TECHNICAL REPORT

DATA MINING

Rully Meidyta (2206130795)

PROGRAM STUDI MAGISTER MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

DEPOK

2023

Pendahuluan

Clustering adalah teknik dalam data mining untuk mengelompokkan data yang serupa berdasarkan karakteristik yang dimiliki. Pada teknik ini, data dikelompokkan berdasarkan kemiripan fitur dan tidak memperhatikan kelas atau label data. Pada umumnya, clustering digunakan untuk mengelompokkan data ke dalam beberapa kelompok untuk tujuan analisis lebih lanjut.

Pada technical report ini, akan menggunakan 3 teknik clustering yang akan diterapkan pada dataset Breast Cancer menggunakan bahasa pemrograman Python yang tersedia di library scikit-learn. Dataset Breast Cancer adalah dataset yang populer dalam pemrosesan data dan banyak digunakan dalam machine learning. Dataset ini berisi informasi mengenai tumor payudara yang bersifat ganas atau jinak. Dengan menerapkan Teknik clustering K-Means, Agglomerative Clustering dan DBSCAN pada data set Breast Cancer. Dan evaluasi teknik clustering akan dilakukan dengan menggunakan Silhoutte Score.

Metode

Dataset

Dataset yang digunakan adalah dataset Breast Cancer yang disediakan oleh library sklearn. Dataset ini memiliki 569 sampel dan 30 fitur yang merepresentasikan ukuran dan bentuk sel pada gambar biopsy payudara. Label yang ada pada dataset ini adalah 0 dan 1, di mana 0 merepresentasikan tumor jinak dan 1 merepresentasikan tumor ganas.

Library

Pada proses ini, kita memuat semua library yang dibutuhkan dalam proses clustering, seperti seaborn untuk visualisasi data, serta library dari Scikit-learn (sklearn) seperti PCA, StandardScaler, KMeans, AgglomerativeClustering, DBSCAN, dan silhouette_score.

```
import seaborn as sns
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import load_breast_cancer
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA
```

Load Dataset

Proses ini memuat dataset breast cancer yang tersedia di library scikit-learn menggunakan fungsi load_breast_cancer(). Kemudian, kita memisahkan antara atribut dan target dari dataset menggunakan data dan target.

```
cancer = load_breast_cancer()
X = cancer.data
y = cancer.target
```

Preprocessing

Sebelum melakukan clustering, terlebih dahulu dilakukan preprocessing pada dataset. Preprocessing yang dilakukan meliputi:

1. Melakukan standardisasi data menggunakan StandardScaler() dari library scikit-learn. Tujuan dari standardisasi ini adalah untuk mengubah skala data sehingga memiliki mean=0 dan variance=1. Hal ini dilakukan untuk memudahkan proses clustering nantinya.

2. Setelah data di standardisasi, kita melakukan reduksi dimensi dengan Principal Component Analysis (PCA) menggunakan PCA() dari library scikit-learn. PCA digunakan untuk mengubah data menjadi representasi baru yang lebih rendah dimensi namun tetap menjaga informasi utama dari data.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

Clustering

Pada technical report ini, akan digunakan 3 teknik clustering, yaitu KMeans, Agglomerative Clustering, dan DBSCAN.

K-Means

KMeans adalah teknik clustering yang mengelompokkan data berdasarkan jarak ke centroid tertentu. Pada teknik ini, jumlah kelompok harus ditentukan sebelumnya. Pada eksperimen ini, KMeans clustering dengan 2 cluster menggunakan KMeans() dari library scikit-learn. fit_predict() digunakan untuk memprediksi cluster dari setiap data berdasarkan hasil clustering yang telah dilakukan.

```
kmeans = KMeans(n_clusters=2)
kmeans.fit(X_pca)
y_kmeans = kmeans.fit_predict(X_pca)

sns.scatterplot(x=X_pca[:,0], y=X_pca[:,1], hue=y_kmeans, palette='viridis')
```

Agglomerative Clustering

Agglomerative Clustering adalah teknik clustering yang mengelompokkan data secara hierarki dengan cara menggabungkan dua kelompok yang paling mirip. Pada teknik ini melakukan Agglomerative clustering dengan 2 cluster menggunakan AgglomerativeClustering().

```
agglo = AgglomerativeClustering(n_clusters=2, linkage='ward')
y_agglo = agglo.fit_predict(X_pca)

sns.scatterplot(x=X_pca[:,0], y=X_pca[:,1], hue=y_agglo, palette='viridis')
```

DBSCAN

DBSCAN adalah teknik clustering yang mengelompokkan data dengan cara menemukan kelompok yang memiliki kepadatan tertentu. Pada teknik ini melakukan DBSCAN clustering dengan epsilon=2 dan minimal samples=2 menggunakan DBSCAN() dari library scikit-learn. fit_predict() digunakan untuk memprediksi cluster dari setiap data berdasarkan hasil clustering yang telah dilakukan.

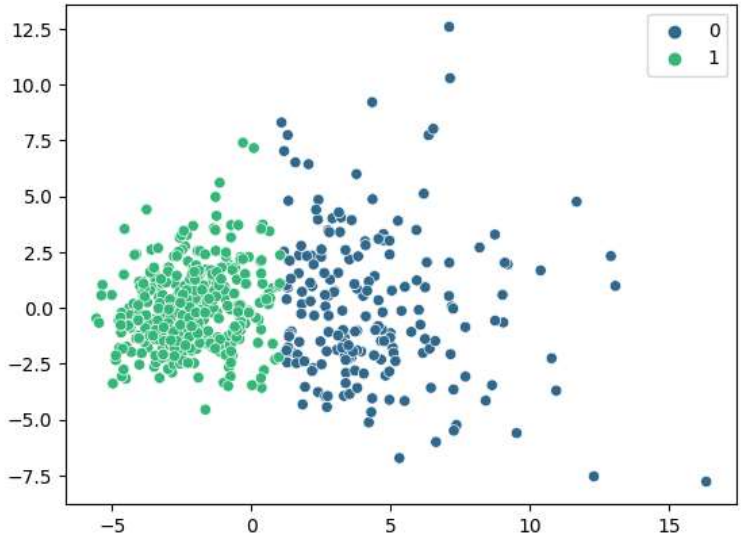
```
dbscan = DBSCAN(eps=2, min_samples=2)
y_dbscan = dbscan.fit_predict(X_pca)

sns.scatterplot(x=X_pca[:,0], y=X_pca[:,1], hue=y_dbscan, palette='viridis')
```

Hasil

Setelah melakukan clustering, hasil yang didapatkan dapat divisualisasikan menggunakan scatter plot. Pada scatter plot, titik-titik yang memiliki warna yang sama merupakan data yang termasuk ke dalam kelompok yang sama.

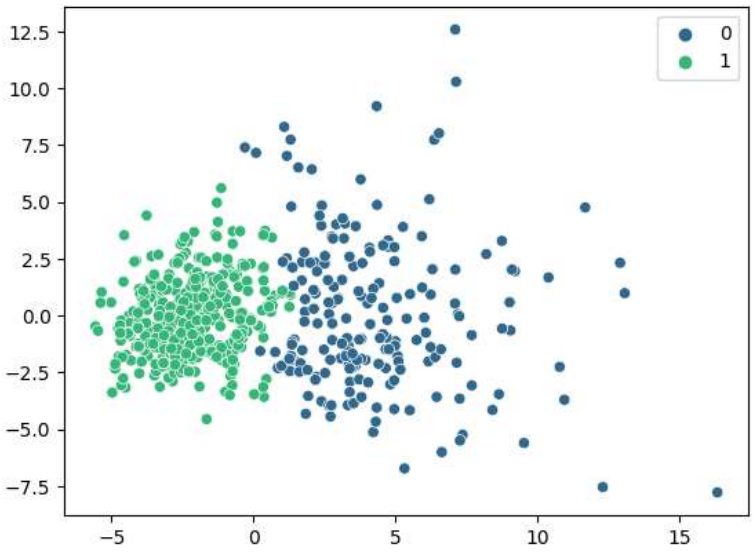
K-Means

Titik-titik pada scatterplot mewakili sampel data. Setiap titik memiliki koordinat pada sumbu x dan y yang merepresentasikan nilai dari dua fitur utama setelah dilakukan PCA. Warna pada setiap titik menunjukkan keanggotaan klaster. Setiap klaster diberikan warna yang berbeda pada scatterplot. Sebagai contoh, dalam scatterplot tersebut, kita dapat melihat bahwa ada dua klaster yang diberikan warna biru dan hijau.Warna pada setiap titik menunjukkan keanggotaan klaster dari sampel data tersebut.Jika kita melihat titik-titik yang berwarna sama, kita dapat mengasumsikan bahwa mereka memiliki karakteristik yang serupa dan tergabung dalam satu klaster.



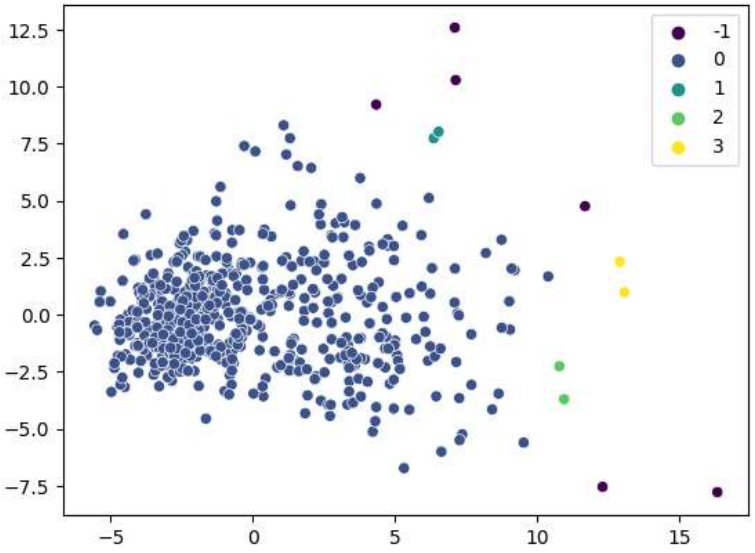
Agglomerative Clustering

Sama seperti pada k-means clustering, scatter plot pada agglomerative clustering di mana setiap titik memiliki koordinat pada sumbu x dan y yang merepresentasikan nilai dari dua fitur utama setelah dilakukan PCA. Dalam scatterplot tersebut, kita dapat melihat bahwa ada dua klaster yang diberikan warna biru dan hijau. Titik-titik yang berwarna sama, kita asumsikan bahwa mereka memiliki karakteristik yang serupa dan tergabung dalam satu klaster.



DBSCAN

Setiap titik pada scatterplot merepresentasikan satu sampel data, dan koordinat titik tersebut menunjukkan nilai dari dua fitur utama untuk sampel tersebut.Terdapat beberapa klaster yang diberikan warna berbeda. Dalam scatterplot tersebut, kita dapat melihat bahwa ada lima cluster yang diberikan warna berbeda. DBSCAN menggunakan density-based clustering, sehingga klaster-klaster dibentuk dari titik-titik yang memiliki kepadatan yang tinggi. Oleh karena itu, pada scatterplot, kita dapat melihat klaster-klaster yang memiliki titik-titik yang saling berdekatan dan memiliki kepadatan yang tinggi.



Evaluasi

Untuk mengevaluasi hasil clustering, dapat dilakukan dengan menggunakan metrik seperti Silhouette Score atau Adjusted Rand Index (ARI). Pada eksperimen ini, akan digunakan metrik Silhouette Score.

Hasil evaluasi dengan Silhouette Score:

- K-Means Silhouette Score: 0.508469019061225
- Agglomerative Clustering Silhouette Score: 0.5046397728338283
- DBSCAN Silhouette Score: 0.44201956139147686

```
kmeans_silhouette = silhouette_score(X_pca, kmeans.labels_)
agg_silhouette = silhouette_score(X_pca, agglo.labels_)
dbscan_silhouette = silhouette_score(X_pca, dbscan.labels_)

print("Hasil evaluasi dengan Silhouette Score:")
print("KMeans Silhouette Score:", kmeans_silhouette)
print("Agglomerative Clustering Silhouette Score:", agg_silhouette)
print("DBSCAN Silhouette Score:", dbscan_silhouette)
```

```
Hasil evaluasi dengan Silhouette Score:
KMeans Silhouette Score: 0.5084690190671225
Agglomerative Clustering Silhouette Score: 0.5046397728338283
DBSCAN Silhouette Score: 0.44201956139147686
```

Dari hasil evaluasi, dapat dilihat bahwa K-Means memberikan nilai Silhouette Score yang paling tinggi dibandingkan dengan Agglomerative Clustering dan DBSCAN.

Kesimpulan

Data yang digunakan dalam code tersebut adalah dataset Breast Cancer dari library sklearn. Sebelum melakukan clustering, data di-scaling terlebih dahulu menggunakan StandardScaler, dan di-reduksi dimensi menjadi dua fitur utama menggunakan PCA. Clustering dengan KMeans menghasilkan dua klaster yang ditampilkan dalam visualisasi scatterplot pertama, sedangkan Agglomerative Clustering menghasilkan dua klaster yang ditampilkan dalam visualisasi scatterplot kedua. DBSCAN menghasilkan lima klaster yang ditampilkan dalam visualisasi scatterplot ketiga. Setelah melakukan clustering, evaluasi dilakukan dengan menggunakan Silhouette Score. Hasil evaluasi menunjukkan bahwa KMeans memiliki Silhouette Score tertinggi dibandingkan dengan Agglomerative Clustering dan DBSCAN. Bahwa dataset Breast Cancer dapat dikelompokkan menjadi dua cluster menggunakan KMeans dengan Silhouette Score yang tinggi. Namun, klastering dengan Agglomerative Clustering dan DBSCAN juga menghasilkan hasil yang cukup baik.