SKRIPSI

PEMETAAN PRODUKSI PERIKANAN TANGKAP DI INDONESIA DENGAN MENGGUNAKAN METODE DBSCAN

MUH. AKBAR IDRIS
E0119305

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SULAWESI BARAT
TAHUN 2023

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama	: Muh. Akbar Idris
Tempat/Tgl.Lahir	: Jeneponto/28 Oktober 2000
NIM	: E0119305
Program Studi	: Matematika (S1)

Menyatakan bahwa tugas akhir dengan judul "Pemetaan Produksi Perikanan Tangkap di Indonesia dengan Menggunakan Metode DBSCAN" disusun berdasarkan prosedur ilmiah yang telah melalui pembimbingan dan bukan plagiat dari karya ilmiah/naskah yang lain. Apabila dikemudian hari terbukti bahwa pernyataan ini tidak benar, maka saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Majene, 1 Maret 2023

Muh. Akbar Idris

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Muh. Akbar Idris

NIM : E0119305

Judul : Pemetaan Produksi Perikanan Tangkap di Indonesia dengan Menggunakan Metode DBSCAN

Telah berhasil dipertahankan di depan Tim Penguji (SK Nomor: 24/UN55.7/HK.04/2023) dan diterima sebagai bagian persyaratan memperoleh gelar sarjana Matematika (S.Mat) pada Program Studi Matematika dan Ilmu Pengetahuan Alam Universitas Sulawesi Barat.

Disahkan oleh:

Dekan FMIPA

Universitas Sulawesi Barat

Prof. Muhammad Abdy, M.Si., Ph.D.

NIP. 196901129199403100l

Tim penguji:

Ketua Penguji : Prof. Muhammad Abdy, M.Si., Ph.D. (.....)

Sekretaris : Rahmawati, S.Si., M.Si. (.....)

Pembimbing 1 : Rahmawati, S.Si., M.Si. (.....)

Pembimbing 2 : Apriyanto, S.Pd., M.Sc. (.....)

Penguji 1 : Musafira, S.Si., M.Sc. (.....)

Penguji 2 : Darmawati, S.Si., M.Si. (.....)

Penguji 3 : Laila Qadrini, S.Si., M.Stat. (.....)

ABSTRAK

Pemetaan produksi perikanan tangkap di Indonesia merupakan salah satu aspek penting dalam memahami dan mengelola sumber daya perikanan secara efektif. Metode DBSCAN (*Density-Based Spatial Clustering Algorithm with Noise*) telah digunakan dalam penelitian ini untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan tingkat produksi perikanan tangkap. Tujuan dari penelitian ini adalah untuk mengidentifikasi pola dan pemetaan dalam produksi perikanan tangkap di berbagai wilayah. Dalam penelitian ini, data produksi perikanan tangkap dari tahun 2017 hingga 2019 digunakan sebagai data. Langkah pertama adalah menentukan parameter DBSCAN, yaitu nilai *epsilon* dan *MinPts*, yang penting untuk menggambarkan kepadatan data dan menentukan batas *cluster*. Selanjutnya, dilakukan pemilihan titik awal data secara acak dalam perhitungan jarak menggunakan metode *euclidean*. Dari hasil pemetaan menggunakan metode DBSCAN, terbentuklah *cluster-cluster* yang mencerminkan tingkat produksi perikanan tangkap yang serupa di provinsi-provinsi yang berdekatan. *Cluster-cluster* ini membantu dalam mengidentifikasi wilayah-wilayah dengan tingkat produksi perikanan yang rendah, sedang, dan tinggi. *Noise*, yaitu data yang tidak termasuk dalam *cluster* apapun, juga diidentifikasi sebagai provinsi-provinsi dengan karakteristik produksi perikanan tangkap yang yang tinggi. *Silhouette coefficient* digunakan sebagai metrik evaluasi untuk mengukur kualitas pembentukan *cluster*. Nilai *silhouette coefficient* memberikan indikasi sejauh mana data dalam *cluster* berdekatan dengan data dalam *cluster* lainnya. Semakin tinggi nilai *silhouette coefficient*, semakin baik pembentukan *cluster* tersebut.

Kata kunci: *pemetaan,produksi perikanan tangkap, DBSCAN, clustering, silhouette*

ABSTRACT

The mapping of capture fisheries production in Indonesia is an important aspect of understanding and effectively managing fisheries resources. The DBSCAN (Density-Based Spatial Clustering Algorithm with Noise) method has been used in this research to cluster the provinces in Indonesia based on their levels of capture fisheries production. The objective of this study is to identify patterns and mapping of capture fisheries production in various regions. In this research, capture fisheries production data from 2017 to 2019 is utilized. The first step is to determine the DBSCAN parameters, namely the epsilon value and MinPts, which are important in representing data density and determining cluster boundaries. Subsequently, random initial data points are selected for distance calculations using the Euclidean method. Based on the mapping using the DBSCAN method, clusters are formed that reflect similar levels of capture fisheries production in neighboring provinces. These clusters help in identifying regions with low, moderate, and high levels of fisheries production. Noise, which refers to data points that do not belong to any clusters, is also identified, indicating provinces with distinct characteristics of high capture fisheries production. The Silhouette coefficient is used as an evaluation metric to measure the quality of cluster formation. The Silhouette coefficient value indicates how closely data within a cluster is related to data in other clusters. A higher Silhouette coefficient value indicates better cluster formation.

Keywords: *mapping, capture fisheries, DBSCAN, clustering, silhouette*

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perikanan merupakan salah satu sektor penting dalam perekonomian Indonesia, terutama perikanan tangkap yang menangkap ikan, udang, serta produk laut lainnya menurut Kementerian Kelatuan dan Perikanan (KPP, 2020). Produksi perikanan tangkap di Indonesia selalu mengalami penaikan tiap tahun, tapi masih banyak masalah dirasakan ketika melakukan upaya pemantauan serta pengelolaan produksi perikanan tangkap Indonesia. Permasalahan dirasakan yakni terbatasnya informasi spasial mengenai lokasi produksi perikanan tangkap di Indonesia. Data hasil tangkapan ikan berasal dari perairan umum dapat dikategorikan sebagai produksi perikanan tangkap (Fajriana, 2021).

Metode *clustering* merupakan teknik dimana dapat menjalankan analisis terhadap data spasial seperti produksi perikanan tangkap. Terdapat beberapa metode *clustering,* misalnya *K-Means, Hierarchical Clustering, K-Medoids, Gaussian Mixture Models* (GMM) dan DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*). Berdasarkan penelitian (Fajriana, 2021) Penerapan algoritma *k-medoids* pada sistem pengelompokan (*clustering*) produksi perikanan tangkap di Kabupaten Aceh Utara berhasil menghasilkan 3 *cluster* data. Sehingga metode ini mampu mengelompokkan data spasial berdasarkan kerapatan data, sehingga mampu mengidentifikasi wilayah-wilayah yang memiliki produksi perikanan tangkap yang rendah, sedang hingga tinggi.

Menurut penelitian (Khurin'in, 2021), berdasarkan indeks distribusi data pengangguran pada Provinsi Jawa Barat menggunakan metode DBSCAN, didapatkan beberapa *cluster* dan *noise* yang tidak dapat dihilangkan yang dapat dijadikan acuan dalam membuat kategori dalam hal ini data yang merupakan termasuk data tinggi untuk pengangguran. Oleh sebab itu penulis tertarik menggunkan metode DBSACAN karena pada penelitian ini semua titik data dibutuhkan termasuk *noise* yang merupakan kategori tinggi dalam pembuatan peta. Kemudian pada penelitian (Siti-Isfandari dkk, 2020) dengan menggunakan

metode analisis spasial *ArcGIS* dan data satelit untuk memetakan sebaran hasil tangkapan ikan pelagis pada Perairan Selat Bali, ikan pelagis merupakan kelompok ikan yang berada di lapisan permukaan air misalnya ikan tuna, cakalang dan tongkol.

Penulis ingin menggunakan metode DBSCAN sebab DBSCAN bisa mendapatkan titik data yang terdistorsi, dan dengan menggunakan *QGIS* dalam pembuatan peta dalam pemetaan karena merupakan *sotfwere* yang hampir sama dengan *ArcGIS* tapi *opensource* atau dengan katalain gratis untuk digunakan. Melalui penjelasan sebelumnya, penulis tertarik menjalankan penelitian berjudul "Pemetaan Produksi Perikanan Tangkap di Indonesia menggunakan Metode DBSCAN".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka permasalahan dalam penelitian ini adalah:

1. Bagaimanakah penerapan metode DBSCAN dalam memetakan produksi perikanan tangkap di Indonesia?
2. Bagaimana hasil pemetaan produksi perikanan tangkap Indonesia menggunakan metode DBSCAN?

1.3 Tujuan Penelitian

1. Untuk menerapkan metode DBSCAN dalam memetakan produksi perikanan tangkap di Indonesia.
2. Untuk mendapatkan hasil pemetaan produksi perikanan tangkap Indonesia menggunakan metode DBSCAN.

1.4 Manfaat Penelitian

Manfaat penelitian sebagai beriku:

1. Menjadi referensi untuk pihak yang terkait dalam melakukan pengelolaan produksi perikanan tangkap di Indonesia dimana peta yang dihasilkan dapat memberikan solusi pada pemerintah Indonesia dalam memberikan bibit ataupun

alat yang dapat membantu dalam produksi perikanan tangkap pada provinsi yang rendah dalam menghasilkan perikanan tangkap.

2. Penelitian memberi solusi bagi pemerintahan Indonesia memahami tingkatan produksi perikanan tangkap pada tiap Provinsi Indonesia ke dalam bentuk informasi yang visual atau gambar dalam bentuk peta sehingga pemerintah Indonesia dapat memberikan bantuan pada provinsi yang produksi perikanan tangkapnya yang rendah.

1.5 Batasan Masalah

Penulis membatasi permasalahan dalam penelitian sebagai berikut:

1. Metode digunakan yakni DBSCAN.
2. Jarak digunakan pada penelitian yakni jarak *Euclidean*.
3. Data digunakan yakni data produksi perikanan tangkap pada tiap Provinsi Indonesia tahun 2017, 2018, serta 2019.
4. Rentang untuk kategori klasifikasi dibuat oleh penulis untuk memberikan batasan pada setiap kategori, dalam hal ini produksi perikanan tangkap di Indonesia dalam satuan ton.

BAB II

KAJIAN PUSTAKA

2.1 Perikanan Tangkap

Perikanan tangkap adalah kegiatan menangkap ikan dan bahan hayati lainnya dari laut, sungai, dan danau, perikanan tangkap menjadi salah satu sektor penting pada perekonomi Indonesia sebab memberikan kontribusi terhadap pendapatan nasional dan devisa negara, dalam sektor perikanan dan kelautan, subsektor perikanan tangkap menjadi salah satu kontributor utama (KPP, 2020). Berdasarkan data BPS Indonesia 2020, sektor perikanan dan kelautan di Indonesia tumbuh sebesar 0,39% dan memberikan kontribusi sebesar 1,11% terhadap Produk Domestik Bruto (PDB) nasional.

2.2 Data *Mining*

Knowledge Discovery in Database atau biasa disebut dengan KDD pada data *mining* bermakna pengumpulan data untuk menemukan hubungan data dalam kumpulan data besar (Rohalidyawati dkk, 2020). Tahapan proses data *mining* dimulai pemilihan data dari sumber data yang menjadi data target, proses dalam pra-pemprosesan data untuk meningkatkan transformasi kualitas data, serta langkah interpretasi dan evaluasi yang menghasilkan *output* berupa data yang baru diharapkan memberikan hasil yang lebih baik (Fayyad, 1996).

Definisi 1 **: Data *Mining* (Fayyad, 1996)**

Diberikan sebuah dataset U yang terdiri dari n titik data dengan m atribut, data *mining* merupakan proses untuk mendapatkan hubungan serta pola tersembunyi pada dataset U . Dalam proses data *mining*, digunakan berbagai teknik dan algoritma matematika seperti *clustering*, klasifikasi, *regresi*, dan asosiasi untuk mengolah data U sehingga dapat diperoleh informasi dan hasil. Proses data *mining* dapat didefinisikan sebagai fungsi f yang mengambil dataset U sebagai *input* dan menghasilkan G sebagai *output*, yaitu $f(U) = G$.

2.3 Analisis *Cluster*

Menurut Sitepu dkk, Analisis *cluster* (*clustering analysis*) merupakan metode analisis data yang bertujuan untuk mengenali dan mengelompokkan objek-objek data ke dalam *cluster-cluster* berdasarkan kesamaan karakteristik atau pola tertentu, analisis multivariat digunakan untuk menganalisis hubungan antara variabel independen dan variabel dependen dalam konteks tertentu. Namun, dalam analisis *cluster*, tidak ada pembagian khusus antara variabel independen dan variabel dependen. Sebaliknya, tujuan analisis *cluster* adalah untuk mengelompokkan sejumlah titik data yang memiliki karakteristik serupa ke dalam *cluster* yang berbeda.

Variabel bebas (*independent variable*) digunakan untuk memprediksi atau menjelaskan variasi pada variabel terikat (*dependent variable*). Variabel bebas pun dikenali menjadi variabel prediktor ataupun variabel penjelas. Dalam analisis regresi, variabel bebas ditempatkan pada sumbu-x dan digunakan untuk memprediksi variabel terikat yang ditempatkan pada sumbu-y. Sementara itu, variabel terikat yakni variabel dimana terpengaruh variabel bebas serta akan diprediksi atau dijelaskan dari variabel bebas tersebut. Variabel terikat pun dikenal menjadi variabel respons ataupun yang diprediksi.

Dalam analisis *clustering*, tidak ada variabel tunggal yang diinterpretasikan sebagai variabel bebas atau terikat. Hal ini disebabkan karena dalam *clustering*, tidak ada variabel yang diprediksi atau dijelaskan oleh variabel lainnya. Tujuan utama *clustering* agar mengelompokkan data ke dalam kelompok yang serupa berdasarkan kesamaan atau kemiripan karakteristik data tanpa memperhatikan variabel mana yang lebih penting atau lebih dipengaruhi oleh variabel lainnya. Sehingga, analisis *clustering* tak terdapat perbedaan antar variabel terikat serta bebas.

***Definisi* 2 : Analisis *Cluster* (Sitepu dkk, 2011)**

Diberikan sebuah dataset U yang terdiri dari n titik data dengan m atribut, analisis *cluster* adalah proses untuk menemukan kelompok atau *cluster* $C = \{C_1, C_2, ..., C_n\}$ yang terdiri dari titik data yang memiliki kemiripan yang tinggi

dalam matriks *similarity* atau *dissimilarity*. Setiap *cluster* C_i dalam C memiliki karakteristik yang berbeda dengan *cluster* lainnya, sedangkan setiap titik data dalam *cluster* memiliki kemiripan yang tinggi dengan titik data lainnya dalam *cluster* yang sama.

2.4 Standarisasi Data

Standarisasi data digunakan untuk menghitung normalisasi *z-score* pada sebuah variabel data secara berurutan agar dapat terhindar akan masalah yang diakibatkan pengguna, nilai skala yang seimbang atau dengan kata lain melalui data asli tak terlampaui jauh akan data lainnya dengan variabel pengelompokan titik data (Walpole dkk, 1995).

***Definisi* 3 : Standarisasi Data *z-score* (Walpole dkk, 1995)**

Misalkan x_i adalah nilai pengamatan pada suatu data, $\bar{x}$ yakni rerata data, serta s yakni standar deviasi data, maka *z-score* dari x_i dapat dihitung menggunakan rumus:

$$Z_i = \frac{x_i - \bar{x}}{s}, \tag{2.1}$$

rumus *z-score* (2.1) didasarkan pada teori statistik dan probabilitas yang menggunakan konsep distribusi normal atau distribusi Gauss. Distribusi ini simetris dan sebagian besar nilai dalam data berada di sekitar rata-rata.

Dalam distribusi normal, nilai *z-score* dapat dihitung sebagai jarak antara nilai pengamatan pada suatu data dan rata-rata data dalam satuan standar deviasi. Dengan menggunakan definisi distribusi normal, penulis dapat membuktikan bahwa *z-score* akan memiliki nilai 0 jika nilai pengamatan pada suatu data sama dengan rata-rata data, nilai negatif jika nilai pengamatan pada suatu data lebih rendah dari rata-rata data, dan nilai positif jika nilai pengamatan pada suatu data lebih tinggi dari rata-rata data.

Contoh 2.4.1

Di kelas matematikanya, Barid mendapat nilai 80, sedangkan standar deviasi serta nilai rata-rata untuk seluruh kelas adalah 75 serta 9,29. Barid mendapat nilai 85 dalam pelajaran Kimia di kelas yang sama. Semua siswa pada kelas tersebut

mempunyai nilai mean 82,5, dan standar deviasinya yakni 7,54. Topik mana yang membuat Barid lebih baik, tepatnya?

Jawab :

Pelajaran Matematika

$$Z_1 = \frac{x_i - \bar{x}}{s}$$
$$= \frac{80 - 75}{9,29}$$
$$= \frac{5}{9,29}$$
$$= 0,538.$$

Pelajaran Kimia

$$Z_2 = \frac{x_i - \bar{x}}{s}$$
$$= \frac{85 - 82,5}{7,54}$$
$$= \frac{2,5}{7,54}$$
$$= 0,331,$$

maka dari perhitungan diatas, nilai terbaik menurut *z-score* pada Barid yaitu pada pelajaran Matematika.

2.5 Jarak *Euclidean*

Jarak *euclidean* adalah perhitungan jarak dua titik dalam ruang geometri yang menghubungan antara sudut dan jarak (Isnarwaty dkk, 2016).

***Definisi* 4 : Jarak *Euclidean* (Isnarwaty dkk, 2016)**

Misal terdapat dua titik $(x_1, x_2, ..., x_n)$ dan $(y_1, y_2, ..., y_n)$. Dalam matematika, jarak *euclidean* didefinisikan sebagai jarak antara dua titik. Dengan kata lain, jarak *euclidean* adalah jarak antara dua titik dalam ruang *euclidean* didefinisikan sebagai segmen garis antara dua titik, dituliskan sebagai berikut:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\,, \tag{2.2}$$

terdapat dua titik $(x_1,x_2,...,x_n)$ dan $(y_1,y_2,...,y_n)$. Jarak *euclidean* antara dua titik tersebut dapat dinyatakan sebagai:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_n-y_n)^2}\,, \tag{2.3}$$

dengan menggunakan sifat *pythagoras* di segitiga siku-siku pada persamaan 2.3. Misalkan a serta b yakni panjang sisi-sisi segitiga siku-siku serta c yakni panjang sisi miringnya. Dalam hal ini, a dan b dapat dianggap sebagai perbedaan antara komponen vektor, sedangkan c merupakan jarak *euclidean* yang diinginkan, dengan demikian, penulis dapat menghitung c dengan menggunakan rumus:

$$c=\sqrt{(a^2+b^2)}\,, \tag{2.4}$$

ini dapat digunakan untuk menentukan jarak terdekat ke titik data. Jumlah kesamaan sampel dapat ditentukan dengan menggunakan rumus jarak *euclidean*.

Contoh 2.5.1

Untuk mengukur jarak antara dua titik A dan B pada bidang koordinat *cartesius* dengan koordinat A(2,3) dan B(5,7), penulis dapat menggunakan rumus jarak *euclidean*!

Penyelesaian:

Jarak *euclidean* antara A dan B $=\sqrt{((5-2)^2+(7-3)^2)}$

$$=\sqrt{(3^2)+(4^2)}$$
$$=\sqrt{(9+16)}$$
$$=\sqrt{25}$$
$$=5\,.$$

Jadi, jarak *euclidean* antara A dan B adalah 5 satuan.

2.6 *Density-Based Spatial Clustering Algorithm with Noise* (DBSCAN)

Density-Based Spatial Clustering Algorithm with Noise disingkat DBSCAN merupakan sebuah algorima pengelompokan berdasarkan kepadatan (*density*) data.

Density adalah jumlah minimum data dalam radius yang termasuk dalam kelas kepadatan yang didapatkan. Ada tiga jenis konsep kepadatan: *noise*, *border* (batas), dan *core* (inti). Data selanjutnya masuk ke *core* ketika radius (ε) ≥ titik *cluster* minimum (*MinPts*). Ketika $\varepsilon \leq$ *MinPts* sera terdapat *neighbor* (jarak antar data) yaitu jadi inti, hingga itu dikatakan batas. Lain jika $\varepsilon \leq$ *MinPts* serta tak adanya *neighbor* dimana jadi inti disebut *noise* (Safitri dkk, 2017).

***Definisi* 5 : DBSCAN (Safitri dkk, 2017)**

Diberikan sebuah dataset $D = \{x_1, x_2, ..., x_n\}$, dimana setiap x_i adalah titik data dalam ruang dengan dimensi dua. DBSCAN memiliki dua parameter, yakni *epsilon* (ε) dan *minimum number of points* (*MinPts*), yang digunakan untuk menentukan daerah-daerah padat dalam dataset D.

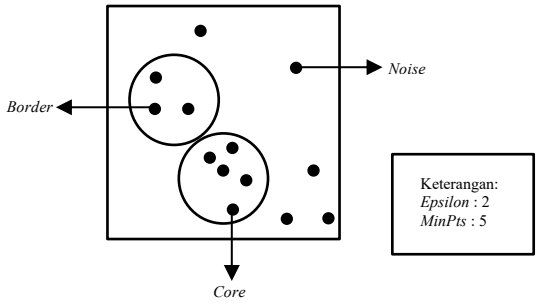
Terdapat beberapa keunggulan pada metode DBSCAN antara lain (Wuryandari dkk, 2017) :

a. Metode DBSCAN dalam menentukan jumlah *cluster*-nya berdasarkan pada nilai paramater *epsilon* dan *MinPts* dengan kata lain metode ini tidak ditentukan jumlah *cluster*-nya ketika melakukan perhitungan pada program komputer.
b. Pada *noise* atau dengan kata lain titik yang menganggu terdapat pada metode DBSCAN.
c. Dua parameter pada metode DBSCAN tidak rawan terhadap pada urutan titik data.

Kemudian menurut Yumono dkk, terdapat beberapa istilah pada metode DBSCAN yaitu:

1. *Core* (Inti):

Berdasarkan titik inti, pusat *cluster* (nilai radius atau ambang batas). Penulis menentukan *MinPts* (*minimum cluster point*), dan saat $\varepsilon \geq$ *MinPts* terpenuhi, maka disebut dengan nama *core* (inti).



Gambar 2.1 ***Core Point***

2. **Batas (*Border*):**

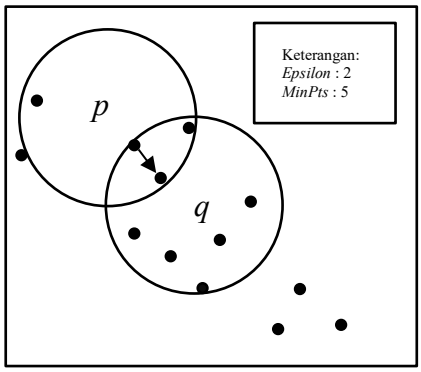
   Titik ≤ *MinPts* disebut titik dimana ada pada satu *cluster.*

3. ***Noise*:**

   Titik diman bukan anggota *core*, kemudian bukan *border*, serta tak termasuk pada anggota *cluster*.

4. ***Directly density-reachable*:**

   Directly Reachable Density adalah jarak antar titik di bawah nilai ε. Titik p dinyatakan titik inti dan q yakni titik dalam ε melalui titik p. Sehingga *Directly Reachable Density* ialah titik terhubung langsung ke titik inti dapat diakses langsung dalam *density*.

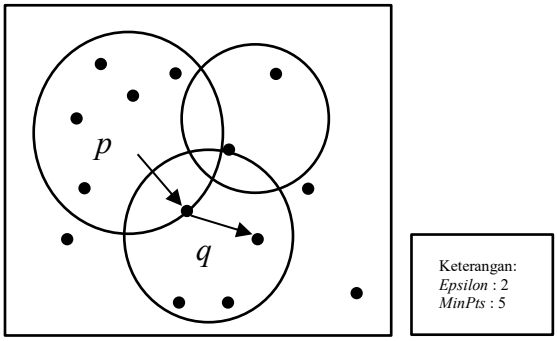


Gambar 2.2 ***Directly Reachable Density***

5. ***Density reachable:***

   Density yang dapat dicapai yakni titik dimana tidak terhubung langsung ke intinya. Titik data q yakni batas serta titik data p yakni intinya tapi ketika

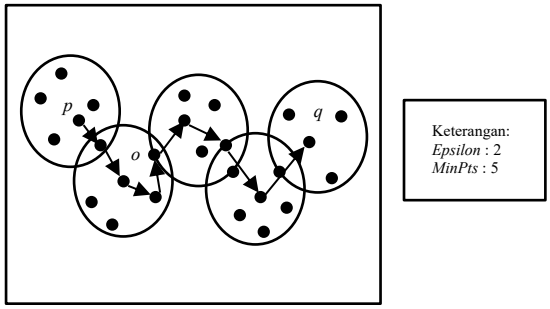
titik data q serta p yakni *border* hingga ia termasuk kedalam *density connected*.



Gambar 2.3 ***Density Reachable Density***

6. ***Density Connected:***

Titik dimana dihubungkan dari titik lainnya. Titik p serta q yakni sisi, titik o yaitu titik dimana menghubungkan q serta p hingga q serta p adalah kepadatan (*density*) yang dapat dicapai melalui o.



Gambar 2.4 ***Density Connected***

Berdasarkan penjelasan diatas lebih lanjut, metode DBSCAN melakukan pencarian titik *core* dan daerah padat dengan melakukan langkah-langkah sebagai berikut:

a. Untuk setiap titik data x_i dalam dataset D, hitung jarak antara x_i dan semua titik data dalam dataset D.

b. Tentukan daerah padat (*density*) dengan memeriksa apakah terdapat setidaknya *MinPts* titik data lain yang berada dalam jarak ε dari suatu titik data x_i. Jika terdapat, maka x_i adalah titik *core* dan semua titik data yang berada dalam jarak ε dari x_i akan menjadi bagian dari daerah padat (*density*) yang sama.

c. Dalam daerah padat (*density*), DBSCAN menandai setiap titik data yang terhubung dengan titik *core* sebagai bagian dari *cluster* yang sama.
d. Setiap titik data yang tidak termasuk dalam daerah padat (*density*) atau tidak terhubung dengan titik *core* disebut sebagai titik *noise*.

Selanjutnya, *Density-Based Spatial Clustering Algorithm with Noise* mengelompokkan titik data menjadi *cluster* dengan melakukan langkah-lankah yakni:

a. Memilih salah satu titik *core* x_i dimana belum termasuk dalam suatu *cluster*.
b. Kelompokkan semua titik data yang terhubung dengan x_i ke dalam sebuah *cluster* baru.
c. Ulangi proses ini untuk setiap titik *core* yang belum termasuk dalam *cluster* hingga semua titik data termasuk dalam suatu *cluster* atau dianggap sebagai titik *noise*.

2.7 Metode Iteratif

Metode iteratif adalah pendekatan dalam memecahkan masalah yang melibatkan perulangan atau pengulangan beberapa langkah atau prosedur tertentu sampai mencapai suatu kondisi akhir atau solusi yang diinginkan. Dalam konteks *clustering*, metode iteratif dapat digunakan untuk menentukan parameter-parameter penting yang dibutuhkan dalam algoritma *clustering*, seperti *epsilon* (ε) pada metode DBSCAN (Zhang dkk, 2019).

***Definisi* 7 : Metode Iteratif**

Misalkan i_n adalah nilai variabel pada iterasi ke-n, f adalah fungsi atau operasi yang diulang pada setiap iterasi, dan $i_{\{n+1\}}$ adalah nilai variabel pada iterasi berikutnya.

$$i_{\{n+1\}} = f(i_n), \tag{2.5}$$

dalam *clustering*, metode iteratif dapat diterapkan untuk menentukan parameter-parameter penting seperti *epsilon* pada DBSCAN, sehingga menghasilkan pengelompokan yang lebih baik dan optimal. Algoritma iteratif pada dasarnya merupakan algoritma yang menggunakan pendekatan iteratif untuk menentukan

pengelompokan yang optimal dengan mengulangi suatu proses atau langkah secara terus-menerus sampai mencapai kondisi penghentian atau solusi optimal.

2.8 Pemilihan *Epsilon* dan *MinPts*

Metode DBSCAN memiliki dua parameter pengolahan data *epsilon* (ε) serta *MinPts*, *epsilon* (ε) yakni jarak atau radius dimana ditentukan melalui titik data ke target, menghitung jumlah *neigbhor* dalam titik data, dengan menggunakan metode iteratif, mulai dari nilai *epsilon* yang kecil, penulis dapat menguji model DBSCAN pada data dan memperbesar nilai *epsilon* secara bertahap hingga jumlah *cluster* atau *noise* pada model terlihat stabil (Zhang dkk, 2019), *MinPts* adalah jumlah titik data jarak minimum untuk membentuk *cluster*.

2.9 Validasi *Silhouette Coefficient*

Setelah mengelompokkan evaluasi hasil *clustering*-nya dengan validasi *cluster.* Validasi dilakukan dengan pengukuran bagaimana mencapai hasil pengelompokkan yang baik. Dalam penelitian ini digunakan validasi *silhouette coefficient.*

***Definisi* 7 : Nilai *silhouette* (Rousseeuw, 1986 dan Kaufman dkk, 2005)**

Misal terdapat i titik data yang dikelompokkan ke dalam *cluster,* dan titik i termasuk ke dalam *cluster* $b(i)$. Sedangkan jarak antara titik data i dan anggota *cluster* $a(i)$ merupakan hasil dari *cluster* pertama, kemudian jarak antara titik data i dan anggota *cluste* terdekat (yang bukan $b(i)$) dikurang dengan $a(i)$ merupakan *cluster* kedua, kemudian dibagi $Max\left[a(i),b(i)\right]$. Berikut rumus menghitung nilai *silhouette* :

$$s(i)=\frac{b(i)-a(i)}{Max[a(i),b(i)]}, \tag{2.6}$$

$a(i)$ yaitu *cluster* pertama antar suatu titik dengan semua titik dalam *cluster* sama serta $b(i)$ yakni *cluster* kedua terdekat antar suatu titik dengan titik pada *cluster* yang berbeda.

Untuk mempermudah, dapat diasumsikan bahwa titik data yang akan diklasifikasikan adalah titik dalam ruang n-dimensi. Selanjutnya, penulis dapat mengasumsikan bahwa setiap titik data adalah titik dalam ruang n-dimensi yang memiliki koordinat $(x_1, x_2, ..., x_n)$.

Asumsikan terdapat C *cluster* pada data. Kemudian asumsikan bahwa memilih satu titik pada suatu *cluster*, lalu hitung jarak rata-rata titik tersebut akan semua titik *cluster* sama. Dalam hal ini, jarak rata-rata bisa dihitung menggunakan persamaan:

$$a(i) = \left(\frac{1}{(N-1)} \right) \times \sum_{i=1}^{n} (d(x_i, C_i)), \tag{2.7}$$

dimana $d(x_i, C_i)$ adalah jarak antara titik data x_i dan titik inti C_i, N adalah jumlah titik data *cluster* yang sama, $\sum_{i=1}^{n}$ adalah penjumlahan nilai untuk setiap i, yang merupakan indeks dari titik data dalam *cluster* yang sama.

Selanjutnya, akan dihitung jarak rata-rata terdekat antara suatu titik dengan titik pada *cluster* yang berbeda. Dalam hal ini, jarak rata-rata terdekat dapat dihitung menggunakan persamaan:

$$b(i) = \underset{j \neq C}{Min} \left(\frac{1}{N_k} \right) \times \sum_{i=1}^{n} (d(x_i, C_j)), \tag{2.8}$$

dimana N_k yakni jumlah titik *cluster* k, C_j yaitu titik inti *cluster* k, dan $d(x_i, C_j)$ adalah jarak antara x_i dan C_j.

Setelah memiliki nilai $b(i)$ serta $a(i)$ untuk tiap titik dalam setiap *cluster*, dapat dihitung nilai *silhouette coefficient* untuk setiap titik menggunakan persamaan (2.6).

Nilai *silhouette* ada dalam rentang $-1 \leq s(i) \leq 1$, kemudian menghitung nilai *silhouette* yang merupakan *mean* nilai *silhouette* dari seluruh titik-titik data di setiap *cluster*,

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{jika } a(i) < b(i), \\ 0 & \text{jika } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{jika } a(i) > b(i). \end{cases}$$

dapat dilihat bahwa nilai *silhouette coefficient* sekitar dari -1 sampai 1. Nilai 1 menyatakan bahwa titik data sangat cocok dalam *cluster*-nya dan terpisah dengan *cluster* yang berbeda, sedangkan nilai -1 menunjukkan bahwa titik data kurang cocok dalam *cluster*-nya dan seharusnya ditempatkan di *cluster* yang berbeda. Nilai 0 menunjukkan bahwa titik data berada di antara dua *cluster* yang berbeda dan tidak dapat dengan pasti ditempatkan pada salah satu *cluster*.

Contoh 2.9.1

Misalkan terdapat data sebagai berikut:

Data 1 = [2, 5], Data 2 = [3, 4], Data 3 = [4, 5]

Data 4 = [6, 2], Data 5 = [7, 3] ,Data 6 = [5, 1]

jika data tersebut dikelompokkan menjadi dua *cluster*, yakni [Data 1, Data 2, Data 3] dan [Data 4, Data 5, Data 6], maka dapat dihitung nilai *silhouette* untuk setiap data sebagai berikut!

Penyelesaian:

Untuk Data 1 hingga Data 6 dapat menggunakan rumus pada persamaan (2.7) untuk $a(i)$ dan $b(i)$ pada persamaan (2.8):

Untuk Data 1:

$a(1)$ adalah rata-rata jarak dari Data 1 ke Data 2 dan Data 3 $= \dfrac{(3+4)}{2} = 3,5.$

$b(1)$ merupakan rata-rata jarak dari Data 1 ke Data 4, Data 5, dan Data 6

$$= \frac{(5+5+4)}{3} = 4,67,$$

$$s(1) = \frac{(4,67 - 3,5)}{Max[3,5.4,67]} = 0,225.$$

Untuk Data 2:

$a(2)$ adalah rata-rata jarak dari Data 2 ke Data 1 dan Data 3 $= \dfrac{(3+1)}{2} = 2.$

$b(2)$ merupakan rata-rata jarak dari Data 2 ke Data 4, Data 5, dan Data 6

$$= \frac{(4+4+3)}{3} = 3,67,$$

$$s(2) = \frac{(3,67-2)}{Max[2.3,67]} = 0,405 .$$

Untuk Data 3:

$a(3)$ adalah rata-rata jarak dari Data 3 ke Data 1 dan Data 2 $= \frac{(4+1)}{2} = 2,5.$

$b(3)$ merupakan rata-rata jarak dari Data 3 ke Data 4, Data 5, dan Data 6

$$= \frac{(4+4+4)}{3} = 4,$$

$$s(3) = \frac{(4-2,5)}{Max[2,5.4]} = 0,375.$$

Untuk Data 4:

$a(4)$ adalah rata-rata jarak dari Data 4 ke Data 5 dan Data 6 $= \frac{(2+3)}{2} = 2,5.$

$b(4)$ merupakan rata-rata jarak dari Data 4 ke Data 1, Data 2, dan Data 3

$$= \frac{(5+4+4)}{3} = 4,33,$$

$$s(4) = \frac{(2,5-4,33)}{Max[4,33.2,5]} = -0,416.$$

Untuk Data 5:

$a(5)$ adalah rata-rata jarak dari Data 5 ke Data 4 dan Data 6 $= \frac{(5+2)}{2} = 3,5.$

$b(5)$ merupakan rata-rata jarak dari Data 5 ke Data 1, Data 2, dan Data 3

$$= \frac{(5+4+4)}{6} = 4,33,$$

$$s(5) = \frac{(4,33-3,5)}{Max[3,5.4,33]} = 0,160.$$

Untuk Data 6:

$a(6)$ adalah rata-rata jarak dari Data 6 ke Data 4 dan Data 5 $= \frac{(3+2)}{2} = 2,5.$

$b(6)$ merupakan rata-rata jarak dari Data 6 ke Data 1, Data 2, dan Data 3

$$= \frac{(3+3+3)}{3} = 3,$$

$$s(6) = \frac{(2,5-3)}{Max[2,5.3]} = -0,166,$$

melalui perhitungan sebelumnya, bisa dipahami jika nilai *silhouette* tertinggi adalah 0,405 yang diperoleh dari Data 2. Sedangkan nilai *silhouette* terendah adalah -0,416 yang diperoleh dari Data 4. Nilai *silhouette* dapat membantu menentukan hasil *clustering* dimana didapatkan tidak ataupun baik. Semakin mendekati 1, sehingga semakin baik hasil *clustering* dimana diperoleh. Sedangkan makin mendekati -1, hingga makin buruk hasil *clustering* yang diperoleh. Jika nilai *silhouette* mendekati 0, maka dapat dikatakan bahwa hasil *clustering* tidak terlalu baik dan perlu dilakukan evaluasi lebih lanjut (Rousseeuw, 1986).

***Definisi* 8 *:Silhouette Coefficient* (Kaufman dkk, 2005)**

Misal terdapat kumpulan data *silhouette* $(s_{(i)} + ... + s_{(n)})$ kemudian data tersebut dibagi dengan banyaknya data ke-n *silhouette.* Lebih lanjut dapat dilihat pada rumus di bawah ini:

$$SC = \frac{s_{(i)} + ... + s_{(n)}}{n}. \tag{2.9}$$

Contoh 2.9.2

Dari contoh 2.9.1 dapat dihitung rata-rata *silhouette coefficient*-nya sebagai berikut!

Penyelesaian:

Diketahui untuk nilai $s(1) - s(6)$ adalah 0,225, 0,405, 0,375, -0,416, 0,160, -0,166.

Dengan menggunakan rumus pada persamaan (2.8) maka,

$$SC = \frac{s_{(i)} + ... + s_{(n)}}{n}$$

$$SC_1 = \frac{0,225 + 0,405 + 0,375 + (-0,416) + 0,160 + (-0,166)}{6},$$

$$SC_1 = 0,098.$$

Jadi, untuk nilai SC_1 nya adalah 0,098.

Berikut ini merupakan tabel kriteria pengukuran *Silhouette Coefficient* ada di tabel 2.1

Tabel 2.1 Kriteria pengukuran *Silhouette Coefficient*

Nilai SC	Kriteria
0,71-1,00	Struktur Kuat
0,51-0,70	Struktur Baik
0,26-0,50	Struktur Lemah
≤0,25	Struktur Buruk

Sumber :Kaufman dkk, 2005.

Silhouette coefficient adalah salah satu metrik evaluasi *clustering* yang digunakan untuk mengevaluasi seberapa baik titik data berada dalam *cluster* yang sama dengan titik data yang serupa dan berbeda dengan titik data dalam *cluster* yang berbeda. Kriteria pengukuran ini memberi nilai sekitar -1 serta 1, nilai 1 menunjukkan bahwa titik data sangat cocok dengan *cluster*-nya, nilai 0 menunjukkan titik data berada tepat di antara dua *cluster*, dan nilai negatif menunjukkan bahwa titik data seharusnya ditempatkan dalam *cluster* yang berbeda.

Metode perhitungan *silhouette coefficient* dilakukan dengan mempertimbangkan jarak titik data serta titik data dalam *cluster* dimana sama serta jarak antar titik data serta titik data dalam *cluster* berbeda. Kriteria pengukuran *silhouette coefficient* dapat digunakan untuk mengevaluasi kualitas *clustering* dalam dua cara: (1) secara keseluruhan, melalui perhitungan rata-rata nilai $s(i)$ dari seluruh titik data pada dataset; dan (2) secara individual, dengan melihat nilai $s(i)$ untuk setiap titik data dalam dataset. Semakin tinggi nilai *Silhouette Coefficient*, semakin baik kualitas *clustering*. Namun, perlu diingat bahwa *Silhouette Coefficient* bukanlah satu-satunya metrik evaluasi *clustering* yang tersedia dan harus digunakan bersama dengan metrik evaluasi lainnya agar menghasilkan ilustrasi lebih komprehensif mengenai kualitas *clustering*.

DAFTAR PUSTAKA

Apuke, Destiny, O. 2017. *Quantitative Research Methods a Synopsis Approach.* Arabian J Bus Manag Review (Kuwait Chapter). 6(10): 40-47.

Badan Pusat Statistik. (2022). Produksi Perikanan Tangkap Umum Menurut Lokasi di Indonesia 2017-2019. https://www.bps.go.id/indicator/56/1519/1/produksi-perikanan-tangkap-di-perairan-umum-menurut-lokasi.html. Diakses pada tanggal 12 November 2022.

Fajriana. (2021). Analisis Algoritma K-Medoids pada Sistem *Cluster*isasi Produksi Perikanan Tangkap Kabupaten Aceh Utara. Jurnal Edukasi dan Penelitian Informatika, 7(2), 92-101.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-54.

Isnarwaty, Putri, D., Irhamah. 2016. Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means. *Jurnal Sains dan Seni ITS*. 8(2): 137-144.

Kaufman, L. dan Rousseeuw, P. J. (2005). "Validation of Clusters". In Finding Groups in Data: An Introduction to Cluster Analysis. *Wiley-Interscience*, New York. hal. 91-101.

Khurin'in, A.I. (2021). Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Tingkat Sebaran Pengangguran Menggunakan Metode *Density-Based Spatial Clustering Algoruthm with Noise* (DBSCAN). Univ. Islam Negeri Sunan Ampel Surabaya, Surabaya.

Kementerian Kelautan dan Perikanan. (2020). Statistik Perikanan Tangkap Indonesia 2019. Jakarta: KKP.

Rohalidyawati, Windy, Rahmawati, R., Mustafid. 2020. Segmentasi Pelanggan E-Money Dengan Menggunakan Algoritma DBSCAN Density-Based Spatial Clustering Application with Noise di Provinsi DKI Jakarta. *Jurnal Gaussian*. 9(2): 162-169.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Safitri, Diah, Wuryandari, T., Rahmawati, R. 2017. Metode DBSCAN Untuk Pengelompokkan Kabupaten/Kota di Provinsi Jawa Tengah. *Jurnal Statistika*. 5(1): 1-6.

Siti-Isfandari, N., & Syawaludin, M. (2020). Pemetaan sebaran hasil tangkapan ikan pelagis di perairan Selat Bali menggunakan metode analisis spasial ArcGIS dan data satelit. Jurnal Kelautan: *Indonesian Journal of Marine Science and Technology*, 13(1), 46-54.

Sitepu, Robinson, Irmeilyana, Gultom, B. 2011. Anlisi Cluster terhadap Tingkat Pencemaran Udara pada Sektor Industri di Sumatera Selatan. *Jurnal Penelitian Sains*. 14(3A): 11-17.

Walpole, R. E. (1995). *Introduction to statistics*. Macmillan Publishing Company.

Wuryandari, A. S., Aini, Q. N., & Kusumadewi, S. (2017). Penerapan Metode DBSCAN pada Data Crime untuk Identifikasi Tingkat Kriminalitas di Wilayah Surabaya. Jurnal Teknologi dan Sistem Komputer, 5(1), 15-21.

Yumono, Andreas, Oslan, Y. dan Dwijono, D. 2009. Implementasi Metode Density-Based Spatial Clustering Applications with Noise untuk Mencari Arah Penyeberan Wabah Deman Berdarah. *Jurnal EKSIS*. 02(01): 11-21.

Zhang, X., Sun, Y., Liu, J., Zhang, Y., & Sun, Y. 2019. *Density-based clustering for imbalanced data based on a novel iterative method*. Information Sciences, 484, 348-367.