Penggunaan Algoritma K-Means Pada Metode Clustering Untuk Menganalisa Tindak Kriminal

Imam Zuhdi Muzakkiy[-1], Khoirul Husein[-2], Kelfin Antonius[-3], Kevin Raihan Hidayat[-4], El Emir Di Haryanto[-5], Iman Paryudi[-6]
Program Studi Teknik Informatika
Fakultas Teknik Universitas Pancasila
Jakarta, Indonesia
4521210010@univpancasila.ac.id[1], 4521210007@univpancasila.ac.id[2], 4521210018@univpancasila.ac.id[3],
4521210025@univpancasila.ac.id[4], 4521210031@univpancasila.ac.id[5], iman.paryudi@univpancasila.ac.id[6]

Abstract — Kriminalitas adalah bentuk tindakan yang merugikan secara ekonomis maupun psikologis dan melanggar hukum yang berlaku di suatu negara. Dapat diartikan juga, tindakan kriminalitas adalah segala sesuatu yang melanggar norma-norma sosial, sehingga terdapat pertentangan dari masyarakat. Studi ini bertujuan untuk mengetahui tingkat kriminalitas pada suatu daerah menggunakan metode K-Means clustering dengan menggunakan perangkat lunak Orange Data Mining Tool. Dengan dibuatnya sistem ini, diharapkan dapat membantu aparat keamanan setempat dalam menentukan daerah rawan kriminal dan meningkatkan keamanan pada daerah rawan tersebut agar dapat mencegah dan mengurangi dampak serta akibat tindakan kriminal. Dan dari hasil studi yang dilakukan, akan menghasilkan pengelompokan daerah rawan kriminal.

Kata Kunci: *Kriminalitas, Orange Data Mining Tool, K-Means Clustering, Data Mining*

I. PENDAHULUAN

Pemeliharaan keamanan dalam masyarakat dipercayakan kepada polisi, sehingga membuat mereka bertanggung jawab untuk pencegahan kejahatan. Untuk melaksanakan tugas ini secara efektif, polisi harus memiliki pengetahuan dan pemahaman yang komprehensif tentang berbagai kegiatan kriminal yang lazim di masyarakat. Pengetahuan ini diperoleh dari data terverifikasi kasus-kasus kriminal masa lalu, yang menjadi dasar bagi polisi untuk memaksimalkan upaya pencegahan kejahatan mereka.

Tujuan dalam penelitian ini yaitu untuk menganalisis dampak dari banyaknya individu yang terlibat dalam konflik dengan polisi di Amerika Serikat. Dataset ini diperoleh dari Data World.

Clustering adalah proses pengelompokan objek data bersama-sama dengan cara yang menekankan kesamaan sekaligus membedakannya dari kelompok lain. Dalam hal ini algoritma yang digunakan adalah algoritma K-Means clustering.

Metode K-Means clustering digunakan untuk mengorganisasikan benda-benda ke dalam kelompok-kelompok berdasarkan kesamaannya. Dibutuhkan banyak hal dan menempatkannya ke dalam kelompok yang berbeda berdasarkan rata-rata atau rata-ratanya. Metode ini mencoba membagi sesuatu menjadi dua kelompok atau lebih. Ini membantu penulis memahami bagaimana berbagai hal berbeda satu sama lain.

II. CLUSTERING

Menurut Suriani[1] clustering adalah metode yang berguna untuk mengidentifikasi dan mengkategorikan data yang memiliki karakteristik serupa. Termasuk dalam kategori penambangan data tanpa pengawasan, artinya tidak memerlukan panduan atau masukan eksternal.
Menurut Dani[2] uji validitas sebagai berikut:

A. Koefisien Silhouette

Metode koefisien silhouette adalah cara untuk mengevaluasi seberapa baik sebuah cluster diatur. Evaluasi ini penting untuk menentukan seberapa akurat data dikelompokkan. Untuk menghitung koefisien silhouette, langkah-langkah spesifik harus diikuti:

1) Temukan jarak rata-rata antara objek i dan semua objek lain dalam grup yang sama.
2) Hitung jarak rata-rata antara setiap objek dan semua data dalam berbagai kelompok, dan pilih nilai minimum untuk setiap objek.
3) Setelah itu, tentukan pengukuran numerik dari koefisien silhouette.

Nilai koefisien siluet dapat berkisar dari -1 hingga 1, dan nilai 1 menunjukkan bahwa objek ke-i berada di cluster yang benar, sehingga hasil clustering akurat. Ketika nilai koefisien siluet adalah 0, tidak pasti apakah objek ke-i milik cluster U atau cluster V karena terletak di antara dua cluster.

Nilai koefisien siluet menunjukkan kualitas struktur cluster, dan nilai -1 berarti cluster yang berbeda mungkin lebih cocok untuk objek tertentu. Kauffman dan Rousseuw (1990) membuat tabel dengan kriteria subyektif untuk pengelompokan kualitas berdasarkan skor siluet.

Tabel 1. Kriteria subjektif kualitas pengelompokkan berdasarkan koefisien silhouette

Nilai Silhouette	Interpretasi
0,71-1,00	Strong Cluster

0,51-0,70	Good Cluster
0,26-0,50	Weak Cluster
0,00-0,25	Bad Cluster

B. Algoritma K-Means

1. Pengertian Algoritma K-Means

   K-Means adalah algoritma canggih untuk menambang data yang dapat mengkategorikan dan mengatur data secara efisien. Ada berbagai metode untuk membentuk klaster, seperti menetapkan pedoman yang menentukan anggota kelompok mana yang menjadi anggota berdasarkan distribusi yang adil di antara bagian-bagiannya[3].

2. Langkah Algoritma K-Means

   Langkah-langkah dari algoritma K-Means yaitu[4][5]:

   1) Tentukan berapa k-cluster yang ingin dibentuk.
   2) Menghasilkan nilai acak untuk pusat cluster awal (titik tengah) hingga k cluster.
   3) Hitung jarak setiap data masukan dari setiap centroid dengan menggunakan rumus jarak (Euclidean distance) hingga ditemukan jarak terdekat dari centroid setiap data.

      Persamaan jarak Euclidean adalah:

      $$d(xi, \mu i) = ((xi - \mu i)2$$

   4) Urutkan setiap datum menurut seberapa dekat dengan centroid (jarak minimum).
   5) Perbarui nilai titik tengah. Nilai centroid baru diperoleh dari cluster rata-rata tersebut dengan rumus berikut:

      $$Ck = 1nk \sum di$$

      Di mana:
      nk = jumlah data dalam cluster
      di = jumlah nilai jarak yang terdapat pada setiap cluster
   6) Ulangi langkah 2-5 sampai tidak ada perubahan pada anggota tiap cluster.
   7) Setelah langkah 6 selesai, rata-rata pusat cluster (μj) pada iterasi terakhir digunakan sebagai parameter untuk menentukan klasifikasi data.

III. DATA MINING

A. Pengertian Data Mining

Data mining adalah komponen dari proses pengungkapan informasi dari Knowledge Discovery dalam basis data Database[6][7].

B. Metode Data Mining

Menurut Mardi[8] metode data mining dikategorikan ke dalam berbagai kelompok sesuai dengan tugas yang dapat diselesaikan, yaitu:

1. Deskripsi (Description)

   Terkadang, peneliti dan analis mungkin ingin mengeksplorasi metode untuk menggambarkan pola dan tren yang diamati dalam data.
2. Estimasi (Estimation)

   Estimasi dan klasifikasi memiliki kesamaan, tetapi estimasi melibatkan pendekatan yang lebih numerik untuk menentukan variabel target.
3. Prediksi (Prediction)

   Prediksi, sebuah proses yang mengantisipasi hasil di masa depan, memiliki kesamaan dengan klasifikasi dan estimasi. Metode dan teknik yang diterapkan dalam klasifikasi dan estimasi memiliki potensi untuk prediksi dalam konteks yang sesuai.
4. Klasifikasi (Classification)

   Dalam bidang klasifikasi, komponen penting adalah identifikasi variabel kategori yang diminati. Misalnya, ketika memeriksa klasifikasi pendapatan, penting untuk membedakan antara kategori pendapatan tinggi, sedang, dan rendah.
5. Pengklusteran (Clustering)

   Clustering melibatkan proses mengkategorikan data, pengamatan, atau item berdasarkan kesamaannya, membentuk kelompok atau kelas. Cluster adalah kumpulan data yang memiliki kesamaan karakteristik di dalam grup, tetapi tidak dengan cluster lain.
6. Asosiasi (Association)

   Data mining melibatkan proses asosiasi, yang bertujuan untuk mengidentifikasi atribut yang sering muncul bersamaan.

C. Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) adalah serangkaian proses untuk menemukan informasi yang berguna dari data. KDD terdiri dari serangkaian langkah transformasi, termasuk data preprocessing dan juga post processing.

Data processing adalah proses mengubah data mentah menjadi format yang sesuai untuk langkah analisis selanjutnya. Selain itu data preprocessing juga digunakan untuk mengidentifikasi atribut dan segmen data yang penting untuk data mining.

Istilah Data mining dan Knowledge Discovery in Databases (KDD) sering digunakan secara bergantian untuk menggambarkan proses penggalian informasi tersembunyi dari database besar. Sebenarnya, kedua istilah ini memiliki arti yang berbeda, tetapi keduanya terkait. Dan salah satu langkah dalam keseluruhan proses KDD adalah Data mining[9].

D. Tahapan Proses KDD Dalam Data Mining

Menurut Muslim *et al*[10] berikut adalah tahapan proses KDD dalam data mining:

1. Data Cleaning

   Data cleaning adalah proses pembersihan data melibatkan penghapusan data yang tidak relevan atau tidak konsisten untuk menghilangkan kebisingan yang tidak diinginkan.

2. Data Integration
   Integrasi data melibatkan penggabungan data dari beberapa database untuk membuat database baru. Data yang diperlukan untuk data mining tidak terbatas pada satu basis data dan dapat bersumber dari banyak basis data.
3. Data Selection
   Basis data hanya akan menggunakan informasi yang diperlukan dan relevan, dan tidak akan mengambil atau menganalisis data yang tidak perlu.
4. Data Transformation
   Transformasi data melibatkan konversi dan penggabungan data ke dalam format tertentu, yang diperlukan untuk penambangan data yang efektif.
5. Data Mining
   Proses mining dapat disebut juga sebagai proses penambangan data, yang merupakan metode utama yang digunakan untuk mengungkap informasi penting yang tersembunyi di dalam data.
6. Pattern Evaluation
   Evaluasi pola adalah proses menemukan pola yang signifikan dalam basis pengetahuan. Ini melibatkan menghasilkan pola unik dari model klasifikasi dan mengevaluasinya untuk menentukan apakah itu mengkonfirmasi hipotesis yang ada.
7. Knowledge Presentation
   Knowledge presentation merupakan melibatkan penyajian informasi tentang metode yang digunakan untuk mengumpulkan pengetahuan, seperti yang dieksplorasi oleh pengguna.

E. Tools Dalam Data Mining

1. Orange
   Menurut indriyanti *et al*[11] Orange adalah jenis perangkat lunak sumber terbuka yang dapat digunakan untuk penambangan data dan pembelajaran mesin. Memungkinkan untuk analisis dan visualisasi data eksplorasi, dan dapat digunakan untuk berbagai tujuan seperti pemodelan prediktif dan sistem rekomendasi. Selain itu, memiliki aplikasi di bidang-bidang seperti penelitian genomik, bioinformatika biomedis, dan pendidikan.
2. Weka
   Menurut Faid *et al*[12] Weka adalah Weka adalah perangkat lunak pembelajaran mesin yang ditulis dalam Java dan dikembangkan di University of Waikato di Selandia Baru. Perangkat lunak ini memiliki banyak algoritma pembelajaran mesin untuk penambangan data. Weka juga memiliki banyak tools untuk pengolahan data, mulai dari preprocessing, klasifikasi, aturan asosiasi dan visualisasi.

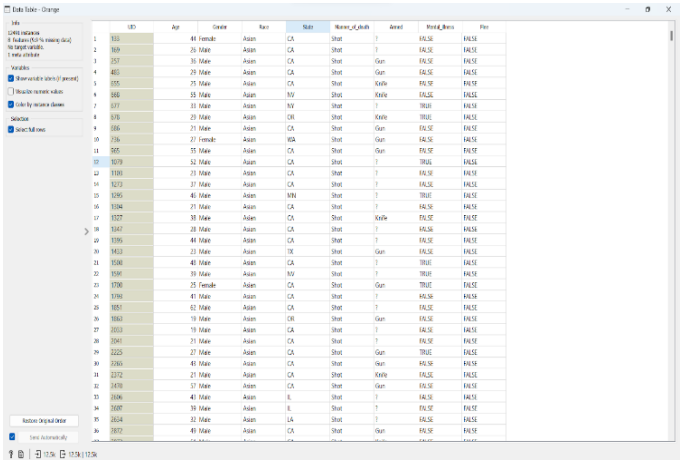
IV. DATA SET

Pada studi ini, digunakan data set dari Data World. Dataset ini terdiri dari data fatalities/korban jiwa yang disebabkan oleh polisi dengan total jumlah data 12.491, dan 12 kolom.

Dengan masing-masing feature:

- UID: data text
- Age: data numerik
- Gender: data kategorikal
- Race: data kategorikal
- State: data kategorikal
- Manner_of_Death: data kategorikal
- Armed: data kategorikal
- Mental_illness: kategorikal
- Flee: data kategorikal
- Name: text data
- City: text data
- Date: text data

Adapun raw data yang terdapat pada data set seperti Gambar 1 di bawah ini

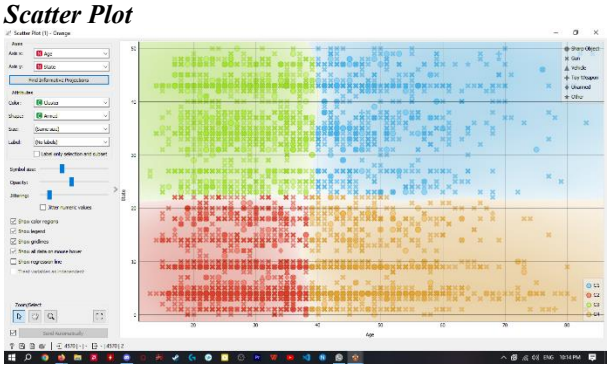


Gambar 1. Raw Data

V. PEMBAHASAN

Dipilih clustering 4 dikarenakan lebih optimal, selanjutnya hasil *clustering* dapat divisualisasikan menggunakan widget *scatter plot*.

A. *Scatter Plot*



Gambar 2. Scatter Plot

Scatter plot dapat terlihat dibagi menjadi 4 daerah, lalu data yang dibandingkan adalah pada sumbu X "Age" atau umur, lalu "State" menjadi sumbu Y. Maknanya disini adalah visualisasi korban jiwa yang terbunuh akibat tindakan kriminal dari umur dan daerah asal nya.

1. Daerah C1

Gambar 3. Scatter Plot Daerah C1

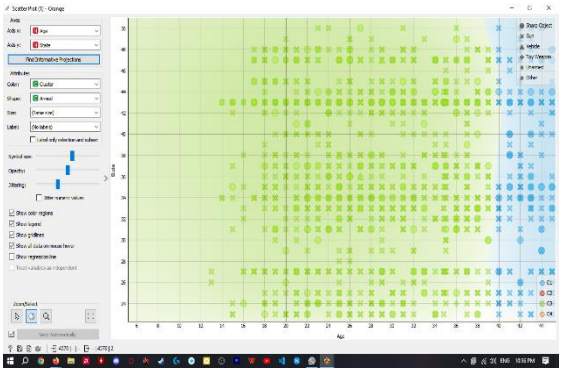
Pada pengelompokan C1, terlihat korban jiwa yang ada adalah dari rentang umur 40-84 atau umur dewasa hingga lanjut usia, serta dari daerah 23-50.

2. Daerah C2

Gambar 4. Scatter Plot Daerah C2

Pada pengelompokkan C2, ditemukan bahwa kelompok disini berisi korban jiwa yang rentangnya dari umur 13-38 atau dari usia remaja hingga dewasa dan lokasi yang terdampak berasal dari daerah atau "State" 0 sampai 22.

3. Daerah C3



Gambar 5. Scatter Plot Daerah C3

Pengelompokkan C3 adalah korban yang berumur 13 sampai 39 atau dari usia remaja hingga dewasa dan lokasi yang terdampak berasal dari daerah atau "State" 23 sampai 50.

4. Daerah C4

Gambar 6. Scatter Plot Daerah C4

Pengelompokkan C4 ini adalah daerah yang terdiri dari korban berumur 39 sampai 83 atau usia dewasa hingga lanjut usia dan daerah atau "state" 0 sampai 22.

B. *Widget Distributions*

Gambar 7. *Widget Distribution*

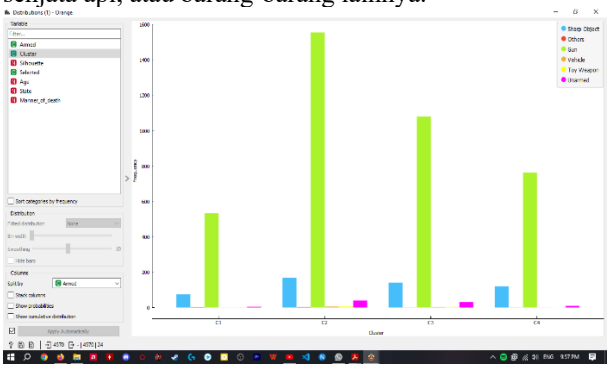
Dari gambar hasil capture Widget Distributions terlihat frekuensi atau korban jiwa terbanyak adalah dari kelompok C2, lalu diikuti nomor 2 terbanyak adalah kelompok C3, nomor 3 terbanyak yakni C4 dan yang terakhir adalah C1. Informasi dari *raw data* yang ada adalah ditemukan bahwa kelompok C2 yang berasal dari daerah-daerah kelompok tersebut menjadi tempat tinggal yang memiliki korban jiwa terbanyak akibat tindakan kriminal termasuk memberontak kepada polisi sehingga harus dilumpuhkan total, yang dimana daerah-daerah tersebut korban jiwa nya terdiri dari orang-orang berumur remaja hingga dewasa, lalu di urutan kedua adalah kelompok C3 yang berumur usia remaja hingga dewasa dan lokasi yang terdampak berasal dari daerah atau "State" 23 sampai 50, urutan ketiga adalah C4 yang terdiri dari korban usia dewasa hingga lanjut usia dan daerah atau "state" 0 sampai 22. Dan urutan terakhir adalah kelompok C1 korban jiwa yang ada adalah dari rentang umur dewasa hingga lanjut usia, serta dari daerah 23-50.

Pengetahuan atau *knowledge* yang dapat diambil adalah bahwa data korban jiwa terbanyak adalah C2, sehingga langkah yang dapat dilakukan oleh negara tersebut adalah memberi pengamanan lebih pada daerah-daerah yang telah disebutkan sebelumnya, atau memberi

langkah pencegahan lainnya seperti membangun sebuah *social service* yang baik sehingga terbentuk komunitas yang saling menghargai dan dapat mengurangi tingkat kriminalitas. memperketat dan mengevaluasi tentang peraturan jual beli dan peredaran senjata api di negara yang terlibat.

C. *Widget Distribution* dengan target feature "*Armed*"

Dengan menggunakan target feature "Armed" yang dijadikan target, informasi baru dapat ditemukan, yakni informasi mengenai pelaku kriminal yang sebelumnya mengancam polisi setempat dengan alat tajam maupun senjata api, atau barang-barang lainnya.



Gambar 8. Plot Cluster

Dari gambar diatas, dapat ditemukan fakta bahwa masing-masing cluster, pemegang senjata api lah yang masih menjadi faktor utama mengapa polisi melumpuhkan si pelaku. Jadi menurut hasil clustering tersebut, aparat keamanan dapat lebih berfokus pada daerah-daerah yang lebih beresiko, contohnya daerah C2 lalu diikuti oleh C3, C4 dan yang terakhir kelompok C1. Dari informasi tersebut, maka dapat diperkirakan bahwa banyak sekali penyebab pelaku berani berbuat kriminal karena memegang senjata api yang berbahaya, jadi dari hasil informasi itu jugalah pemerintah setempat dapat setidaknya berencana untuk lebih membatasi atau memperketat aturan peredaran senjata api dan memberi sanksi yang tegas bagi penjual illegal.

VI. KESIMPULAN

Telah dilakukan pemodelan dengan menggunakan metode K-Means clustering pada *Police Fatalities* data set. Data set ini diambil dari (Data World). Terdapat data dengan 9 atribut, namun dari 12.491 data tersebut masih terdapat missing data sehingga perlu dilakukan imputasi dengan metode *average/frequent*. Dari hasil pemodelan didapatkan bahwa dengan menggunakan metode tersebut, data lebih optimal jika dikelompokkan menjadi 4 cluster karena skor silhouette 4 cluster adalah yang tertinggi, serta hasil yang dihasilkan 1 dari 4 cluster yang ada menjadi urutan tertinggi daerah rawan kriminal.

DAFTAR PUSTAKA

[1] L. Suriani, "Pengelompokan Data Kriminal Pada Poldasu Menentukan Pola Daerah Rawan Tindak Kriminal Menggunakan Data Mining Algoritma K-Means Clustering", *Jurnal Sistem Komputer Dan Informatika (JSON)*, vol. 1, no. 2, pp. 151-157, 2020. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/JSON/article/view/1955

[2] A. T. R. Dani, S. Wahyuningsih, and N. A. Rizki, "Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu", vol. 1, no. 2, pp. 66-67, 2019. Available: https://ejurnal.ung.ac.id/index.php/jjom/article/view/2354/1498

[3] W. M. P. Dhuhita, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita", *Jurnal Informatika*, vol. 15, no. 2, pp. 160-174, 2015. Available: https://media.neliti.com/media/publications/103765-ID-clustering-menggunakan-metode-k-mean-unt.pdf

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations", In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, pp. 281-297, 1967. Available: https://books.google.co.id/books?hl=id&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&ots=nP_kJWKatQ&sig=FJ1wk_MZ5JK4bkEAYdWtcYR-KYc&redir_esc=y#v=onepage&q&f=false

[5] C. Purnamaningsih, R. Saptono, and A. Aziz, "Pemanfaatan Metode K-Means Clustering dalam Penentuan Penjurusan Siswa SMA", *Jurnal ITSMART*, vol. 3, no. 1, pp. 27-33, 2014. Available: https://jurnal.uns.ac.id/itsmart/article/view/644/1652

[6] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara", In *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, vol. 1, no. 1, pp. 762-767, 2019. Available: http://seminar-id.com/prosiding/index.php/sainteks/article/view/228

[7] P. M. S. Tarigan, J. T. Hardinata, H. Quniawan, M. Safii, and R. Winanjaya, "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus: Toko Sinar Harahap)", *Jurnal Janitra Informatika dan Sistem Informasi*, vol. 2, no. 1, pp. 9-19, 2022. Available: http://www.janitra.org/index.php/home/article/view/142

[8] Y. Mardi, "Data Mining: Klasifikasi Menggunakan Algoritma C4. 5", *Jurnal Edik Informatika Penelitian Bidang Komputer Sains dan Pendidikan Informatika*, vol. 2, no. 2, pp. 213-219, 2017. Available: https://core.ac.uk/download/pdf/229192374.pdf

[9] Karsito and W. M. Sari, "Prediksi Potensi Penjualan Produk Delifrance Dengan Metode Naive Bayes Di Pt. Pangan Lestari", *Jurnal SIGMA*, vol. 9, no. 1, pp. 67-

78, 2018. Available: https://jurnal.pelitabangsa.ac.id/index.php/sigma/article/view/465

[10] M. A. Muslim, B. Prasetiyo, E. L. H. Mawarni, A. J. Herowati, Mirqotussa'adah, S. H. Rukmana, and A. Nurzahputra, DATA MINING ALGORITMA C4.5, Universitas Negeri Semarang, 2019. [Online]. Available: http://lib.unnes.ac.id/33080/6/Buku_Data_Mining.PDF

[11] Indriyanti, N. Ichsan, H. Fatah, T. Wahyuni, and E. Ermawati, "IMPLEMENTASI ORANGE DATA MINING UNTUK PREDIKSI HARGA BITCOIN", *Jurnal Responsif: Riset Sains dan Informatika*, vol. 4, no. 2, pp. 118-125, 2022. Available: https://www.ejurnal.ars.ac.id/index.php/jti/article/view/762

[12] M. Faid, M. Jasri, and T. Rahmawati, "Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi", *Teknika*, vol. 8, no. 1, pp. 11-16, 2019. Available: http://repository.ikado.ac.id/72/