Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Algoritma *Mean Shift* untuk Menentukan Segmentasi Pelanggan pada Penjualan Toko *Online*

(*Mean Shift Algorithm to Determine Customer Segmentation in Online Store Sales*)

Ryan Reliovani[1], Nina Nadia Syafitri Husein[2], Kamal Zaki Abdurrafi[3], Cecep Rafqi Al Husni[4], Muhammad Azka Khowarizmi[5]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ryan.reliovani@gmail.com
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ninanadiasyafitrihusein@gmail.com
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, kamalzakiab@gmail.com
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ceceprafqi19@gmail.com
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, azkakhowarizmi@gmail.com

Abstrak

Segmentasi pasar merupakan salah satu hal yang sangat krusial bagi sebuah bisnis ataupun usaha, dengan adanya segmentasi pasar sebuah toko atau perusahaan dapat mengetahui kemampuan daya beli, kebutuhan dan karakteristik pelanggan. Tujuan dari penelitian ini adalah untuk mengetahui nilai segmentasi pelanggan pada penjualan di sebuah toko online yang berbasis di Inggris dengan penjualan utamanya merupakan hadiah unik untuk berbagai acara dimana pelanggan toko merupakan pedagang grosir dari berbagai negara. Data mining dengan teknik Clustering dimanfaatkan dalam penelitian ini. Algoritma yang digunakan untuk membangun klaster adalah algoritma Mean Shift, dengan estimasi nilai bandwith 1.55, nilai quantil =0.4 dan n_samples = 500 didapatkan 3 klaster yang divisualisasi menggunakan model scatter plot.

Kata kunci: algoritma mean shift, data mining, segmentasi pelanggan

Abstract

Market segmentation is one of the most important things for a business or business, with market segmentation a shop or company can see the purchasing power, needs and customers of customers. The purpose of this study was to determine the value of customer segmentation in an online shop based in the UK where the main sales are unique gifts for various events where the shop's customers are wholesalers from various countries. Data mining with clustering techniques is used in this study. The algorithm used to build clusters is the Mean Shift algorithm, with an estimated bandwidth value of 1.55, the quantile value = 4, epsilon = 4% and n_samples = 5000, there are 3 clusters visualized using a scatter plot model.

Keywords: *customer segmentation, data mining, mean shift algorithm*

1 Pendahuluan

Segmentasi pasar merupakan hal yang sangat penting dalam menentukan strategi marketing sebuah perusahaan atau industri kedepannya. Dengan mengetahui segmentasi pasar, perusahaan dapat

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

meningkatkan efektivitas pemasaran, dan mengeluarkan produk yang memang dibutuhkan oleh pelanggan. Segmentasi dengan menggunakan metode FRM (*Frequency, Recency, Monitory*) digunakan untuk mencari probabilitasnya terhadap toko online tersebut. *Frequency* sendiri merupakan besarnya jumlah transaksi pelanggan berdasarkan kurun waktu tertentu, *Recency* adalah besarnya jumlah hari dari tanggal terakhir transaksi ke hari ini. Sedangkan *Monetary* adalah besarnya nilai total transaksi yang dibayarkan pelanggan dalam kurun waktu tertentu.

Pendekatan klasterisasi pada teknik data mining [1], [2], yang dapat digunakan sebagai solusi untuk menganalisis segmentasi pasar tersebut berdasarkan aktivitas pelanggan. Salah satu algoritma yang dapat digunakan adalah algoritma Mean Shift [3]. Terdapat beberapa penelitan terdahulu yang terkait, antara lain: *object tracking* dengan menggunakan algoritma mean shift [4][5]; (2) klasterisasi gambar atau citra dnegan menggunakan algoritma mean shift [6]; dan kombinasi algoritma mean shift dan k-means untuk klasterisasi data batu bara [7].

Tujuan dari percobaan ini sendiri adalah untuk mengetahui apakah algoritma *mean shift* dapat mensegmentasikan pasar berdasarkan dataset penjualan di sebuah toko online sebanyak 541909 data dengan jangka waktu penjualan satu tahun dari 01/12/2010 hingga 09/12/2011. Berdasarkan latar belakang diatas maka didapatkan rumusan masalah yang dibahas dalam percobaan ini adalah bagaimana implementasi algoritma mean-shift clustering untuk menentukan segmentasi pasar pada penjualan toko online?

2 Metodologi

Clustering atau klasterisasi adalah sebuah metode untuk mengelompokan data. Clustering sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum [8]. Beberapa manfaat Clustering yang pertama merupakan metode segmentasi data yang sangat berguna dalam memprediksi dan menganalisa masalah bisnis tertentu, contohnya segmentasi pasar, marketing dan pemetaan zonasi wilayah. Kedua, dapat mengidentifikasi obyek dalam berbagai bidang seperti computer vision dan image processing.

Mean Shift termasuk kedalam kategori clustering algorithm dengan unsupervised learning yang menetapkan titik data ke cluster secara berulang dengan menggeser titik ke arah mode (mode adalah kepadatan titik data tertinggi di wilayah tersebut, dalam konteks Meanshift) [9]. Dengan demikian, ini juga dikenal sebagai mode-seeking algorithm. Mean shift algorithm memiliki aplikasi di bidang image processing dan computer vision [10]. Tidak seperti K-Means Cluester Algorithm yang populer, Mean-shift tidak memerlukan penentuan jumlah cluster terlebih dahulu. Jumlah cluster ditentukan oleh algoritma sehubungan dengan data.

3 Hasil dan Pembahasan

3.1 Pengumpulan Data

Pada percobaan kali ini kami menggunakan data transaksi online retail yang didalamnya memiliki data berupa nomor struk, kode stok, deskripsi, jumlah, tanggal struk, harga unit, id pelanggan, dan negara. Data yang berbentuk .csv dimasukkan ke dalam dataset. Setelah data dimasukkan kedalam dataset, dataset tersebut kemudian melalui proses penyaringan dimana baris yang memiliki data *null* akan dihilangkan. Dataset yang sudah melalui proses penyaringan memiliki data sebanyak 541909 baris dan 8 kolom. Contoh data yang digunakan dalam eksperimen ini terdapat pada Gambar 1.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

Gambar 1 Contoh Data

3.2 Pre-processing Data

Pada data terdapat kolom *CustomerID* memiliki data yang berbentuk *float*. Data tersebut harus diubah tipe datanya menjadi *String* agar proses pengolahan data dapat berjalan dengan baik dan untuk meminimalisir kesalahan dalam pemrosesan data. Setelah data tersebut sudah diubah menjadi *String*, kami membagi dataset diatas menjadi 3 bagian diantaranya:

1. Moneter (nilai dari total transaksi yang dilakukan)
2. Frekuensi (banyaknya transaksi yang dilakukan)
3. Resensi (banyaknya hari dari hari ini sampai tanggal pembelian terakhir)

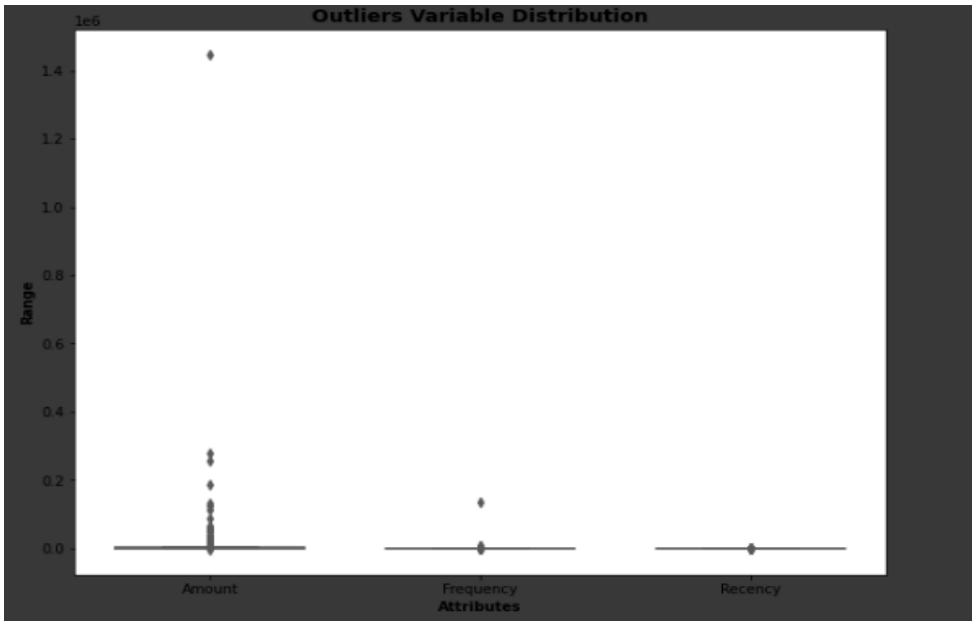
Gambar 2 adalah dataset yang sudah diubah menjadi tiga bagian yang sudah disebutkan diatas.

	CustomerID	Amount	Frequency	Recency
0	12346.0	0.00	2	325
1	12347.0	4310.00	182	1
2	12348.0	1797.24	31	74
3	12349.0	1757.55	73	18
4	12350.0	334.40	17	309

Gambar 2. Dataset yang sudah diproses menjadi 3 nilai yang dimiliki oleh CustomerID

Dataset yang sudah diproses memiliki data *outliers* yang harus dihilangkan. Data *outliers* dapat mempengaruhi akurasi dari proses *clustering* yang akan dilakukan seperti terdapat pada Gambar 3.

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)



Gambar 3. Distribusi *outliers* pada masing masing jenis data

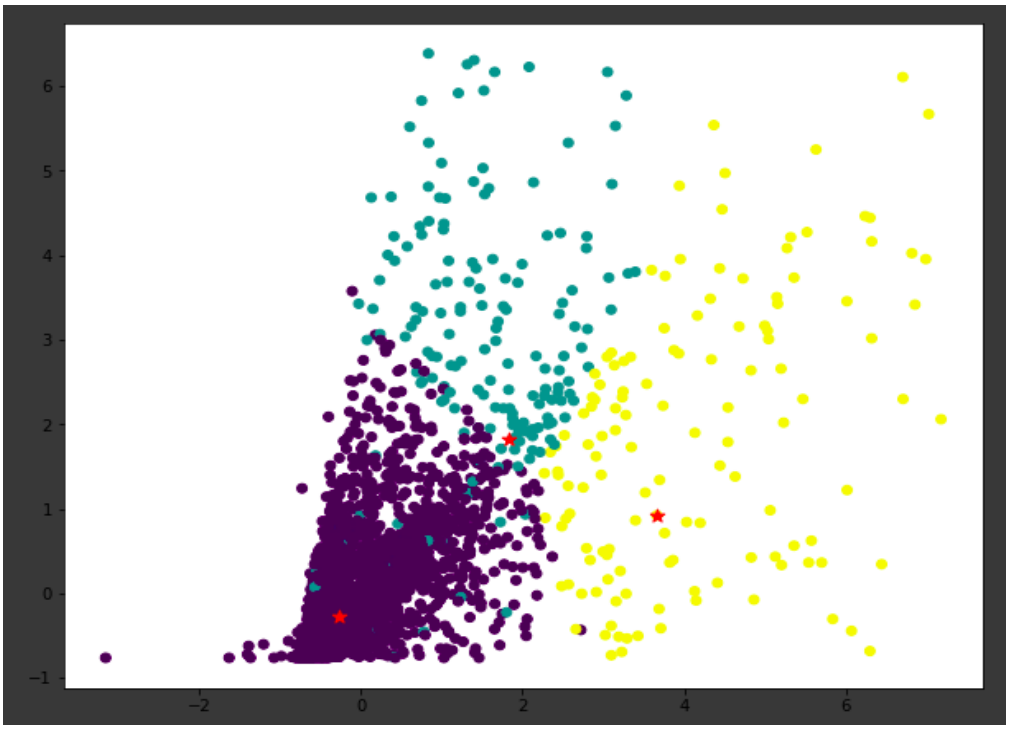
Dapat dilihat bahwa untuk jenis data *Amount* dan *Frequency* memiliki outlier dengan jarak yang cukup jauh. Sehingga data data tersebut perlu dipisahkan. Tetapi kami juga melakukan pemisahan outlier pada *recency* dengan harapan mendapatkan tingkat akurasi yang optimal. Perlu dilakukan scaling terhadap dataset diatar agar masing masing data memiliki tingkat prioritas yang sama. Gambar 4 adalah contoh data set hasil pre-processing.

	Amount	Frequency	Recency
0	-0.759639	-0.771795	2.295613
1	1.916220	1.117217	-0.910045
2	0.356175	-0.467454	-0.187782
3	0.331534	-0.026685	-0.741847
4	-0.552027	-0.614377	2.137309

Gambar 4. Dataset yang terdiri atas data Amount, Frequency, dan Recency yang sudah dilakukan scaling

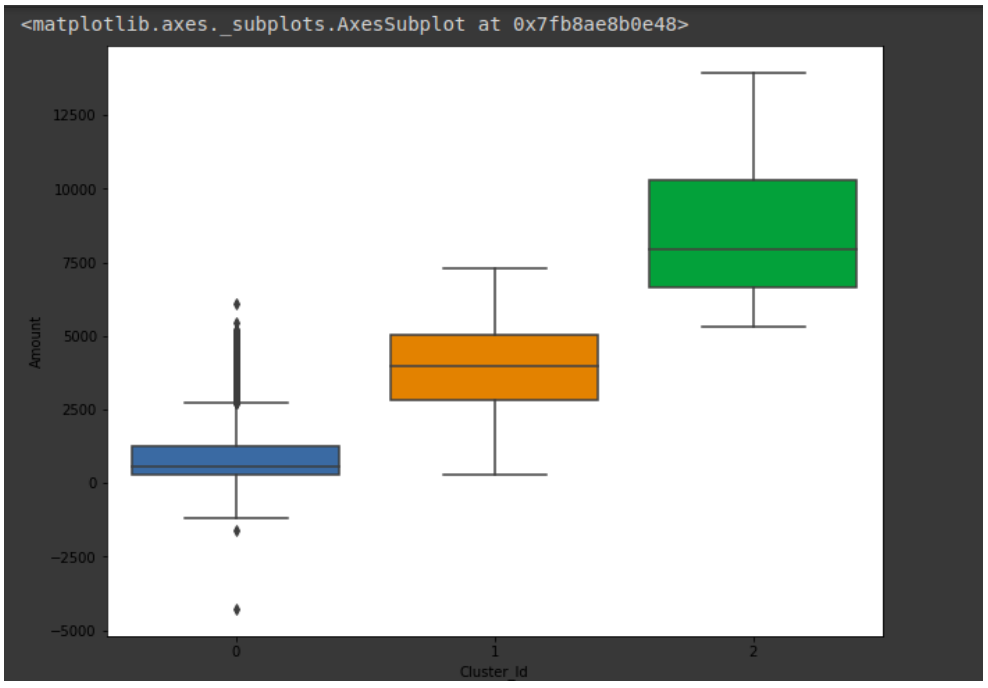
3.3 Implementasi Algoritma

Kami menghitung estimasi bandwidth yang akan digunakan oleh algoritma meanshit berdasarkan data yang sudah di pre-process dengan nilai treshold quantile = .4 dan n_samples = 500. Didapatkan nilai bandwidth sebesar 1.55. Setelah nilai bandwith didapatkan, nilai tersebut dijadikan input terhadap parameter bandwidth yang ada pada algoritma Mean Shift. Dari algoritma tersebut menghasilkan sebanyak tiga klaster data yang divisualisasikan dalam scatter plot berikut. Nilai klaster optimal yang sudah didapatkan dimasukkan kedalam fungsi meanshift. Hasil dari algoritma tersebut direpresentasikan menjadi 3 box plot. Masing masing box plot merepresentasikan satu jenis data (terdapat pada Gambar 5).



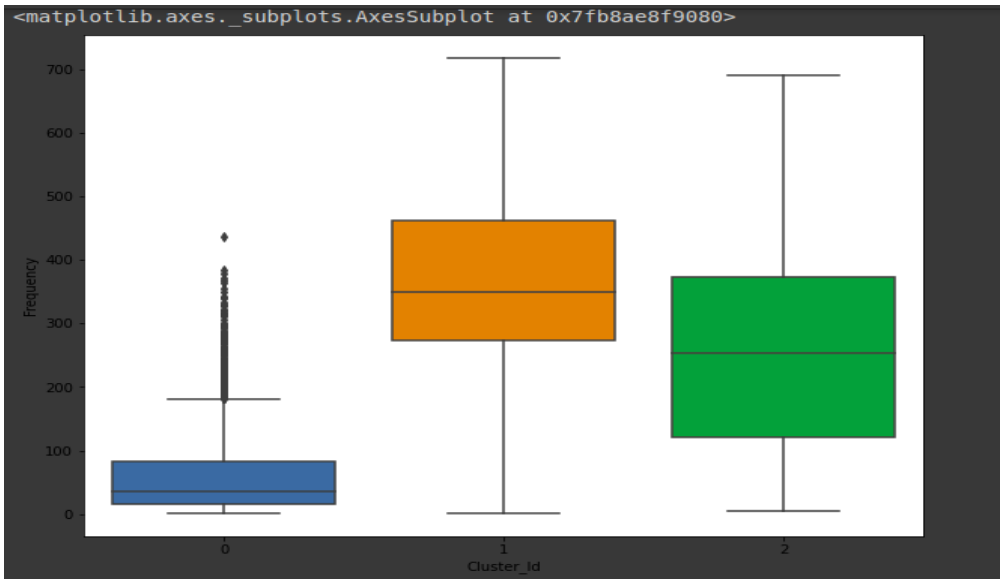
Gambar 5. Visualisasi *scatter plot* dari data hasil algoritma Mean Shift

Data dari hasil proses algoritma Mean Shift juga disajikan dalam bentuk box plot. Data ditampilkan berdasarkan jenisnya masing masing. Gambar 6-8 menunjukkan hasil klister pada masing-masing variable-nya.

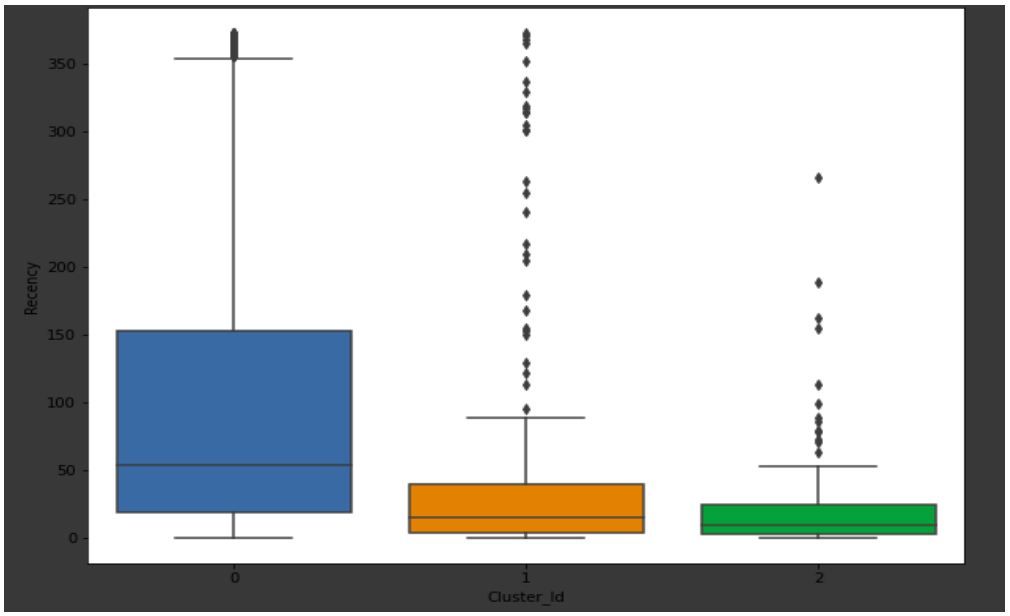


Gambar 6. Visualisasi Box Plot untuk data *Amount* pada masing masing klister

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)



Gambar 7. Visualisasi Box Plot untuk data *Frequency* pada masing masing klaster



Gambar 8. Visualisasi Box Plot untuk data *Recency* pada masing masing klaster

4 Simpulan

Berdasarkan percobaan kami mengenai implementasi algoritma mean shift untuk clustering. Dapat dilihat bahwa algoritma meanshift clustering untuk dataset ini dapat melakukan klasterisasi penjualan di sebuah toko online dengan baik, hal ini dapat dilihat dari hasil visualisasi data yang menunjukan tiga klaster yang terbentuk dengan masing - masing cluster_id 0, 1, dan 2. Dimana perbandingan cluster_id dan amount menggunakan boxplot model menunjukkan cluster tertinggi yakni cluster_id 2 dengan warna boxplot hijau, untuk perbandingan cluster_id dan frequency didapatkan cluster dengan frequeny transaksi tertinngi yakni cluster_id 1 dengan warna bloxplot orange, dan untuk hasil perbandingan cluster_id dengan recency didapatkanhasil cluster dengan recency tertinggi yakni cluster 0 dengan warna boxplot biru.

Referensi

[1] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.
[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction do Data Mining*. 2005.
[3] Y. Z. Zhao, H. Wang, and G. C. Yin, "Research on mean shift algorithm," in *Advanced Materials Research*, 2013, vol. 756–759, pp. 4021–4025, doi: 10.4028/www.scientific.net/AMR.756-759.4021.
[4] G. Simi Margarat and S. Sivasubramanian, "Basketball tracking using mean shift algorithm," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, pp. 339–344, 2019, doi: 10.35940/ijrte.C4159.098319.
[5] J. Liu, M. Wang, L. Kong, and X. Yang, "Through-wall tracking using mean-shift algorithm," in *2017 IEEE Radar Conference, RadarConf 2017*, 2017, pp. 1274–1277, doi: 10.1109/RADAR.2017.7944400.
[6] S. Bo and Y. Jing, "Image clustering using mean shift algorithm," in *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*, 2012, pp. 327–330, doi: 10.1109/CICN.2012.128.
[7] R. M. Awangga, S. F. Pane, K. Tunnisa, and I. S. Suwardi, "K means clustering and meanshift analysis for grouping the data of coal term in puslitbang tekmira," *Telkomnika*, vol. 16, no. 3, pp. 1351–1357, 2018.
[8] V. Kumar, P.-T. Tan, and M. Steinbach, "Cluster analysis: basic concepts and algorithms," in *Introduction to data mining*, 2006, pp. 488–568.
[9] S. Rao, L. Weifeng, J. C. Principe, and A. De Medeiros Martins, "Information theoretic mean shift algorithm," in *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, MLSP 2006*, 2006, pp. 155–160, doi: 10.1109/MLSP.2006.275540.
[10] D. Demirović, "An implementation of the mean shift algorithm," *Image Process. Line*, vol. 9, pp. 251–268, 2019, doi: 10.5201/ipol.2019.255.