

Nama Dosen : Teguh Iman Hermanto, M.Kom
 Mata Kuliah : Machine Learning 1
 Pembahasan : Hierarchical Clustering
 Pokok Pemb : - Membangun Model Hierarchical Clustering
 - Simulasi Hierarchical Clustering
 - Profiling Hasil Clustering

1. Load Library dan dataset pada file Notebook

```

1 import pandas as pd
2
3 from sklearn.preprocessing import StandardScaler
4 from scipy.cluster.hierarchy import linkage
5 from scipy.cluster.hierarchy import dendrogram
6 from scipy.cluster.hierarchy import cut_tree
7
8 import matplotlib.pyplot as plt
9 import seaborn as sns

```

```

1 # Membaca dataset
2 data = pd.read_csv('dataset_nasabah.csv')

```

`data.head()`

✓ 0.0s

Python

	ID Nasabah	Usia	Jenis Kelamin	Status Merokok	BMI	Jumlah Klaim	Total Biaya Klaim
0	1	58	Pria	Ya	30.7	1	4744854
1	2	26	Pria	Ya	28.2	0	0
2	3	19	Pria	Ya	22.3	4	5780796
3	4	53	Pria	Tidak	22.1	4	22669060
4	5	69	Pria	Ya	30.0	2	11323814

2. Buatlah 5 variasi Exploratory data analysis dan berikan penjelasannya yang terdiri dari

- Pie Plot
- Box plot
- Violin plot
- Scatter plot
- Count plot

3. Buat model Hierarchical Clustering

```

1 # Pra-pemrosesan data
2 # Mengonversi kolom kategorikal menjadi numerik
3 data['Jenis Kelamin'] = data['Jenis Kelamin'].map({'Pria': 0, 'Wanita': 1})
4 data['Status Merokok'] = data['Status Merokok'].map({'Tidak': 0, 'Ya': 1})

```

`data.head()` ✓ 0.0s Python

	ID Nasabah	Usia	Jenis Kelamin	Status Merokok	BMI	Jumlah Klaim	Total Biaya Klaim
0	1	58	0	1	30.7	1	4744854
1	2	26	0	1	28.2	0	0
2	3	19	0	1	22.3	4	5780796
3	4	53	0	0	22.1	4	22669060
4	5	69	0	1	30.0	2	11323814

```

1 # Select features for clustering
2 features = ['Usia', 'Jenis Kelamin', 'Status Merokok', 'BMI', 'Jumlah Klaim', 'Total Biaya Klaim']
3 X = data[features]

```

```

1 # Standardize the data
2 scaler = StandardScaler()
3 X_scaled = scaler.fit_transform(X)

```



```
1 # Create linkage matrix for dendrogram
2 linked = linkage(X_scaled, method='ward')
```



```
1 # Plot dendrogram
2 plt.figure(figsize=(10, 7))
3 plt.title("Dendrogram for Hierarchical Clustering")
4 dendrogram(linked, truncate_mode='lastp', p=30, leaf_rotation=90, leaf_font_size=10)
5 plt.xlabel("Cluster Size")
6 plt.ylabel("Distance")
7 plt.show()
```



```
1 # membuat cluster label
2 cluster_labels = cut_tree(linked, n_clusters=5).reshape(-1, )
3 cluster_labels
```



```
1 # gabungkan cluster label dengan data
2 data['Cluster_Labels'] = cluster_labels
```

`data.head()`

✓ 0.0s

Python

	ID Nasabah	Usia	Jenis Kelamin	Status Merokok	BMI	Jumlah Klaim	Total Biaya Klaim	Cluster_Labels
0	1	58	0	1	30.7	1	4744854	0
1	2	26	0	1	28.2	0	0	0
2	3	19	0	1	22.3	4	5780796	0
3	4	53	0	0	22.1	4	22669060	1
4	5	69	0	1	30.0	2	11323814	0

4. Buatlah profiling masing-masing cluster

- a. Buatlah profiling menggunakan scatterplot, lakukan perbandingan dengan merubah nilai x dan y lalu berikan kesimpulannya

```
1 plt.figure()
2 sns.scatterplot(data=data,
3                 x='Usia',
4                 y='Total Biaya Klaim',
5                 hue='Cluster_Labels',
6                 palette='viridis')
7 plt.title("Hubungan Usia dan Total Biaya Klaim")
8 plt.xlabel("Usia")
9 plt.ylabel("Total Biaya Klaim")
10 plt.show()
```

- b. Buatlah profiling menggunakan boxplot, lakukan perbandingan dengan merubah nilai y lalu berikan kesimpulannya

```
1 sns.boxplot(data=data,
2             x='Cluster_Labels',
3             y='Usia',
4             palette='Set3')
5 plt.title("rata-rata usia berdasarkan cluster")
6 plt.xlabel("Cluster")
7 plt.ylabel("Usia")
8 plt.show()
```