LAPORAN TUGAS DATA SCIENCE

Dataset: Customer Segmentation Dataset
Nama : Adam Akbarul Dimas
NIM : 20230040207
Kelas : TI23C

1. Pendahuluan

Unsupervised Learning merupakan metode pembelajaran mesin yang digunakan untuk menemukan pola tersembunyi dalam data tanpa label. Salah satu teknik yang sering digunakan adalah Agglomerative Clustering. Pada tugas ini dilakukan analisis pengelompokan pelanggan menggunakan dataset Customer Segmentation.

2. Dataset

Dataset yang digunakan adalah Customer Segmentation Dataset yang berisi karakteristik pelanggan berupa usia (Age), pendapatan tahunan (Annual Income), dan skor pengeluaran (Spending Score). Dataset ini digunakan untuk mengelompokkan pelanggan berdasarkan kemiripan karakteristik.

3. Data Cleaning dan Preprocessing

Tahapan data cleaning dilakukan dengan memeriksa nilai yang hilang (missing values). Hasil pemeriksaan menunjukkan bahwa dataset tidak memiliki missing values. Selanjutnya dilakukan standarisasi data menggunakan StandardScaler untuk menyamakan skala fitur.

4. Analisis Deskriptif dan Visualisasi Awal

Analisis deskriptif dilakukan untuk memahami distribusi data. Visualisasi awal berupa histogram dan analisis korelasi digunakan untuk melihat pola hubungan antar variabel sebelum proses clustering dilakukan.

5. Agglomerative Clustering

Model Agglomerative Clustering dibangun menggunakan empat metode linkage yaitu Single, Complete, Average, dan Ward Linkage. Jumlah cluster ditentukan sebanyak 4 cluster untuk setiap metode agar hasil dapat dibandingkan secara adil.

6. Dendrogram

Dendrogram dibuat menggunakan library scipy.cluster.hierarchy untuk membantu menentukan jumlah cluster yang optimal. Berdasarkan pemotongan (cutting threshold) pada dendrogram, jumlah cluster yang paling optimal adalah 4 cluster karena menunjukkan pemisahan data yang jelas.

7. Evaluasi Cluster

Evaluasi kualitas cluster dilakukan menggunakan Silhouette Score. Hasil evaluasi menunjukkan bahwa Ward Linkage menghasilkan kualitas cluster terbaik karena memiliki nilai silhouette paling tinggi dan cluster yang lebih kompak.

8. Analisis dan Interpretasi Cluster

Cluster 1: Pelanggan dengan usia menengah dan pengeluaran tinggi, berpotensi sebagai target premium. Cluster 2: Pelanggan usia muda dengan pengeluaran sedang. Cluster 3: Pelanggan dengan

pendapatan tinggi namun pengeluaran rendah, berpotensi untuk upselling. Cluster 4: Pelanggan dengan pendapatan dan pengeluaran rendah.

9. Kesimpulan

Berdasarkan hasil analisis, Ward Linkage merupakan metode terbaik dalam kasus ini. Agglomerative Clustering terbukti efektif untuk segmentasi pelanggan dan dapat digunakan sebagai dasar pengambilan keputusan bisnis.