

KOMPARASI METODE CLUSTERING K-MEANS, DBSCAN DAN HIERARCHICAL UNTUK ANALISIS PENYAKIT HEPATITIS C

Aditya Dwi B.^{1*}, Nur Irvan Rizqi², Syahrul Mubarak³, Dwi Rolliawati⁴

^{1,2,3,4}Sistem Informasi, Sains dan Teknologi, Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya, Indonesia

E-mail: ¹h96219036@student.uinsby.ac.id, ²h76219030@student.uinsby.ac.id, ³h06219016@student.uinsby.ac.id, ⁴dwi_roll@uinsby.ac.id

*Penulis Korespondensi

Abstract – In recent years, a disease that is often avoided by humans is hepatitis. Hepatitis is a disease that spreads to the human liver which becomes inflamed so that the function of the liver is stagnant. With the sluggish function of the liver, it will affect other organs in humans and result in blood flow not reaching the liver so that blood pressure becomes abnormal and blood vessels rupture. This study uses the clustering method with datasets contained in the UCI Repository and then compares the three algorithms namely Hierarchical, K-Means, and DBSCAN which aims to find out which algorithm is the best of the three algorithms. the results of this study indicate that the algorithm Hierarchical is the best with a value of 0.7779 without using any scaler.

Keywords: Hepatitis C, clustering, K-Means, Hierarchical, DB.Scan

Abstrak – Beberapa tahun terakhir ini penyakit yang sering dihindari manusia adalah hepatitis. Hepatitis yaitu penyakit yang menjangar di organ hati manusia yang mengalami peradangan sehingga fungsi daripada hati itu tersendat. Dengan tersendatnya fungsi organ hati maka akan mempengaruhi organ yang lain pada diri manusia dan berakibatkan aliran darah yang mengalir tidak sampai ke organ hati sehingga tekanan darah menjadi tidak normal dan pembuluh darah menjadi pecah. penelitian ini menggunakan metode klustering dengan dataset yang terdapat di UCI Repository lalu membandingkan ketiga algoritma yakni Hierarchical, K-Means, dan DBSCAN yang bertujuan untuk mengetahui mana algoritma terbaik dari ketiga algoritma tersebut. Hasil penelitian ini menunjukkan bahwa algoritma hierarchial mendapatkan hasil yang terbaik dengan nilai 0.7779 tanpa menggunakan *scaler*.

Kata Kunci: Hepatitis C, Metode Klustering, K-Means, Hierarchical, DB.Scan

PENDAHULUAN

Pada saat ini perkembangan teknologi semakin cepat dan pesat sehingga berkembang juga keahlian dalam mengumpulkan serta mengolah data yang didapat [1]. penggunaan informasi dan pengetahuan yang tersimpan di data tersebut pada sekarang ini disebut *data mining* [2]. dengan memanfaatkan data mining yang didukung oleh data yang sesuai maka untuk mengelompokkan suatu data menjadi lebih mudah. Pada sektor bidang pendidikan, pemerintahan dan juga kesehatan saat ini sudah mulai memanfaatkan data mining terutama dalam sektor bidang kesehatan yang dibuat untuk mengelompokkan pasien yang terpapar penyakit sesuai diagnosa.

Beberapa tahun terakhir ini penyakit yang sering dihindari manusia adalah hepatitis. Hepatitis yaitu penyakit yang menjangar di organ hati manusia yang mengalami peradangan sehingga fungsi daripada hati itu tersendat [3]. Dengan tersendatnya fungsi organ hati maka akan mempengaruhi organ yang lain pada diri manusia dan berakibatkan aliran darah yang mengalir

tidak sampai ke organ hati sehingga tekanan darah menjadi tidak normal dan pembuluh darah menjadi pecah [4]. Hepatitis ada beberapa macam yakni A, B, C, D dan E. Hepatitis A dan E merupakan kondisi yang sangat luar biasa dikarenakan dapat menular secara *fecal oral* atau berhubungan dengan kepribadian yang bersih (PHBS) lalu untuk Hepatitis B, C, dan D jarang bisa ditularkan karena sudah menjadi kronis dan kanker hati, sebanyak 2 miliar orang didunia telah terdiagnosa penyakit jenis hepatitis B dan sekitar 240 juta lainnya mengidap penyakit B kronik, lalu penderita Hepatitis C diperkirakan sebanyak 170 juta orang, berdasarkan data yang ada 1,5 Juta orang di bumi setiap tahunnya meninggal karena penyakit hepatitis [5].

Berdasarkan latar belakang yang telah dipaparkan maka ilmu pengetahuan tentang data mining diperlukan karena pada saat ini digunakan penelitian untuk mengelompokkan data penyakit Hepatitis C sesuai penyebabnya dengan hasil tes darah yang sudah dijalani. Dengan itu metode yang dipakai yaitu metode *clustering*

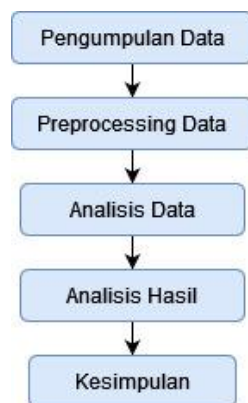
dikarenakan metode *clustering* dapat dipergunakan untuk mengelompokkan suatu data. pada penelitian ini menggunakan metode *clustering* dengan algoritma K-Means, DB.Scanner dan Hierarchical yang mana nanti hasilnya akan dikomparasikan sesuai tingkat Akurasi.

Terdapat penelitian terdahulu yang menggunakan dataset Hepatitis C, pertama hasil penelitian dari septian [6] menunjukkan bahwa perbandingan metode klasifikasi menggunakan algoritma C4.5 dan algoritma naive bayes dalam memprediksi penyakit hepatitis C kedua algoritma tersebut akurat akan tetapi tingkat akurasi tertinggi dimiliki algoritma C4.5 dengan mempunyai nilai AUC 0.846. kedua yakni hasil penelitian dari dinata dkk [7] menunjukkan dataset hepatitis C menggunakan metode reduksi atribut dari *information gain* berhasil memperbaiki cluster k-means. terakhir yakni penelitian dari saputra dan chusyairi [8] menunjukkan bahwa algoritma *Fuzzy C-Means Clustering* merupakan algoritma yang optimal dalam mengelompokkan data imunisasi bayi dan anak.

Oleh karena itu pada penelitian yang dilakukan saat ini yaitu dengan melakukan komparasi dari 3 metode yaitu *K-means*, *DBScan*, dan *Hierarchical*. Clustering pada Dataset Hepatitis C menunjukkan bahwa metode *Hierarchical* dapat melakukan *Clustering* dengan performa tertinggi.

METODOLOGI PENELITIAN

Penelitian ini menggunakan *software Knime* dalam melakukan clustering data dan menggunakan metode K-means, DBScan, dan Hierarchical dalam melakukan analisis data. Dataset yang digunakan adalah data sekunder yang bersumber dari *kaggle* yaitu *Hepatitis C Prediction Datasets*. Dataset sekunder adalah data yang telah tersedia dan dapat digunakan kembali [9]. Hasil perhitungan dari ketiga metode ini dibandingkan untuk mencari metode mana yang memiliki tingkat keakuratan yang lebih tinggi.



Gambar 1. Tahapan Penelitian

Berikut adalah deskripsi mengenai tahapan penelitian pada gambar diatas.

1. Pengumpulan Data

Pengumpulan Data Dilakukan dengan menggunakan data sekunder yaitu dataset dari Kagie data *Hepatitis C Prediction*.

2. Preprocessing Data

Preprocessing data pada setiap metode analisis Dalam pre-processing, data dipersiapkan [17]. kluster disamakan yaitu dengan menggunakan *node cell replacer*, *missing value*, *string to number*, dan *normalizer* yang setiap kegunaannya akan dibahas lebih lanjut pada bagian hasil dan pembahasan. namun pada analisis hierarchical ditambahkan numeric distance untuk menentukan jarak *Euclidean* pada kolom numerik.

3. Analisis Data pada penelitian Komparasi Metode Clustering data ini menggunakan 3 metode yaitu Metode K-Means, DBScan, dan Hierarchical Analisis Clustering metode K-Means, Analisis Clustering metode DBScan Analisis Clustering metode Hierarchical

4. Analisis Hasil

5. Kesimpulan

K Means yaitu metode mengelompokkan data non struktur yang dipakai untuk mengaitkan subjek data ke dalam kelompok, Setiap data pada kelompok mempunyai jarak dekat antar centroidnya [10]. dalam menentukan alur algoritma K means hal yang perlu diperhatikan sebagai berikut:

1. Menentukan Jumlah clustering
2. Mendistribusikan data ke kelompok secara random
3. Menghitung inti metode clustering dari data yang diperoleh dari tiap tiap metode clustering.
4. Mendistribusikan tiap data ke centroid terdekat
5. kembali ke alur ke 3 jika ada data yang bergeser atau ada nilai centroid yang berubah pada setiap nilai yang telah ditentukan
6. selesai

jika menentukan *centroid* pada tiap kelompok harus diambil rata rata (*Mean*) di semua nilai data yang ada di dalam fiturnya. jika nilai *M* menyatakan banyaknya data pada setiap kelompok maka *i* menyatakan fitur ke *i* di sebuah kelompok [11]. rumus menghitung centroid sebagai berikut:

$$C_i = \frac{1}{M} \sum_{j=1}^M x_j$$

menghitung jarak antar titik terpendek

$$D=(x_1,x_2)=\sqrt{\sum_{j=1}^p |x_{2j} - x_{1j}|^2}$$

mendistribusikan titik keanggotaan

$$a_{ji} = \begin{cases} 1, & d = \min(D(x_j, C_i)) \\ 0, & \text{lainnya} \end{cases}$$

Fungsi objektif

$$F = \sum_{j=1}^N \sum_{i=1}^K a_{ji} D(x_j, C_i)$$

DBScan merupakan metode yang berfungsi untuk mengoptimalkan area dengan kerapatan yang cukup tinggi ke dalam cluster serta mengumpulkan cluster berbentuk acak ke dalam database spasial yang memuat *noise* [12]. prinsip dalam menggunakan algoritma DBScan sebagai berikut [13]:

1. *Neighborhood* yang terdapat di dalam radius E dikatakan *e neighborhood* dari suatu objek data
2. Jika *e neighborhood* dari objek adalah jarak terpendek dalam cluster maka objek tersebut dikatakan objek Core
3. sebuah objek p dinamai density reachable dari objek q dengan respek ϵ dan minimal objek dalam suatu set objek D jika terjadi rantai objek P_i P_i P_{i+1} dimana $p_i = q$ dan $p_{i+1} = p_i$, di mana P density reachable secara langsung dari p dengan respek ke C dan $MinPts$, untuk $1 \leq i \leq n$. p . anggota D .
4. Suatu objek p adalah *density connected* ke objek q dengan respek ke e jika menemui suatu objek o maka anggota itu D yang mana objek p & q itu *density reachable* dari o dengan respek ke C dan $MinPts$

Hierarchical *clustering* merupakan metode yang berfungsi untuk mengelompokkan suatu document clustering, teknik ini dapat menghasilkan suatu partisi yang terstruktur dimana kumpulan tersebut termuat *cluster* yang memiliki nilai individual serta ada sebuah *cluster* yang mempunyai nilai yang punya cluster didalamnya yang dinamakan *single cluster* [14].

Adapun dataset yang kami gunakan adalah *Hepatitis C virus - Blood based Detection* yaitu data kandungan dalam darah dari donor darah, pasien hepatitis, pasien hepatitis dengan *fibrosis* dan pasien dengan *sirosis* [16].

Dalam dataset tersebut, terdapat 14 parameter yang ada di dalamnya. Penjelasan mengenai parameter tersebut dapat dilihat pada Tabel 1.

Tabel 1. Parameter Dataset

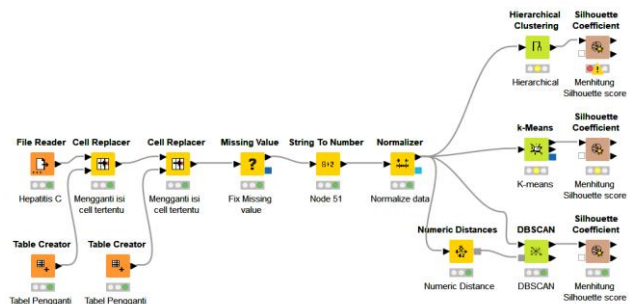
No	Parameter	Deskripsi
1	Id	Nomor urutan data
2	Category	Kategori penyakit
3	Age	Umur pasien
4	Sex	Jenis kelamin Pasien
5	ALB	Kadar albumin Pasien
6	ALP	Kadar alkalin fosfat Pasien
7	ALT	Kadar alanine transaminase Pasien
8	AST	Kadar aspartate aminotransferase Pasien

9	BIL	Kadar bilirubin Pasien
10	CHE	Kadar cholinesterase Pasien
11	CHOL	Kadar Kolesterol Pasien
12	CREA	Kadar creatinin Pasien
13	GGT	Kadar gamma-glutamyl transferase Pasien
14	PROT	Kadar Protein Pasien

HASIL DAN PEMBAHASAN

Proses analisis komparasi algoritma dilakukan dengan menggunakan program, KNIME. KNIME (*Konstanz Information Miner*) merupakan sebuah platform yang dapat digunakan untuk mengintegrasikan, memproses, dan menganalisis data dari berbagai macam sumber. Dengan KNIME pengguna dapat membangun alur kerja analisis data langsung tanpa harus menuliskan kode [15]. Dalam Rumpun IT, Algoritma merupakan tahapan dalam melakukan perhitungan atau sebagai pemecahan suatu masalah yang di kemukakan secara runtut [18].

Proses analisis dilakukan dengan memodelkan setiap metode clustering untuk memberikan gambaran alur menggunakan *node* yang telah disediakan dan menjalankannya.









Gambar 2. Workflow KNIME

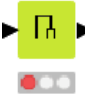




Gambar 2 merupakan tampilan dari workflow pada KNIME yang diawali dengan membaca dataset dengan node CSV Reader. Sebelum proses dimulai, pre-processing perlu dilakukan terlebih dahulu dengan menggunakan node Cell replacer yang akan mengganti value dari cell sesuai dengan table yang telah ditentukan, setelah itu node Missing value digunakan untuk mengisi cell atau data yang kosong, node string to number digunakan untuk mengubah data angka yang terbaca sebagai string menjadi angka atau number lalu yang terakhir ada node Normalizer yang digunakan untuk menormalisasi data. Pada algoritma DBSCAN, terdapat satu node tambahan yang diperlukan sebelum dapat memproses data, yaitu Numeric Distances untuk menentukan jarak Euclidean dari data numerik yang ada.

Selanjutnya, analisis dilakukan menggunakan tiga node berbeda sesuai dengan algoritma clustering terkait. Dari masing-masing algoritma tersebut, diperlukan validasi menggunakan *node Silhouette Coefficient*.

Penjelasan dari masing-masing node yang digunakan pada workflow untuk penelitian ini dapat dilihat pada Tabel 2 berikut.

Tabel 2. Node yang digunakan

Nama	Node	Deskripsi
CSV Reader		Untuk membaca file CSV
Table Creator		Untuk membuat tabel
Cell Replacer		Untuk mengganti value dari cell tertentu
Missing Value		Untuk mengisi cell kosong
String to Number		Untuk mengubah data string menjadi number
Normalizer		Untuk menormalisasi nilai dengan format numerik

Hierarchical Clustering		Untuk mengelompokkan data yang diinput berdasarkan hirarki yang terbentuk
K-Means		Untuk menjalankan clustering menggunakan algoritma K-means
Numeric Distances		Untuk menentukan jarak Euclidean pada kolom numerik
DBSCAN		Untuk menjalankan clustering menggunakan algoritma DBSCAN
Silhouette Coefficient		Untuk menghitung koefisien silhouette berdasarkan jumlah klaster yang telah ditemukan

Tabel 3. Silhouette Score Hierarchical

Preprocessing	Cluster					
	2	3	4	5	6	7
Tanpa Scaler	0.7779	0.6983	0.5788	0.5476	0.5306	0.4608
min-max scaler	0.5190	0.4178	0.5315	0.5175	0.4816	0.4743
Z-score	0.7565	0.7567	0.7092	0.6152	0.6065	0.5306
Decimal Scaling	0.7170	0.6894	0.6413	0.4109	0.475	0.4765

Tabel 3 menunjukkan silhouette score untuk metode Hierarchical dari beberapa jumlah cluster, mulai dari 2 cluster hingga 7 cluster. Perolehan silhouette score tertinggi dari masing-masing perlakuan pre-processing

data pada metode Hierarchical berturut-turut adalah 0.7779 untuk pre-processing tanpa scaler, 0.5315 untuk Min-max scaler, 0.7567 untuk z-score normalization, dan 0.7170 untuk decimal scaling. Semua perolehan skor

tertinggi terjadi pada cluster 2. Dari keseluruhan skor pada masing-masing proses Pre-processing tersebut, skor tertinggi yang diperoleh pada metode Hierarchical

clustering adalah 0.7779 yang terdapat pada Pre-processing tanpa *scaler*.

Tabel 4. Silhouette Score K-Means

Preprocessing	Cluster					
	2	3	4	5	6	7
Tanpa Scaler	0.5343	0.4366	0.4227	0.3996	0.3526	0.3501
min max scaler	0.5358	0.5751	0.4273	0.4042	0.3596	0.3342
Z-score	0.1854	0.2032	0.1584	0.1574	0.1662	0.1439
Decimal Scaling	0.7169	0.5998	0.4304	0.4693	0.4472	0.3358

Tabel 4 menunjukkan silhouette score untuk metode clustering K-Means dari beberapa cluster, mulai dari 2 cluster hingga 7 cluster. Adapun perolehan silhouette score tertinggi dari masing-masing pre-processing data pada metode K-Means berturut-turut adalah 0.5343 untuk pre-processing tanpa scaler, 0.5751 untuk Min-Max scaler, 0.2032 untuk Z-score normalization, dan 0.7169 untuk Decimal scaling. Perolehan skor tertinggi terjadi pada cluster 2 untuk perlakuan tanpa scaler dan Decimal scaling, sedangkan perolehan skor tertinggi pada Z-score dan Min Max Scaler terjadi pada cluster 3. Dari keseluruhan skor pada masing-masing Pre-processing

tersebut, skor tertinggi yang diperoleh pada metode K-Means adalah 0.7169 pada dengan Decimal scaling.

Metode DBSCAN memerlukan dua parameter input sebelum proses *clustering* dilakukan, yaitu Epsilon dan Minimum Points. Epsilon adalah jarak maksimal antara dua data dalam satu cluster yang diizinkan, dan minimum points adalah banyaknya data minimal dalam jarak epsilon agar terbentuk suatu cluster. Metode jarak yang digunakan dalam DBSCAN adalah *Euclidean Distance* atau jarak euclidian. Pada penelitian ini kami menggunakan epsilon.

Tabel 5. Silhouette Score DBSCAN

Preprocessing	Minimum Points									
	10		15		20		25		30	
	Cluster	Score	Cluster	Score	Cluster	Score	Cluster	Score	Cluster	Score
Tanpa Scaler	1	0	1	0	1	0	1	0	1	0
min max scaler	14	0.5358	14	0.5342	14	0.5342	14	0.5342	14	0.5336
Z-score	1	0	1	0	1	0	1	0	1	0
Decimal Scaling	14	0.7169	14	0.7169	14	0.7169	14	0.7169	14	0.7169

Dari tabel 5 dapat diketahui bahwa hasil metode DBSCAN dengan Decimal Scaling dan dengan menggunakan 14 cluster pada studi kasus ini memiliki nilai Silhouette Score paling tinggi. Min-max scaler dan tanpa scaler tidak memiliki nilai Silhouette karena scaler tersebut hanya menghasilkan 1 cluster, sedangkan syarat validasi *Silhouette Score* adalah memiliki setidaknya 2 cluster.

Tabel 6. Komparasi Hasil Silhouette Score

Metode	scaler	Cluster	Score
K-Means	Decimal scaling	2	0.7169
Hierarchical	Tanpa Scaler	2	0.7779
DBSCAN	Standard	14	0.7169

Scaler

KESIMPULAN

Hasil penelitian dengan skenario K-Means menunjukkan *silhouette score* dengan 2 cluster menghasilkan nilai tertinggi; Decimal scaling dengan score 0.7169. Untuk Hierarchical dengan 2 cluster tanpa scaler dengan score 0.7779. Maka performa terbaik untuk analisis terhadap data hepatitis c menggunakan darah adalah dengan menggunakan algoritma Hierarchical Clustering. Hal ini dibuktikan dengan perolehan *silhouette score* yang dihasilkan oleh Hierarchical selisih cukup tipis dengan K-Means di posisi teratas.

DAFTAR PUSTAKA

- [1] A. Sudirman *et al.*, *Sistem Informasi Manajemen*. Yayasan Kita Menulis, 2020.
- [2] N. L. W. S. R. Ginantra *et al.*, *Data mining dan penerapan algoritma*. Yayasan Kita Menulis, 2021.
- [3] H. Tandra, *VIRUS CORONA BARU COVID-19: Kenali, Cegah, Lindungi Diri Sendiri & Orang Lain*. Rapha Publishing, 2021.
- [4] Buku ajar ilmu penyakit hati, Ali Sulaiman, Nurul Akbar, Laurentius A. Lesmana, Sjaifoellah Noer, Perpustakaan Nasional RI <https://opac.perpusnas.go.id/DetailOpac.aspx?id=1118609> (accessed 2022-11-27 23:10:42)
- [5] “Dinas Kesehatan Kabupaten Mojokerto.” <http://dinkes.mojokertokab.go.id/artikel/hepatitis> (accessed Nov. 27, 2022).
- [6] W. D. Septiani, “Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis.” *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.
- [7] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, “Reduksi atribut menggunakan information gain untuk optimasi cluster algoritma k-means,” *J Edukasi Dan Penelit Inf.*, vol. 6, no. 1, pp. 48–53, 2020.
- [8] P. R. N. Saputra and A. Chusyairi, “Perbandingan Metode Clustering dalam Pengelompokan Data Puskesmas pada Cakupan Imunisasi Dasar Lengkap,” *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 4, no. 6, pp. 1077–1084, 2020.
- [9] M. Abdurahman, “Sistem Informasi Data Pegawai Berbasis Web Pada Kementerian Kelautan Dan Perikanan Kota Ternate,” *J. Ilm. Ilk. - Ilmu Komput. Inform.*, vol. 1, no. 2, Art. no. 2, Jul. 2018, doi: 10.47324/ilkominfo.v1i2.10.
- [10] N. Indriani Widiastuti, “K Means,” 2017.
- [11] D. A. Pratiwi, R. M. Awangga, and M. Y. H. Setyawan, *Seleksi Calon Kelulusan Tepat Waktu Mahasiswa Teknik Informatika Menggunakan Metode Naïve Bayes*, vol. 1. Kreatif, 2020.
- [12] P. Silitonga, “Analisis Pola Penyebaran Penyakit Pasien Pengguna Badan Penyelenggara Jaminan Sosial (Bpjs) Kesehatan Dengan Menggunakan Metode Dbscan Clustering,” *J. Times*, vol. 5, no. 1, pp. 36–39, 2016.
- [13] C. D. Manning, P. Raghavan, and H. Schütze, “Probabilistic information retrieval,” *Introd. Inf. Retr.*, pp. 220–235, 2009.
- [14] G. A. Pradnyana and N. A. S. ER, “Perancangan Dan Implementasi Automated Document Integration Dengan Menggunakan Algoritma Complete Linkage Agglomerative Hierarchical Clustering,” *J. Ilmu Komput.*, vol. 5, no. 2, pp. 1–10, 2012.
- [15] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold, “KNIME for reproducible cross-domain analysis of life science data,” *J. Biotechnol.*, vol. 261, pp. 149–156, Nov. 2017, doi: 10.1016/j.jbiotec.2017.07.028.
- [16] “Hepatitis C virus - Blood based Detection.” <https://www.kaggle.com/datasets/amritpal333/hepatitis-c-virus-blood-biomarkers> (accessed Dec. 26, 2022).
- [17] “Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine | Journal Information System Development (ISD).” <https://ejournal-medan.uph.edu/index.php/ISD/article/view/381> (accessed Feb. 10, 2023).
- [18] “Perbandingan Metode Moving Average (Ma) Dan Neural Network Yang Berbasis Algoritma Backpropagation Dalam Prediksi Harga Saham | Journal Information System Development (ISD).” <https://ejournal-medan.uph.edu/index.php/ISD/article/view/372> (accessed Feb. 10, 2023).