Nama : Rika Syahriani

NPM : 2206130782

Mid Term : Laporan Data Mining

Option 2 : Data Clustering

1. Pendahuluan

Kanker adalah pertumbuhan sel yang tidak normal dimana sel menyerang sel normal di jaringan sekitarnya yang disebabkan oleh perubahan mutasi pada gen dalam sel. Salah satu kanker yang banyak ditemui terutama dikalangan wanita yaitu kanker payudara. Kanker payudara adalah kanker yang tumbuh dan berkembang pada sel-sel jaringan disekitar payudara.

Data mining adalah metode analisis data yang telah berhasil dibanyak bidang salah satunya adalah *clustering*. Metode K-Means *Clustering* membagi data menjadi cluster/kelompok sehingga data dengan karakteristik yang sama dikelompokkan ke dalam cluster yang sama dan data dengan karakteristik berbeda ke dalam cluster lain. *Clustering* bertujuan untuk meminimalkan variasi dalam cluster dan memaksimalkan variasi antar cluster (Agusta, 2007). Pada laporan ini, penulis memilih pendekatan *K-Means*, DBSCAN, dan *Agglomerative*.

2. Tinjauan Pustaka

Data mining adalah suatu proses untuk mengambil data yang besar, beraneka ragam dan kompleks dan meghasilkan informasi yang berguna dan dapat dimengerti. Konsep-konsep dalam data mining seperti *pre-processing data*, pengelompokan data (*clustering*), klasifikasi data dan asosiasi aturan (*association rules*) (Witten dkk, 2016).

Clustering adalah metode untuk mengelompokkan beberapa objek yang mirip ke dalam kategori. *Cluster* adalah kumpulan data yang mirip satu sama lain dan berbeda dengan data di cluster lain. Tujuan dari *clustering* adalah untuk memaksimalkan kesamaan data pada cluster yang sama dan meminimalkan kemiripannya dengan data pada cluster lain (Larose, 2005).

K-Means merupakan salah satu algoritma *clustering* yang populer bertujuan untuk membagi data menjadi beberapa kelompok. Algoritma k-means, centroid dihitung sebagai rata-rata dari semua objek yang termasuk dalam kelompok tersebut dan bekerja dengan cara melakukan iterasi beberapa tahap yaitu tahap assignment, update dan konvergen (Han dkk, 2011).

DBSCAN adalah algoritma pengelompokan berbasis kepadatan. Algoritme memperluas area padat menjadi *cluster* dan menentukan *cluster* tidak teratur dengan

noise dari database spasial. DBSCAN memiliki 2 parameter yaitu Eps (radius maksimum dari neighborhood) dan MinPts (jumlah minimum titik dalam Eps dari satu titik).

Agglomerative hierarchical clustering adalah metode pengelompokan yang dimulai dengan setiap objek sebagai cluster terpisah dan kemudian secara iteratif menggabungkan dua cluster yang paling mirip hingga hanya tersisa satu *cluster*. *Clustering* dilakukan berdasarkan jarak antara dua *cluster*. *Agglomerative hierarchical clustering* dibagi menjadi dua jenis, yaitu single linkage dan complete linkage, tergantung pada cara pengukuran jarak antar kelompok (Han dkk, 2011).

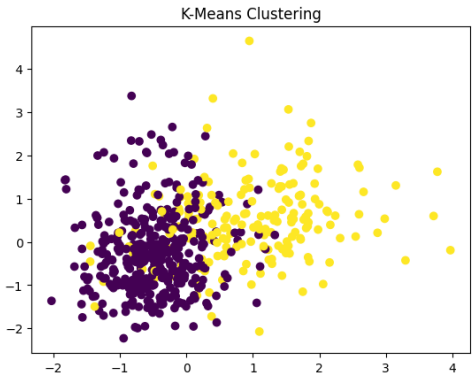
3. Hasil dan Analisis

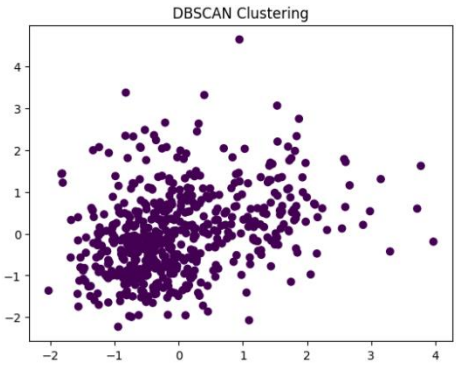
Berdasarkan hasil analisis clustering pada dataset kanker payudara penulis akan menjelaskan ketiga pendekatan yang dilakukan yaitu pendekatan *K-Means*, DBSCAN, dan *Agglomerative.* Langkah-langkah yang dilakukan sebagai berikut:

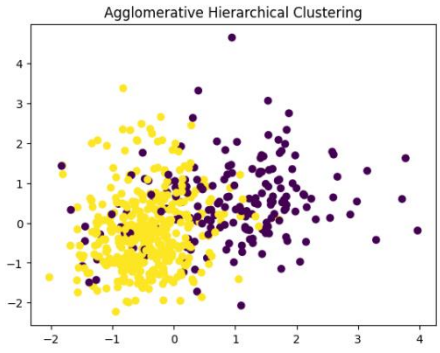
1. Mengimpor beberapa library yang diperlukan seperti *pandas, numpy, sklearn.cluster, sklearn.preprocessing, sklearn.datasets,* dan *matplotlib.pyplot*.
2. Memuat dataset kanker payudara menggunakan fungsi `load_breast_cancer` dari modul `*sklearn.datasets*` dan menyimpannya dalam variabel `data`.
3. Menormalkan data menggunakan StandardScaler dari *sklearn.preprocessing*. Setelah data dinormalisasi, kita menerapkan tiga algoritma pengelompokan (clustering) yaitu *K-Means*, DBSCAN, dan *Agglomerative Hierarchical Clustering*.
4. Pada pengelompokan K-Means, *n_clusters* diinisialisasi dengan nilai 2 karena dataset memiliki dua label kelas (*malignant* dan *benign*). Kemudian, memprediksi label kelas menggunakan *fit predict* dan menyimpannya dalam variable *kmeans labels*.
5. Pada pengelompokan DBSCAN, menginisialisasi *eps* dengan nilai 1 dan *min samples* dengan nilai 5. Nilai *eps* menentukan jarak maksimum antara dua sampel agar dianggap sebagai bagian dari satu cluster. Nilai *min samples* menentukan jumlah minimum sampel untuk membentuk sebuah cluster. Setelah itu memprediksi label kelas menggunakan *fit predict* dan menyimpannya dalam variabel DBSCAN labels.
6. Pada pengelompokan *Agglomerative Hierarchical*, menginisialisasi *n clusters* dengan nilai 2. Kemudian, kita memprediksi label kelas menggunakan *fit predict* dan menyimpannya dalam variabel *agglo labels*.

7. Menggambarkan hasil pengelompokan menggunakan *scatter plots*. Fungsi *plt.scatter* digunakan untuk menggambar *scatter plot* dengan sumbu *x* diisi dengan nilai-nilai pada kolom *0* dan sumbu *y* diisi dengan nilai-nilai pada kolom 1. Argumen c diisi dengan label kelas hasil pengelompokan. Setiap label kelas akan ditampilkan dengan warna yang berbeda.

Output plot berguna untuk memvisualisasikan bagaimana setiap model clustering berhasil memisahkan dan mengelompokkan data, serta memberikan pemahaman visual yang lebih mudah tentang pola dalam data. Berikut ini adalah hasil visualisasi bahasa pemrograman *Python* (Google Colab):







setiap gambar grafik menunjukkan titik-titik data yang diplot dengan sumbu x dan y, dan diwarnai berdasarkan cluster yang telah dibentuk oleh masing-masing model. Setiap cluster diberi warna yang berbeda.

Pada gambar grafik pertama diatas adalah hasil dari menggunakan model *K-Means*. Dimana pada titik-titik pada gambar grafik tersebut warnanya sesuai dengan kelompok cluster yang dihasilkan oleh model *K-Means*, setiap warna yang berbeda mewakili kelompok cluster.

Pada gambar grafik kedua adalah hasil menggunakan model DBSCAN. Gambar grafik menunjukkan titik-titik yang terletak diluar cluster ditandai sebagai noise dengan warna hitam, sedangkan titik-titik yang dikelompokkan diwarnai sesuai dengan kelompok cluster yang dihasilkan oleh model DBSCAN.

Pada gambar grafik ketiga adalah hasil menggunakan model *Agglomerative hierarchical clustering,* titik-titik diwarnai sesuai dengan kelompok cluster.

4. Kesimpulan

Dalam melakukan analisis data, ketiga model clustering yang digunakan adalah K-Means, DBSCAN dan Agglomerative. Hasil dari ketiga model clustering ditampilkan dalam bentuk scatter plot dengan setiap cluster diberi warna yang berbeda. Dengan visualisasi ini, dapat dilihat bagaimana data telah dikelompokkan oleh setiap model clustering.

Referensi

Agusta, Y. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika, 47- 60

Han, J., Kamber, M., & Pei, J. 2011. Data Mining : Concepts and techiniques.Elsevier. Hermawati. F.A. 2013. Data Mining. Yogyakarta: ANDI.

Larose D, T., 2005, Discovering knowledge in data : an introduction to data mining, Jhon Wiley & Sons Inc.

Witten, I. H., Frank, E., & Hall, M.A. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed). Morgan Kaufmaan Publishers.