

Dataset Iris menggambarkan tiga jenis bunga Iris (Setosa, Versicolour, dan Virginica) dengan empat atribut, yaitu panjang sepal, lebar sepal, panjang petal, dan lebar petal. Code ini menerapkan Analisis Komponen Utama (PCA) untuk mengidentifikasi kombinasi atribut yang menjelaskan sebagian besar variasi dalam data, dan hasilnya divisualisasikan dalam dua komponen utama pertama. Selain itu, Analisis Diskriminan Linear (LDA) juga digunakan untuk mengidentifikasi atribut yang menjelaskan perbedaan antar kelas. Berbeda dengan PCA, LDA adalah metode terpandu yang menggunakan label kelas yang sudah diketahui. Hasil PCA dan LDA juga menggambarkan proporsi variasi yang dijelaskan oleh masing-masing komponen. Dua plot hasil PCA dan LDA menunjukkan sebaran data Iris dalam dua dimensi yang membantu dalam memahami perbedaan dan pola antar kelas.

## SVD

[https://scikit-learn.org/stable/auto\\_examples/text/plot\\_document\\_clustering.html](https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html)

merupakan contoh penggunaan API scikit-learn untuk mengelompokkan dokumen berdasarkan topik menggunakan pendekatan Bag of Words. Dalam code ini, dua algoritma clustering ditunjukkan, yaitu KMeans dan varian yang lebih scalable, MiniBatchKMeans. Selain itu, latent semantic analysis (LSA) digunakan untuk mengurangi dimensi dan menemukan pola tersembunyi dalam data.

Dataset yang digunakan dalam contoh ini adalah "The 20 newsgroups text dataset," yang terdiri dari sekitar 18.000 posting di 20 topik newsgroups. Namun, untuk tujuan ilustrasi dan mengurangi biaya komputasi, hanya dipilih subset 4 topik yang berisi sekitar 3.400 dokumen. Dalam code ini, dilakukan evaluasi clustering dengan metrik seperti homogeneity, completeness, V-measure, Rand-Index, dan lainnya. Evaluasi dilakukan dengan membandingkan hasil clustering dengan label kelas yang sudah diketahui.

Dua metode ekstraksi fitur digunakan, yaitu TfidfVectorizer dan HashingVectorizer, dengan hasil perbandingan yang diberikan. Selain itu, code ini juga melakukan reduksi dimensi menggunakan LSA (Latent Semantic Analysis) untuk meningkatkan efisiensi clustering dan evaluasi hasil clustering. Hasil clustering dievaluasi dengan berbagai metrik, dan hasilnya disajikan dalam bentuk grafik.