Laporan Data Mining

Tim :

- Fadlil Aliffiana Hasan (A11.2021.13557)
- Mario Ignatius Surya Nugraha (A11.2021.13361)
- Natanael James Santoso (A11.2021.13533)

UNIVERSITAS DIAN NUSWANTORO

Januari 2024

Laporan Analisis Data

1. Dataset
   Tampilan dari beberapa data :

	Nama	Alamat	Kelamin	Umur
0	Anggun Kholifatul Khasanah	Karangrejo	P	11.0
1	Anggun Afifah Nurulaini	Pandanharum	P	3.0
2	Faizul Naam Hafizhan	Karangrejo	L	2.0
3	Murni	Gabus	P	59.0
4	Salsabila Aufa Puji	Kalipang	P	11.0
..	...	...	...	...
84	Faleshia Zoya	Bendoharjo	P	5.0
85	Hafidzah Prominensa	Gabus	P	10.0
86	Oktaviana Galih Pratiwi	Pandanharum	P	21.0
87	Khaliza Shesha Safitri	Bendoharjo	P	1.0
88	Alfiatun	Pandanharum	P	2.0

[89 rows x 4 columns]

   Dari data tersebut merupakan data dari Puskemas GABUS 2, data tersebut berisi data pasien yang pernah mengalami penyakit demam berdarah. Dari data tersebut akan di kelompokan berdasarkan dengan jenis kelamin dan umur. Dari pengelompokan tersebut nantinya akan dibuat visualisasi untuk lebih memahami isi data. Dataset ini merupakan dataset private.
   Atribut :
   - Nama
     Berisi nama dari pasien
   - Alamat
     Alamat atau tempat tinggal pasien
   - Kelamin
     Jenis kelamin pasien. Laki-laki atau Perempuan.
   - Umur
     Umur pasien (Atribut ini berupa nominal).

2. Permasalahan dan Tujuan
   Dari data di atas, ada beberapa pasien yang terkena penyakit demam berdarah, dari sekian banyaknya pasien yang menderita DBD, dimisalkan ada dokter yang ingin mengetahui berapa banyak jumlah pasien pria dan wanita yang juga berdasarkan umur dari pasien tersebut, bagaimana kira-kira dokter tersebut mengetahui hal tersebut ?
   Tujuan dari ekseperimen ini untuk menentukan hasil cluster dari 2 atribut data yaitu ‘Umur’ dan ‘Kelamin’ menggunakan hierarchical clustering yang akan diukur juga bagaimana hasil cluster tersebut.

3. Model dan Alur Tahapan Eksperimen
    - Langkah awal yaitu import beberapa library yang akan di gunakan yaitu seperti NumPy, Matplotlib, Pandas.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

    - Proses selanjutnya yaitu membaca data yang akan digunakan, yang sudah saya berikan contohnya tadi, file data tersebut berupa file csv dengan nama dbd.

```
dataset = pd.read_csv('dbd.csv', delimiter=";", nrows=89)
```

    - Dikarenakan ada beberapa data yang masih kosong, maka kita harus mengisi data yang kosong tersebut, di sini saya menggunakan cara otomatis.

```
dataset['Kelamin'].fillna(dataset['Kelamin'].mode().iloc[0], inplace=True)
dataset['Umur'].fillna(dataset['Umur'].median(), inplace=True)
```

    - Mengubah nilai Kelamin menjadi nilai numerik.

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
dataset['Kelamin'] = label_encoder.fit_transform(dataset['Kelamin'])
```

```
print(dataset)
```

```
                        Nama       Alamat  Kelamin  Umur
0   Anggun Kholifatul Khasanah   Karangrejo        1    10
1      Anggun Afifah Nurulaini  Pandanharum        1     2
2         Faizul Naam Hafizhan   Karangrejo        0     1
3                        Murni        Gabus        1    25
4          Salsabila Aufa Puji     Kalipang        1    10
..                         ...          ...      ...   ...
84               Faleshia Zoya   Bendoharjo        1     4
85         Hafidzah Prominensa        Gabus        1     9
86     Oktaviana Galih Pratiwi  Pandanharum        1    19
87      Khaliza Shesha Safitri   Bendoharjo        1     0
88                    Alfiatun  Pandanharum        1     1

[89 rows x 4 columns]
```

    - Menerapkan Hierarchical Clustering dengan total cluster 2 pada data 'Umur' dan 'Kelamin'.

```
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.cluster import AgglomerativeClustering

linkage_matrix = linkage(dataset[X], method='ward')
agglomerative = AgglomerativeClustering(n_clusters=2, linkage='ward')
dataset['cluster'] = agglomerative.fit_predict(dataset[X].values)
```

```
print(dataset['cluster'])
```

```
0     0
1     0
2     0
3     1
4     0
     ..
84    0
85    0
86    1
87    0
88    0
Name: cluster, Length: 89, dtype: int64
```

- Menambahkan kolom 'cluster' ke dalam dataset berdasarkan hasil clustering.

```
majority_cluster = dataset['cluster'].mode()[0]
majority_cluster_data = dataset[dataset['cluster'] == majority_cluster]
```

```
print(majority_cluster_data)
```

	Nama	Alamat	Kelamin	Umur	cluster
0	Anggun Kholifatul Khasanah	Karangrejo	1	10	0
1	Anggun Afifah Nurulaini	Pandanharum	1	2	0
2	Faizul Naam Hafizhan	Karangrejo	0	1	0
4	Salsabila Aufa Puji	Kalipang	1	10	0
5	Mirza	Gabus	0	7	0
..	...	...	...	...	...
83	Rizqia Putri Ramadhani	Tunggulrejo	1	2	0
84	Faleshia Zoya	Bendoharjo	1	4	0
85	Hafidzah Prominensa	Gabus	1	9	0
87	Khaliza Shesha Safitri	Bendoharjo	1	0	0
88	Alfiatun	Pandanharum	1	1	0

[65 rows x 5 columns]

- Menunjukkan presentase berapa banyak pria atau wanita yang terkena penyakit DBD. 1 (Perempuan) dan 0 (Laki-laki) .

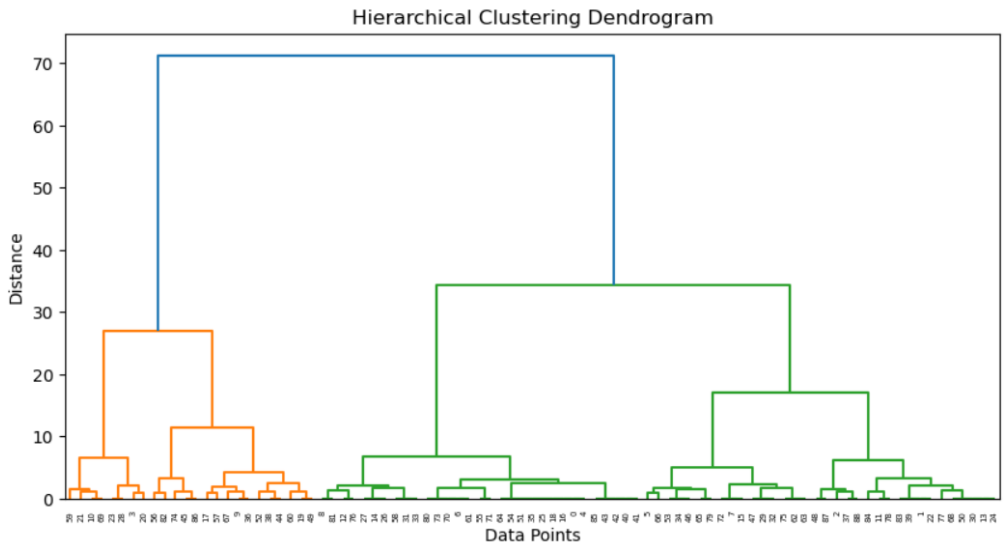
```
gender_counts = majority_cluster_data['Kelamin'].value_counts()
total_samples = len(majority_cluster_data)
```

```
for gender, count in gender_counts.items():
    percentage = (count / total_samples) * 100
    print(f"{gender}: {percentage:.2f}%")
```

```
1: 69.23%
0: 30.77%
```

- Menampilkan model dendrogram hasil Hierarchical Clustering.

```
plt.figure(figsize=(10, 5))
dendrogram(linkage_matrix)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Data Points')
plt.ylabel('Distance')
plt.show()
```



- Menghitung score Davies Bouldin (Semakin rendah maka hasil dari clustering semakin baik).

```
from sklearn.metrics import davies_bouldin_score

davies_bouldin = davies_bouldin_score(dataset[X], dataset['cluster'])
print(f"Score Davies Bouldin : {davies_bouldin:.4f}")
```

```
Score Davies Bouldin : 0.1441
```

4. Kesimpulan dan Rekomendasi
   Dari hasil cluster yang sudah di buat maka bisa di simpulkan dari data, ada banyaknya jumlah Perempuan yang menderita DBD daripada Pria, dan juga hasil cluster yang dihitung dengan Davies Bouldin yang dimana menghasilkan hasil yang baik, karena mendekati 0. Untuk Rekomendasi, sebaiknya dari segi data, mungkin bisa di buat menjadi lebih lengkap lagi, dan bisa menggunakan algoritma clustering atau bahkan algoritma klasifikasi jika ada data penyakit lainnya.