**LAPORAN PRAKTIKUM**

**DATA SCIENCE**

**DATA HIGH DIMENSIONALITY**

**REDUCTION (t-SNE)**



**LAURA VEGAWANI PASARIBU**

**11422009**

**ST Teknologi Rekayasa Perangkat Lunak**

**INSTITUT TEKNOLOGI DEL**

**FAKULTAS VOKASI**

**t-SNE**

t-distributed stochastic neighbord embending(t-SNE) adalah sebuah algoritma dalam masin learning untuk mengurangi jumlah dimensi dataset yang akan digunakan. t-SNE merupakan teknik yang bersifat non-linear yang cocok untuk masalah high-dimensional data kedalam 2 atau 3 dimensi sehingga dapat divisualisasikan menggunakan scatter plot. t-SNE memodelkan objek dengan dimensi yang tinggi berdasarkan tingkat kemiripan data point melalui nilai jarak (distance).

**Mnist Dataset**

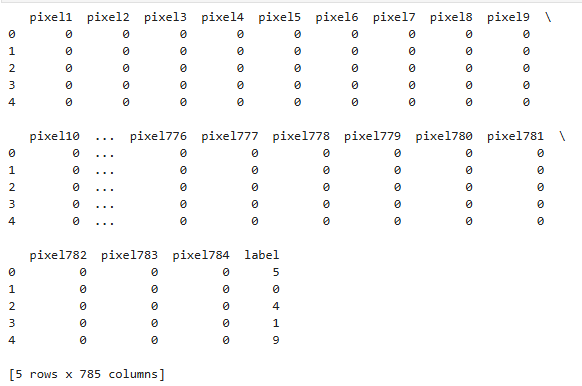
1. Perhatikan details dari dataset mnist yang digunakan seperti dibawah ini :



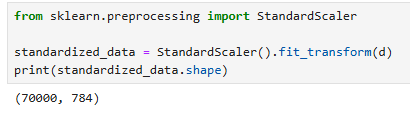
Penjelasan :

Kode ini bertujuan untuk mengunduh, memproses, dan menampilkan dataset MNIST yang berisi gambar digit tulisan tangan. Pada code menggunakan fetch\_openml dari sklearn.dataset untuk mengambil dataset dari OpenMl. Data piksel diubah menjadi bentuk table menggunakan pandas.DataFrame agar lebih mudah diolah.

1. Dari dataset tersebut diketahui jumlah baris dan feature yang digunakan



1. Pada tahap ini dataset dengan jumlah baris 5 dan kolom 785 akan dilakukan Standar scalling. Jelaskan perubahan atau transisi dari step 2 yang menghasilkan (70000, 784)!

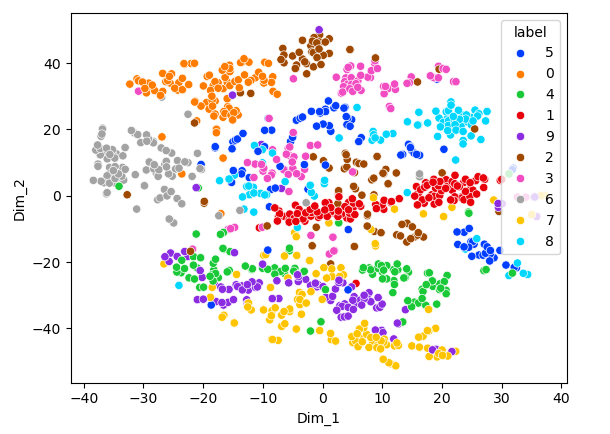


Penjelasan :

Pada step 2 dataset memiliki 785 kolom (784 fitur pikes + 1 kolom label), jumlah baris 70.000, label angka (0-9) masih termasuk dalam dataset namun pada step 3 dataset mengalami standarisasi menggunakan StandardScaler(), yang menesuaikan setiap fitur sehingga memiliki rat-rata 0 dan standar deviasi 1. Label tidak termasuk dalam scaling karena hanya fitur numerik (piksel) yang distandarisasi. Hasilnya array dengan dimensi (70.000, 784), yang dimana hanya fitur piksel yang diproses, tanpa kolom label. Pada step 3 kolom label dihapus sebelum proses standardisasi, karena tidak relevan untuk normalisasi fitur numerik, StandardScaler() hanya bekerja pada fitur numerik (784 kolom piksel), sehingga label harus dipisahkan sebelum transformasi.

1. Perhatikan standarisasi yang digunakan label dan data!



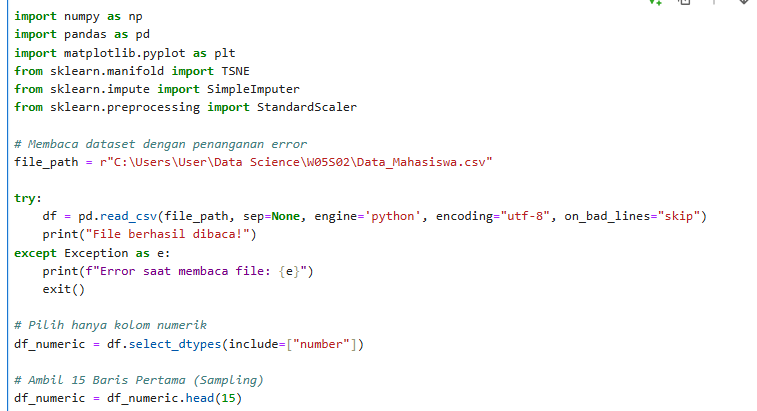


Penjelasan :

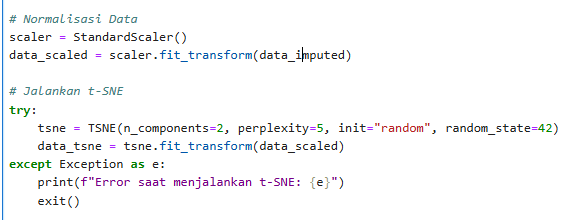
Visualisasi hasil dari proses dimensi reduksi menggunakan algoritma t-SNE (t-Distributed Stochastic Neighbor Embedding) pada data 1000 sampel. t-SNE adalah teknik visualisasi data dimensi tinggi ke dalam ruang berdimensi rendah, biasanya 2D atau 3D, sehingga struktur data dapat diamati dengan lebih baik. Pada visualisasi ini, setiap titik merepresentasikan satu sampel data, dan warna titik menunjukkan label kelas dari sampel tersebut. Dapat dilihat bahwa data terbagi ke dalam beberapa klaster yang berbeda, yang mengindikasikan adanya struktur atau pola tertentu dalam data.

**Tugas**

1. Silahkan melakukan dimensionality reduction dengan dataset DataMahasiswa pada praktikum sebelumnya
2. Jelaskan langkah-langkah dan perubahan data yang terjadi selama proses dimensionality reduction!
3. Silahkan melakukan perhitungan untuk :
4. Ambilah sampling dari dataset tersebut 15 baris pertama

****

1. The t-SNE algorithm finds the similarity measure between pairs of instances in higher and lower dimensional space. After that, it tries to optimize two similarity measures.

****

1. t-SNE models a point being selected as a neighbor of another point in both higher and lower dimensions. It starts by calculating a pairwise similarity between all data points in the highdimensional space using a Gaussian kernel. The points far apart have a lower probability of being picked than the points close together.

Jawab :

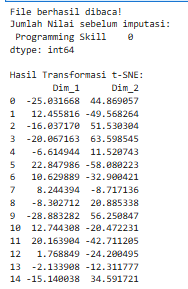
Terjadi secara internal dalam fit\_transform(data\_scaled), di mana Gaussian kernel digunakan untuk menghitung probabilitas bahwa satu titik akan dipilih sebagai tetangga titik lain.

1. Map higher-dimensional data points onto lower-dimensional space while preserving the pairwise similarities.

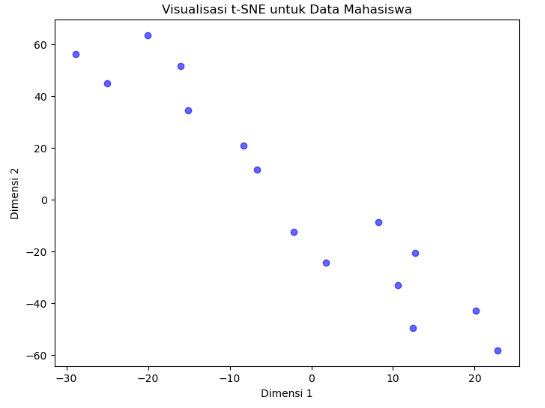
****

1. It is achieved by minimizing the divergence between the original high-dimensional and lowerdimensional probability distribution. The algorithm uses gradient descent to minimize the divergence. The lower-dimensional embedding is optimized to a stable state.

****

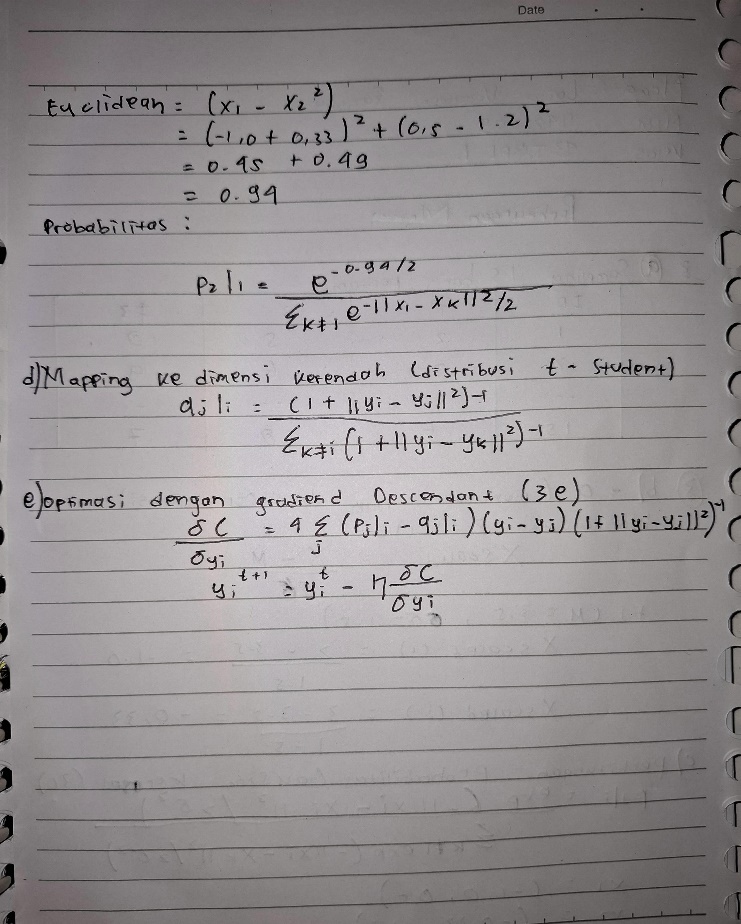
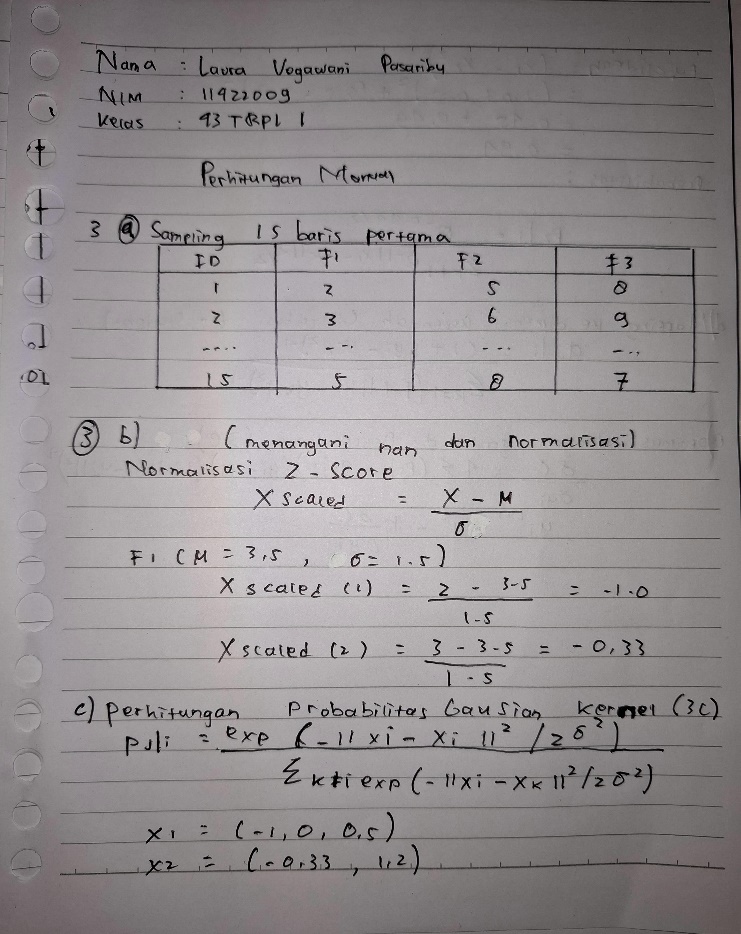
****

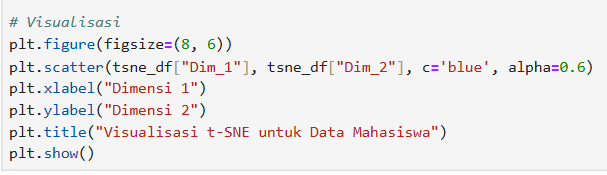
**Output**

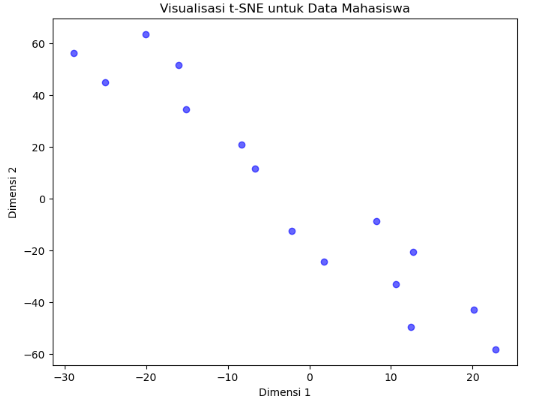
****

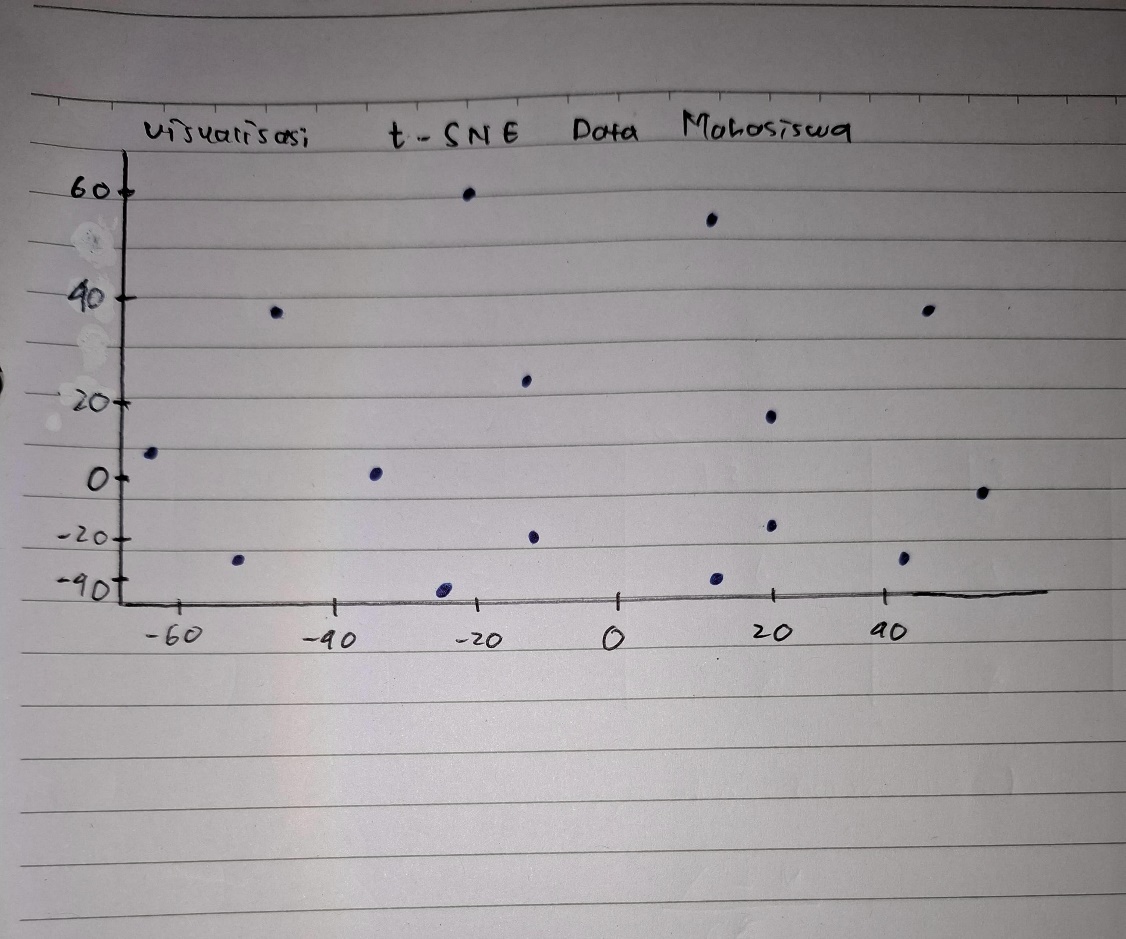
Penerapan t-SNE pada dataset DataMahasiswa berhasil mereduksi dimensi tinggi menjadi dua dimensi untuk memudahkan analisis dan visualisasi. Proses ini diawali dengan pengambilan sampel 15 baris pertama, penanganan nilai NaN dengan imputasi mean, serta normalisasi data menggunakan standarisasi Z-score. t-SNE bekerja dengan menghitung kesamaan antar data di dimensi tinggi menggunakan Gaussian kernel, lalu memetakannya ke dimensi rendah dengan mempertahankan hubungan lokal antar titik melalui t-Student distribution. Proses optimasi dilakukan menggunakan gradient descent untuk menghasilkan representasi yang stabil. Hasil visualisasi menunjukkan pola distribusi data yang lebih jelas, membantu dalam mengidentifikasi struktur tersembunyi dan klaster dalam dataset. Metode ini sangat berguna dalam memahami hubungan antar data yang sulit terlihat dalam dimensi tinggi.

1. Silahkan melakukan perhitungan manual untuk 3a s/d 3e dan visualisasikan dalam bentuk gambar!

****

****

****

****