Laporan Ujian Tengah Semester *Data Mining*

Dosen: Risman Adnan, M.Si.

Disusun oleh: Christiani Turnip (2206130694)

1. Pendahuluan

Kanker adalah suatu penyakit yang ditandai dengan tumbuhnya sel abnormal di dalam tubuh dimana sel tersebut dapat menyerang sel normal di jaringan sekitarnya. Kanker payudara merupakan salah satu kanker yang paling umum terjadi terutama pada wanita. Menurut *World Health Organization* (WHO), ada sekitar 2,3 juta kasus kanker payudara dan sekitar 685.000 kematian akibat kanker payudara setiap tahunnya. Kanker payudara terjadi ketika sel abnormal tumbuh di jaringan payudara membentuk benjolan yang kemudian dapat menyebar ke bagian tubuh lainnya melalui aliran darah. Faktor risiko penyebab terjadinya kanker payudara meliputi usia, riwayat keluarga, gaya hidup, terapi hormonal, dan faktor lingkungan.

Data mining adalah tahapan pada *knowlegde discovery in database* (KDD) yang mempunyai teknik analisis data berjumlah besar dan kompleks sehingga menghasilkan *output* berbentuk pola dari data tersebut. *Clustering* adalah proses pengelompokan data menjadi beberapa *cluster* guna mengidentifikasi pola data di setiap *cluster*. Terdapat beberapa pendekatan yang digunakan dalam metode *clustering*. Pada laporan ini, penulis memilih pendekatan metode *K-Means*, DBSCAN, dan *Agglomerative*.

2. Tinjauan Pustaka

2.1 *Data mining*

Data mining adalah analisis pemeriksaan kumpulan data untuk menemukan hubungan yang tidak terduga dan meringkas data dengan cara berbeda yang dapat dipahami dan berguna bagi pemilik data. *Data mining* adalah bidang multidisiplin yang menggabungkan pembelajaran mesin, pengenalan pola, statistik, basis data, dan teknik visualisasi untuk memecahkan masalah pengambilan informasi dari basis data besar (Han et all., 2012). Lebih lanjut, *data mining* adalah informasi yang disembunyikan dalam basis data, diproses untuk menemukan pola dan teknik statistik matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi dari basis data (Prasetyo, 2014).

2.2 *Clustering*

Clustering adalah proses mengatur objek data ke dalam serangkaian kategori terkait, yang disebut *cluster* (Yedla, 2010). *Clustering* merupakan salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok-kelompok objek dengan karakteristik yang sama dalam data yang cukup besar (Deka dkk., 2014). *Clustering* bertujuan untuk memaksimalkan kesamaan data pada *cluster* yang sama dan meminimalkan kemiripannya dengan data pada *cluster* lain. *Clustering* memungkinkan kita untuk mengklasifikasikan, menemukan pola distribusi umum, dan menemukan hubungan antar atribut data. *Data mining* berfokus pada menemukan metode untuk kelompok *database* besar secara efisien dan efektif. Beberapa persyaratan *clustering* pada *data mining* meliputi skalabilitas, kemampuan untuk menangani

berbagai jenis atribut, kemampuan untuk menangani dimensi besar, kemampuan untuk menangani data yang *noise*, dan kemampuan untuk menerjemahkan dengan mudah.

2.3 *K-Means*

K-Means clustering merupakan salah satu teknik *clustering* yang populer dan sering digunakan dalam analisis data. *K-means clustering adalah* algoritma pengelompokan yang menentukan kelompok data terkait dengan menghitung jarak *Euclidean* antara setiap data dan pusat *cluster* yang ditentukan secara acak (Kouzani, 2017). Langkah-langkah algoritma *K-Means clustering* adalah sebagai berikut.

1. Tentukan jumlah *cluster* (k) yang diinginkan.
2. Pilih *centroid* secara acak untuk setiap *cluster*.
3. Hitung jarak antara setiap data dengan setiap *centroid*.
4. Tetapkan data ke kelompok yang memiliki *centroid* terdekat.
5. Hitung kembali posisi *centroid* untuk setiap kelompok dengan menggunakan rata-rata.
6. Ulangi langkah 3 sampai 5 hingga tidak ada lagi perubahan kelompok.

2.4 DBSCAN

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) adalah algoritma clustering yang mengelompokkan data berdasarkan kerapatan dan jarak antar data. Algoritma ini dapat menemukan *cluster* yang tidak beraturan dan mengecualikan area dngan kepadatan data yang rendah sebagai *noise* (Liu et al., 2017). DBSCAN menggunakan dua parameter utama, yaitu epsilon (ε) dan jumlah minimum titik yang dibutuhkan untuk membentuk *cluster* (minPts). Suatu titik didefiniskan sebagai inti jika setidaknya ada minPts titik lain dalam radius epsilon (Ester et al., 1996). Selain itu, suatu titik dinyatakan sebagai titik batas jika titik inti dalam dalam radius epsilon tetapi tidak memenuhi jumlah titik dalam minPts radius epsilon. Titik yang tidak termasuk dalam kelas inti atau batas disebut *noise*.

DBSCAN *clustering* memiliki beberapa kelebihan dan kekurangan. Kelebihan dari algoritma DBSCAN, antara lain:

1. Dapat menangani data yang memiliki kepadatan yang berbeda dan tidak teratur.
2. Dapat menemukan *cluster* yang memiliki bentuk yang tidak teratur.
3. Tidak memerlukan jumlah *cluster* yang diinginkan sebagai *input*.

Sedangkan kelemahan dari algoritma DBSCAN, antara lain:

1. Sensitif terhadap parameter epsilon dan minPts.
2. Tidak efektif dalam mengelompokkan data yang berada di dalam *cluster* yang berdekatan.

2.5 *Agglomerative*

Agglomerative clustering adalah metode pengelompokan hierarkis yang dimulai dengan setiap titik data sebagai satu *cluster* dan kemudian menggabungkan *cluster* berdasarkan jarak antar titik data. Pada setiap langkah, dua *cluster* dengan jarak terpendek digabungkan sehingga semua titik data digabungkan menjadi satu *cluster*. *Agglomerative clustering* terbagi menjadi dua jenis yaitu *single-linkage clustering* dan *complete-linkage clustering* (Jain et al., 1988). *Single-linkage clustering* mengukur jarak antara dua *cluster* sebagai jarak terdekat antara titik data dalam satu *cluster* dan titik data dalam *cluster* lain. Sedangkan *Complete-linkage clustering* mengukur jarak antara dua *cluster* sebagai jarak terjauh antara satu titik data di satu *cluster* dengan titik data di *cluster* lainnya. Ada juga jenis *agglomerative clustering* lainnya

yaitu *Average-linkage clustering* yang mengukur jarak antara dua *cluster* sebagai jarak rata-rata antara semua titik data dalam satu *cluster* dan semua titik data dalam *cluster* lainnya.

Agglomerative clustering memiliki beberapa kelebihan dan kekurangan. Kelebihan dari algoritma *Agglomerative clustering*, antara lain:

1. Tidak memerlukan jumlah *cluster* sebagai *input*.
2. Cocok untuk data yang memiliki struktur hierarkis atau terdi dari *subcluster*.

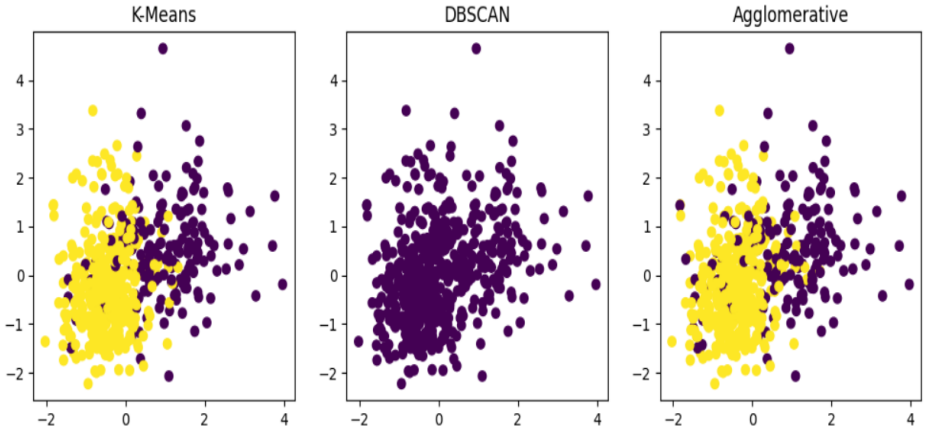
Sedangkan kelemahan dari algoritma DBSCAN, antara lain:

1. Sensitif terhadap jarak antar titik data dan pemilihan metrik jarak.
2. Tidak efisien untuk data yang memiliki jumlah titik yang sangat banyak.

3. Hasil dan Analisis

Pada laporan ini dilakukan analisis clustering pada dataset kanker payudara dengan menggunakan tiga metode clustering yaitu *K-Means*, DBSCAN dan *Agglomerative*. Dataset yang digunakan berasal dari perpustakaan *Scikit-Learn* dari 30 fitur yang digunakan untuk memprediksi apakah suatu tumor jinak atau ganas. Implementasi dari tiga model *clustering* yaitu *K-Means*, DBSCAN dan *Agglomerative clustering* untuk mengelompokkan data kanker payudara pada dataset Scikit. Pertama, dataset kanker payudara *Scikit* dimuat dan diproses menggunakan *StandardScaler* untuk mengubah data menjadi *z-score* sehingga memiliki rata-rata 0 dan varians 1. Kemudian dilakukan *clustering* menggunakan tiga model, yaitu K-Means dengan jumlah *cluster* 2, DBSCAN dengan epsilon 0,5 dan sampel minimal 5, dan *Agglomerative clustering* dengan 2 *cluster*.

Visualisasi hasil clustering kemudian dilakukan menggunakan *scatter plot* dengan tiga *subplot* berbeda yang masing-masing merepresentasikan model *K-Means*, DBSCAN dan *Agglomerative clustering*. Setiap titik pada *scatter plot* merepresentasikan data dari kumpulan data kanker payudara yang dikelompokkan dalam *cluster* yang berbeda. Warna setiap titik merepresentasikan *cluster* tertentu di setiap model *cluster*. Dari hasil visualisasi terlihat perbedaan hasil *clustering* dari ketiga model yang dapat digunakan untuk analisis data kanker payudara lebih lanjut. Di bawah ini hasil visualisasi menggunakan bahasa pemrograman *Python* (*Google Collab*).



Dengan memvisualisasikan hasil *clustering*, terlihat bahwa ketiga metode *clustering* tersebut menghasilkan pemisahan yang relatif baik antara kelas jinak dan ganas. Namun, terdapat perbedaan jumlah *cluster* yang dihasilkan oleh masing-masing metode. *K-Means* dan *Agglomerative* menghasilkan dua *cluster* yang secara jelas memisahkan kelas jinak dan ganas, sedangkan DBSCAN menghasilkan banyak *cluster* tanpa perbedaan yang jelas antara kedua kelas tersebut.

Sehingga, *K-means* dan *agglomerative* merupakan metode clustering terbaik untuk membedakan antara kelas jinak dan ganas pada dataset kanker payudara. Kedua metode *clustering* menghasilkan pemisahan yang jelas antara kedua kelas, dan tidak ada tumpang tindih antara *cluster* yang dihasilkan.

4. Kesimpulan

Berdasarkan hasil visualisasi dan analisis dataset kanker payudara dengan metode *clustering* (*k-means*, DBSCAN, dan *agglomerative*) menggunakan bahasa pemrograman *Python* (*google collab*), dapat disimpulkan bahwa metode *clustering* terbaik yang dapat digunakan untuk membedakan antara kelas jinak dan ganas pada dataset kanker payudara adalah metode *k-means* dan *agglomerative*.

Referensi

[1] A. S. Devi, I. K. G. D. Putra, and I. M. Sukarsa, "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 6, no. 3, p. 185, 2015, doi: 10.24843/lkjiti.2015.v06.i03.p05.

[2] A. Nur Khormarudin, "Teknik Data Mining: Algoritma K-Means Clustering," *J. Ilmu Komput.*, pp. 1–12, 2016, [Online]. Available: https://ilmukomputer.org/category/datamining/.

[3] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[4] E. Susilowati, A. T. Hapsari, M. Efendi, and P. Edi, "Diagnosa Penyakit Kanker Payudara Menggunakan Metode K - Means Clustering," *J. Sist. Informasi, Teknol. Inform. dan Komput.*, vol. 10, no. 1, pp. 27–32, 2019.

[5] A. R. Aritonang, Sutarman, and P. Sihombing, "Analisis Subspace Clustering Menggunakan DBSCAN dan SUBCLU Untuk Proyeksi Pekerjaan Alumni Perguruan Tinggi," *Teknovasi*, vol. 02, pp. 33–60, 2015.

[6] J. Han, M. Kamber dan J. Pei, Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2012.

[7] E. Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: Andi, 2014.

[8] Yedla, M., Pathakota, S. R. and Srinivasa, T. M. (2010). "Enhancing K-means Clustering Algorithm with Improved Initial Center." International Journal of Computer Science and Information Technologies. 1. 121.

[9] Deka Dwinavinta Candra Nugraha, Zumrotun Naimah, Makhfuzi Fahmi dan Novi Setiani. (2014). "Klasterisasi Judul Buku dengan Menggunakan Metode K-Means." Seminar Nasional Aplikasi Teknologi Informasi. ISSN: 1907-5022. G-2.

[10] Kouzani, A. Z., Hu, M., & Rashid, H. A. F. (2017). Clustering in Biomedical Engineering In Clustering in Bioinformatics and Drug Discovery (pp. 185-201). Springer.

[11] Liu, C., & Fei, J. (2017). Data Clustering: Algorithms and Applications. CRS Press.

[12] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for dicovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231.