JSiI | Jurnal Sistem Informasi
Vol. 11 No. 2 September 2024, Hal. 56-63

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

KLUSTERISASI PENYEBAB KEMATIAN DI INDONESIA DENGAN PENERAPAN ALGORITMA K-MEANS

Agus Nursikuwagus[1], Suherman[2], Hendry Gunawan[3], Ilham Alamsyah[4]
[1,4]*Universitas Komputer Indonesia*
Jln. Dipatiukur No.112-114 Bandung – Indonesia
[2,3]*Jurusan Sistem Informasi Fakultas Teknologi Informasi Universitas Serang Raya*
Jln. Raya Cilegon Serang – Drangong Kota Serang
[1]agusnursikuwagus@email.unikom.ac.id
[2]suherman.unsera@gmail.com
[3]*hendrygunawan@unsera.ac.id*
[4]ilham.10519119@mahasiswa.unikom.ac.id

Abstrak

Kasus angka kematian yang terjadi di Indonesia dapat di kelompokan dalam beberapa kategori seperti natural disaster, nonnatural disaster, dan social disaster. Pemisahan suatu instans pada dataset sering menjadi hambatan Ketika melibatkan instans yang banyak. Penemuan karakteristik yang serupa akan menjadi tantangan untuk mendapatkan kluster terbaik. Penentuan jumlah kluster yang efektif terhadap dataset yang dimiliki menjadi permasalahan lain Ketika melakukan proses kluster. Berdasarkan permasalahan dan tantangan yang diperoleh, maka untuk menjawab hal ini dilakukanlah pemodelan clustering dengan bantuan algoritma clustering. Metode yang digunakan pada pengklusteran ini adalah K-Means. Metode ini telah menjadi usulan dari berbagai penelitian yang menyatakan sukses dalam melakukan clustring. Penentuan K terbaik yaitu dengan bantuan elbow curve, dengan melihat titik elbow pada hasil generasi kurva dari dataset. Rangkaian penyelesaian penelitian ini adalah dengan mengikuti flow of process datamining yang dimulai dengan Data Preprocessing, Data modeliing, dan visualization hasil. bertujuan untuk mengetahui klusterisasi penyebab kematian di Indonesia berdasarkan kategori yang di sebutkan di atas. Dataset yang digunakan adalah sebanyak 648 instans yang diambil dari rentang 2000 – 2020 mengenai kasus kematian pada 34 provinsi di Indonesia. Data preprocessing adalah melakukan *cleansing data*, pembersihan *outlier*, *missing value*, data *transformation*. Pembersihan outlier yaitu menggunakan bantuan Box Plot, sedangkan transformation menggunakan fungsi transormasi data diskrit menjadi data numerik. Pada data modelling, algoritma K-means dengan K atau banyaknya diperoleh dari hasil Elbow Curve. Selain proses clustering, penggalian pola juga dilakukan dengan metode *classification* yang hasilnya ditunjukan dengan akurasi sebesar 63%. Meninjau dari hasil *classification*, bahwa klasifikasi kematian yang berasal dari sumber sosial, tidak dapat diprediksi dengan akurat. Klasifikasi sumber kematian dari Sosial tidak berhasil dipolakan oleh mesin learning. Matrik konfusi menunjukkan hanya 55 instans yang benar untuk bencana alam, bencana non alam dan penyakit sebesar 353 yang benar, dan untuk bencana sosial tidak berhasil diprediksi. Dari hasil ini, maka dapat diperoleh tantangan baru yaitu memperbaiki akurasi dengan mempertimbangan *Imbalance Class*, dan *Resampling* yang belum digunakan pada penelitian ini.

Kata kunci*: Unsupervise*, *clustering, K-Means*, *euclidean distance, elbow curve*,.

I. PENDAHULUAN

Indonesia tergolong negara dengan angka kematian terbanyak berdasarkan survei statista.com.. Angka kematian yang meningkat ini disebabkan oleh berbagai faktor yang terjadi di negara Indonesia, mulai dari kematian akibat kecelakaan, bencana alam, hingga akibat penyakit. Setiap kematian biasanya disebabkan oleh beberapa faktor yang mempengaruhi, kasus kematian tersebut dikelompokkan menjadi beberapa kategori berdasarkan faktor yang mempengaruhi kematian diantaranya seperti bencana yang berasala dari alam maupun non alam, serta sosial.

Informasi yang disampaikan pada paragraf pertama dikumpulkan dalam bentuk dataset penyebab kematian di Indonesia. Berdasarkan hasil koleksi diperoleh dataset dengan jumlah 649 instans, dengan informasi mengenai latar belakang kematian di Indonesia. Berdasarkan data yang dikumpulkan dan melihat potensi penggalian pola data yang ada pada dataset, maka penelitian ini diarahkan kepada penelitian pengelompokan instans yaitu dengan metode *Clustering*.

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Clustering merupakan salah satu metode dalam mesin *learning* yang dapat membangun informasi berdasarkan kedekatan jarak antara instans [1], [2], [3]. Hal ini ditujukan untuk menyasar instans – instans pada dataset yang memiliki karakteristik fitur yang sama. Hal lainnya adalah juga untuk mengetahui faktor-faktor fitur dataset yang berhubungan dengan angka kematian di Indonesia [4], [5].

Clustering merupakan salah satu metode data mining yang menyelesaikan penggalian data dengan mengandalkan kedekatan atau kemiripan antar instans pada dataset. Ekstraksi informasi dilakukan dengan menyasar setiap instans dan mengumpulkan setiap instans dengan karakteristik yang sama [6], [7].

Data mining adalah serangkaian tindakan yang digunakan untuk mendapatkan ekstraksi informasi dan mengidentifikasi informasi terkait dan berguna dari berbagai database yang sangat besar dengan menggunakan teknik matematika, statistik, kecerdaan buatan, dan pembelajaran mesin [6]. Data mining mencakup mengekstraksi nilai – nilai yang penting dari kumpulan data, yang terdiri dari pengetahuan *posterior. Classification, clustering, reinforcement*, dan *association* adalah beberapa komponen data mining. Dua metode pemodelan *clustering* dan klasifikasi dapat digunakan untuk menyelidiki dataset [7].

Pendekatan metode clustering pada data mining dijadikan suatu kendali proses untuk menemukan kelompok instans tersebut. Berdasarkan permasalahan yang ditemukan pada kajian dataset yang dimiliki, maka kontribusi yang diberikan pada penelitian adalah :

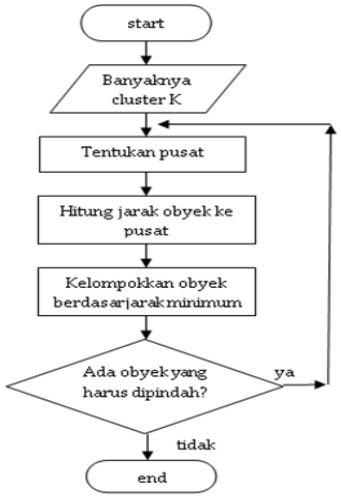
- Penggunaan *feature selection* pada dataset dapat memperkuat proses *clustering*, sehingga fitur-fitur yang memiliki pengaruh kuat saja yang akan digunakan [8], [9], [10].
- Penentuan jumlah K pada algoritma K-Means memberikan dampak pada proses pemisahan karakteristik data. Perolehan klustering yang didapat memberikan hasil kluster yang *well-seperated* [7], [11], [12], [13], [14] .
- Penentuan jumlah K dengan bantuan metode *elbow* dan *silhoutte* memberikan kepastian K yang efektif yang digunakan pada K-Means[12], [13] .

Penyusunan dan penulisan artikel ini disusun berdasarkan urutan yang disarankan pada penyajian subbab. Hal ini dilakukan agar pembaca dapat mudah mengikuti dan memahami isi dari artikel. Adapun urutannya disusun sebagai berikut bagian awal berisikan pendahuluan, bagian kedua adalah metodologi penelitian yang merupakan alur jalannya penelitian. Bagian ketiga berisi temuan dan diskusi tentang aspek luar dari penelitian. Bagian keempat adalah kesimpulan, yang menguraikan tujuan penelitian.

II. METODOLOGI PENELITIAN

Setiap penelitian yang dilakukan diperlukan suatu metodologi penelitian untuk membantu mengarahkan dalam penyelesaian penelitian. Pada penelitian ini digunakan metode kuantitatif dengan mengumpulkan jumlah kasus tentang penyebab kematian di Indonesia dari berbagai sumber [15], [16], [17], [18], [19]. Metode penelitian ini memiliki tahapan-tahapan untuk menyelesaikan proses, seperti pengumpulan data penyebab kematian, mengolah kasus data penyebab kematian, dan memproses klasifikasi penyebab kematian.

Alur dari penelitian yang dilakukan dapat ditunjukkan dengan model *pipeline*. Pipeline penelitian ini mengikuti algoritma K-Means yang terdapat pada Gambar 1 [14], [20].



Gambar 1. Flowchart K-Means Algorithm

A. *Clustering*

Widodo menyampaikan mengenai *clustering* yang merupakan metode yang dapat mengelompokkan data berdasarkan karakteristik yang berdekatan dengan pusat kluster atau *center point*. Cluster sendiri merupakan grup atau sekumpulan objek atau segmentasi data yang serupa [7], [11], [13], [14].

Pengelompokan data digunakan untuk membentuk kelas baru dengan mengidentifikasi kelompok-kelompok yang memiliki fitur khusus. Objektif yang terletak dalam kelompok yang terbentuk tidak hanya memiliki karakteristik yang berbeda dari yang terletak dalam kelompok lainnya, tetapi juga memiliki karakteristik yang mirip satu sama lain. K-Medoids, Fuzzy C-Means, DBSCAN, dan K-Means adalah beberapa algoritma *clustering* [20], [21]. Setiap algoritma memiliki kemampuan unik, bersama dengan keunggulan dan kelemahan.

B. *K-Means*

K-Means adalah teknik yang menggunakan dasar pengelompokkannya menggunakan jarak (distance). Penggunaan algoritma K-Means harus dimulai dengan menentukan jumlah K (kluster). Jumlah K dapat diketahui dengan Teknik Elbow Curve dan akan memberikan nilai K optimal pada posisi elbow-nya. Penggunaan jarak pada K-Means dapat bervariasi seperti Eucledean Distance, Jaccard Distance, Hamming Distance, dan Cosine Distance [22].

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

K-means clustering adalah salah satu algoritma yang paling sederhana dalam hal pengelompokkan. Kerja yang dilakukan pada K-Means adalah proses pemodelan menggunakan pembelajaran tidak diawasi dan mengelompokkan data menggunakan sistem partisi. Secara umum, dua jenis pengelompokan yang digunakan baik secara *hierarchical* maupun *non-hierarchical* [22], [23]. Metode K-Means tentunya memiliki keuntungan dan kekurangan dalam melakukan proses pengelompokkannya [6], [13], [23].

Kelebihan k-means [6], [12], [13]: Sangat mudah untuk dimulai dan digunakan. Belajar lebih cepat dan sangat fleksibel. Sangat umum digunakan. Prinsipnya sederhana dan dapat dijelaskan dalam non-statistik.

Kekurangan K-Means [6], [7], [24] meliputi Titik K diinisialisasikan secara tidak sengaja sebelum algoritma dijalankan, sehingga pengelompokan data yang dihasilkan dapat berbeda-beda. Namun, pengelompokan yang dihasilkan menjadi tidak ideal jika nilai inisialisasi yang diperoleh secara acak kurang baik. Hal yang dihadapi ketika melakukan clustering dengan K-Means adalah terperangkap dalam kasus yang dikenal sebagai kutukan dimensi. Hal ini juga akan terjadi dalam kasus di mana data pelatihan memiliki ukuran dimensi yang besar. Sebagai contoh, jika data pelatihan memiliki dimensi dua maka hanya terdiri dari dua atribut sehingga mudah untuk memprosesnya. Namun, jika terdiri dari dua puluh atribut, maka dimensinya akan mencapai dua puluh. Kendala ini akan menjadi sumber kesulitan ketika setiap instan harus ditentukan jarak dengan center point-nya. Perhitungan jarak akan menjadi lama karena harus mencari akar dari seluruh atribut yang banyak. Penentukan K menjadi salah satu penentu agar proses lebih efektif dan mendapatkan hasil yang diharapkan. Pada penentuan K tidak ada jaminan bahwa akan ada kumpulan cluster terbaik.

C. *Data Mining*

Gartner group menyampaikan definisi mengenai Data mining yaitu suatu Teknik yang diajukan bereksperimen dapat menemukan pola, hubungan, pola, dan kebiasaan baru. Teknik ini didukung dengan disiplin ilmu matematika, statistika, database, dan visualisasi data. Pada definisi yang lain dari David et .al, menyampaikan bahwa data mining merupakan suatu Teknik analisis terhadap data yang sifatnya bulk/besar guna menggali pola, dan hubungan antara data.

Data mining merupakan multidisiplin ilmu yang dapat terdiri dari statistika, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengklasifikasikan, mengklusterisasi, asosiasi, dan *reinforcement learning*. Meninjau definisi yang telah diberikan dan elemen penting terkait data mining, maka data mining dapat dikenali yaitu [6]:

- Data mining merupakan proses mandiri yang diupayakan dapat melakukan proses terhadap data yang sudah ada.
- Dataset yang digunakan selalu memiliki volume yang besar.
- Data Mining dapat menggali pola data sehingga dapat menjadi suatu kluster, klasifikasi, asosiasi, dan semi-klasifikasi.

D. *Angka Kematian*

Secara definisi angka kematian merupakan rasio antara jumlah orang meningal per tahun dibagi dengan seribu orang. Bila nilainya di atas Sembilan belas berarti dianggap tinggi. Bila nilainya diantara 14 – 18, maka dianggap sedang. Bila nilainya di bawah 13 berarti dianggap rendah.

E. *Bencana Alam*

Bencana alam merupakan aktivitas yang terjadi karena peristiwa alam, contohnya kekeringan, gempa, banjir, tsunami, tanah longsor, gunung meletus, dan angin topan.

F. *Bencana Non Alam*

Bencana alam merupakan bencana yang terjadi diluar peristiwa alam, seperti peperangan, pengeboman suatu wilayah, penebangan hutan sembarangan, dan lainnya.

G. *Bencana Sosial*

Bencana yang disebabkan oleh serangkaian peristiwa yang dilakukan oleh manusia, seperti teror, konflik antar kelompok atau antarkomunitas, disebut bencana sosial.

H. *Scatter Plot*

Scatterplot merupakan fungsi untuk menggambarkan sebaran data dalam bentuk koordinat kartesian. Scatterplot dapat menunjukan kedekatan antara data yang merupakan hubungan karakteristik yang serupa. Grafik ini juga dapat memberikan pengetahuan mengenai arah pergerakan data menurut kluster yang ditetapkan. Titik yang dipetakan merupakan titik X dan Y yang dikomposisikan dari atribut yang didefinisikan pada datasert [25]. Penggunaan dataset tidak terlepas dari kekurangan dan kelebihan. Kelebihan scatterplot dapat menunjukan jangkauan data yang jelas. Penggambarann lainnya lagi yaitu titik maksimum dan titik minimum dapat diketahui dengan jelas. Selain itu dapat menunjukkan hubungan positif maupun negative antara data. Scatterplot juga selain memiliki kelebihan, ada juga kekurangan dari grafik. Adapun kekurangannya adalah terbatas dalam penampilan yang hanya didasari oleh dua atribut saja. Data yang diplot hanya berkisar data kuantitaf saja, dan tidak bisa menggunakan data diluar yang sudah digunakan.

III. HASIL DAN PEMBAHASAN

Pada bagian ini memaparkan mengenai hasil yang diperoleh. Hasil dan pembahasan merupakan satu kesatuan yang memaparkan output dari penelitian serta pembahasan dari output tersebut.

A. Analisis Dataset

Pada bagian ini menjelaskan mengenai análisis yang dilakukan dengan teknik klustering. Terkait análisis dataset, maka diperlukan sajian dataset mengenai penggalian dataset lebih dalam. Rangkaian yang dilakukan pada proses ini adalah meliputi data *cleansing*, data *transformation*, *missing value*, *imbalance class* jika sudah ada *label class*. Sekarang, hampir segala aktivitas melibatkan teknologi, dan teknologi ini pasti terkait dengan perbaikan data yang terus bertambah setiap

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

waktu. Data hanya akan menjadi sia-sia jika mereka dibiarkan menumpuk. Namun, data dapat diolah dan digunakan untuk mendapatkan informasi bermanfaat. Oleh karena itu, satu tahap pengolahan data yang sangat penting adalah analisis data. Perihal análisis dapat dipahami bahwa teknik análisis dapat dilakukan secara *qualitative análysis* dan *quantitative análysis*.

Analisis kualitatif merupakan análisis yang melakukan pendekatan penelitian dengan menggunakan teknik selain matematika dan statistika. Tujuan analisis kualitatif adalah untuk menemukan makna dari data. Peneliti biasanya menyajikan hasil analisis dalam bentuk angka yang akan diuraikan dan disajikan. Análisis dengan teknik kualitatif yaitu melakukan pembacaan data dari sumber yang telah dikumpulkan baik berupa gambar, tulisan, pesan, dan lainnya.

Análisis kuantitatif merupakan proses yang dikerjakan dengan berbantuan penggunaan ilmu matematika dan statistika. Analisis kuantitatif memiliki dua pendekatan yaitu *descriptive análysis* dan *inferential análisis*.

Dataset yang di peroleh untuk penelitian ini adalah informasi penyebab angka kematian Indonesia. Proses yang dilakukan pada dataset ini meliputi *cleansing data*, *transformation* data, dan *missing value.* Proses ini akan dilakukan dengan bantuan software ORANGE [26], [27].

B. Attribute Dataset

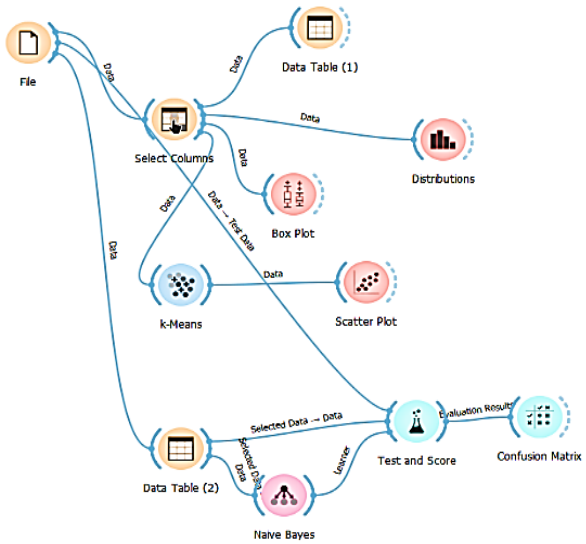
Kumpulan data yang telah diperoleh melalui studi literatur berisikan tujuh atribut dan terdapat 649 instans. Pada Table I ditunjukan atribut dan tipenya untuk pengklusteran.

TABLE I. ATRIBUTE DATASET

No	Atribut		
	Nama	**Jenis**	**Deskripsi**
1	Type	Categorical	Tipe kematian berdasarkan kategori
2	Year	Numeric	Tahun Periode Kematian
3	Data Redudancy	Categorical	penyimpanan data yang sama secara berulang dalam beberapa file
4	Total Death	Numeric	Total kematian selama periode
5	Source	Categorical	Sumber didapatkannya angka kematian
6	Source URL	Categorical	Alamat url dari source
7	Cause	Meta	Penyebab kematian
8	Page at Source	Meta	Jumlah halaman di

C. Data Mining Process

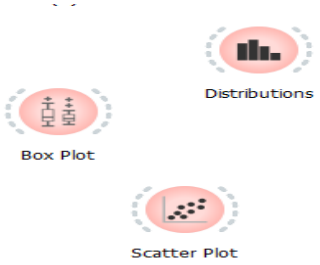
Proses pada data mining selalu mengedepankan penggalian pola informasi pada dataset. Dukungan ilmu matematika, statistika, dan teknik kecerdasan buatan diupayakan untuk menemukan pola hubungan yang ada pada dataset [28]. Pipeline penelitian yang dilakukan bisa dilihat pada Gambar 2. Rangkaian proses yang disusun mengikuti kaidah proses siklus data mining. Siklus tersebut meliputi *data preprocessing*, data *modelling*, dan visualisasi. Pada Gambar 2 terdapat bagian data *preprocessing* seperti fungsi *select column*, *Box-plot*, dan *distribution*. Pada bagian data *modelling* dilakukan tugas seperti penggunaan métode K-means, dan Naive Bayes. Pada bagian visualisasi digunakan grafik *scatterplot*, penampilan evaluasi dengan *confusión matrix*.



Gambar 2. Desain Model Clustering Penyebab Kasus Kematian di Indonesia [27], [29]

D. Proses Pengujian Clustering

Untuk menguji model klasifikasi yang telah dibuat sebelumnya, kumpulan data yang diuji diperlukan untuk mengetahui hasil klasifikasi. Untuk mengetahui hasil pengujian clustering, widget visualisasi aplikasi Orange diperlukan, seperti pada Gambar 3.



Gambar 3. Widget yang digunakan clustering dataset [26], [27], [29]

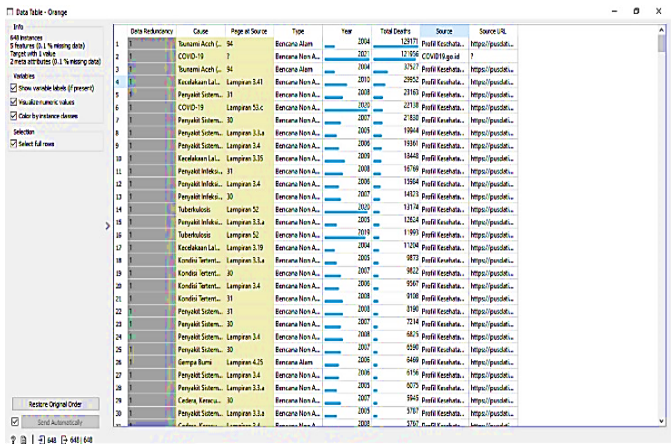
Berdasarkan Gambar 3 Sekumpulan data yang didapat sebelumnya diuji menggunakan tiga model *tool visualize*,yaitu *Distributions*, *Box plot*, dan *Scatter plot*.

Berikut adalah hasil data set apa bila di visualkan menggunakan table, seperti pada Gambar 4.

Received: July 15th, 2024| Published: September 15th, 2024

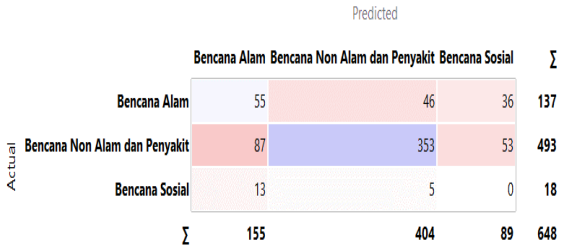
JSiI | Jurnal Sistem Informasi
Vol. 11 No. 2 September 2024, Hal. 56-63

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181



Gambar 4. Dataset

E. *Confusion Matrix*



Gambar 5. Jumlah angka kematian di Indonesia dari 2000-2020 [30], [31], [32]

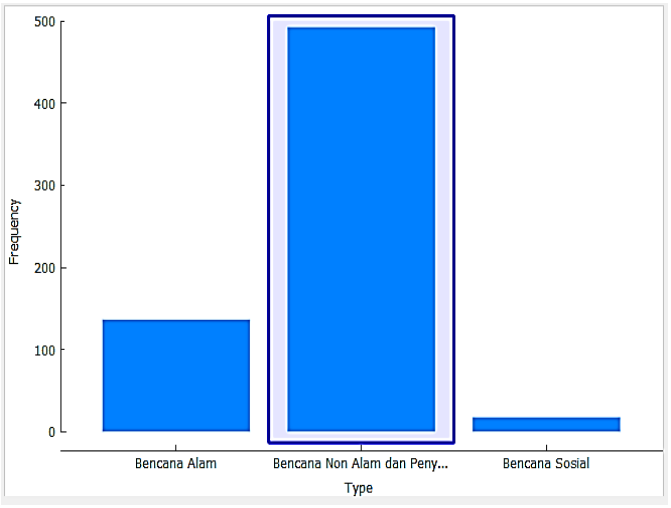
Berdasarkan Gambar 5, *confusion matrix* menunjukan bahwa becana non alam memiliki angka kematian tertinggi dibanding dua kategori lainnya. Penyebab kematian akibat bencana non alam berjumlah 497 kasus, penyebab kematian bencana alam memiliki 137 kasus kematian, dan bencana sosial memiliki jumlah penyebab kematian paling sedikit yaitu berjumlah 18 kasus kematian.

F. *Distribution*

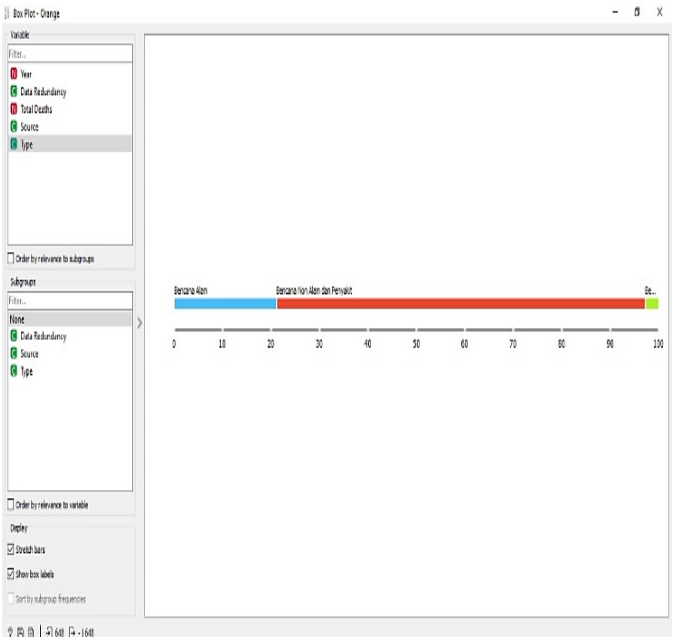
Berdasarkan gambar 6, menunjukan bahwa penyebab kematian akibat becnana non alam memiliki angka kematian tertinggi dibandingkan dengan kematian penyebab dari bencana alam dan bencana sosial.

G. *Box Plot*

Pada gambar di atas menunjukan ada 3 cluster yang memiliki 3 warna yaitu C1 warna biru menujukan kematian disebabkan bencana alam, C2 warna merah yang menunjukan kematian disebabkan oleh bencana alam dan penyakit, dan C3 kematian yang di sebabkan oleh bencana social di tunjukan oleh warna kuning. Dari ke 3 cluster tersebut warna C2 memiliki garis merah yang paling Panjang menandakan bahwa C2 memiliki tingkat angka kematian yang lebih tinggi dibandingkan dua lainnya.



Gambar 6. Hasil Visualisasi dari Distribution [33]



Gambar 7. Box Plot [30], [31], [32]

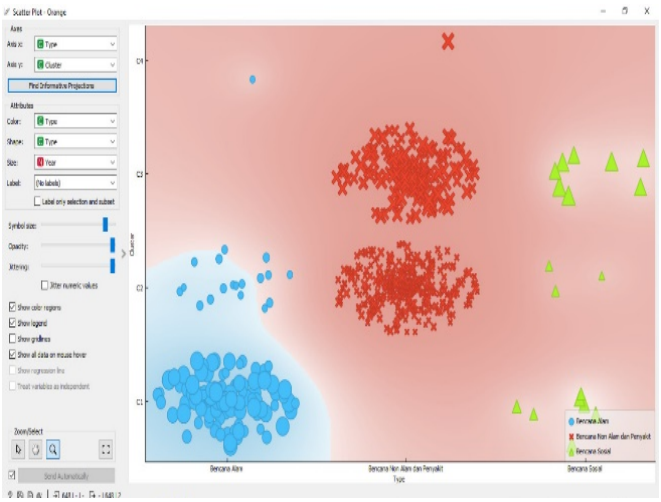
H. *Scatter Plot*

Pada Gambar 8 adalah hasil scatter plot dari kasus data penyebab kematian di Indonesia pada icon x merah di gambar menujukan banyaknya kematian disebabkan bencana non alam dan penyakit dan menjadikan nomer 1 penyebab kematian tertinggi di indonesia di susul oleh bencana alam sebagai penyebab kematian di Indonesia dan nomer 3 urutan terakhir disebabkan bencana social sebagai penyebab kematian di Indonesia.

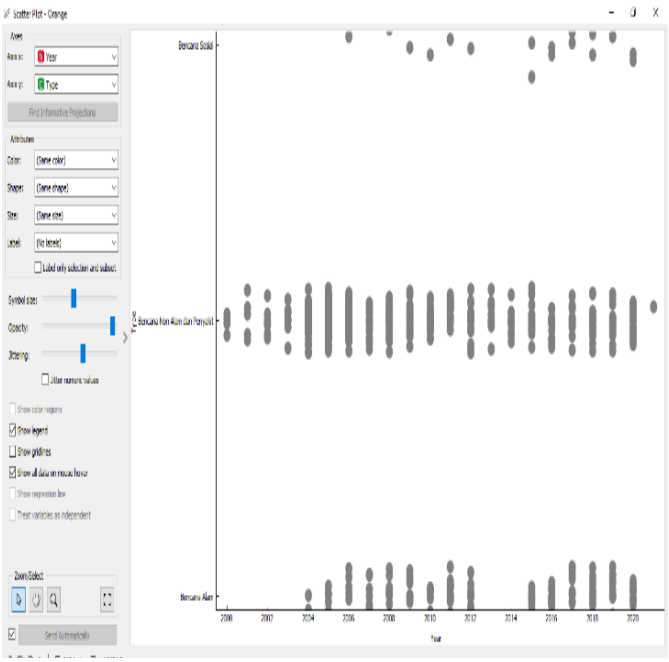
Pada Gambar 9 adalah hasil dari scatter plot dari cluster 1 (tinggi) bencana non alam dan penyakit yang memiliki angka kematian tertinggi sejak dari tahun 2000-2020, sedangkan pada cluster bencana alam menjadi penyebab kematian pada tingkat cluster 2 (sedang), dan terakhir dengan tingkat cluster 3 (rendah) yaitu bencana sosial berada di paling rendah dalam kasus kematian di Indonesia.

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

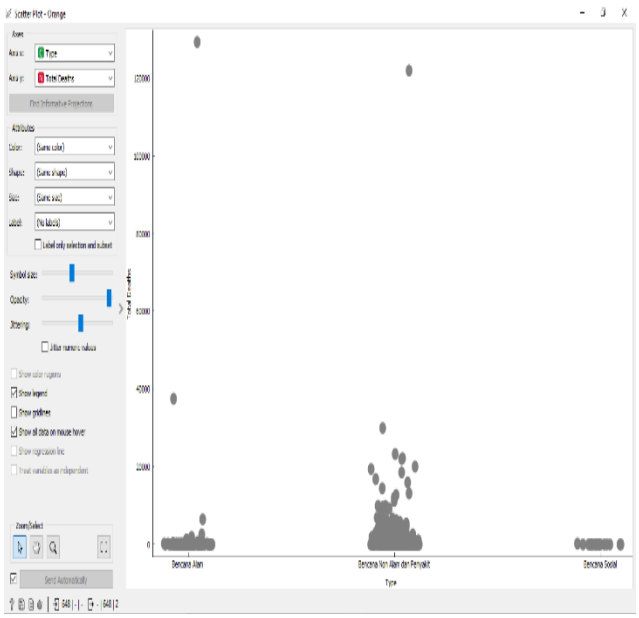


Gambar 8. Scatter Plot [25]



Gambar 9. Scatter Plot dengan cluster Non Alam dan Penyakit [25]

Gambar 10 merupakan hasil dari scatter plot dari jumlah kematian terbanyak yang digrupkan menjadi tiga grup yaitu bencana alam, penyakit, bencana non alam, dan bencana sosial. Dilihat dari data yang ditampilkan bencana alam menjadi penyebab kematian tertinggi sebesar 129171 angka kasus kematian pada bencana tsunami di aceh yang terjadi pada tahun 2004. Sedangankan tingkat rendah berada pada bencana non alam dan penyakit sebanyak 121956 angka kasus kematian kasus Covid-19 pada tahun 2021.



Gambar 10. Scatter Plot dengan tiga cluster [25]

IV. KESIMPULAN

Kesimpulan dari hasil penelitian Data Mining dengan Teknik Clustering Menggunakan Algoritma K-Means Pada Kasus Data Penyebab Kematian di Indonesia adalah:

- Penggunaan *feature selection* seperti Year, Data Redudancy, Total Death, Source, Source URL, Cause, Page at Source menghasilkan kluster yang *well-separated*.
- Penentuan jumlah K dengan K=3 pada algoritma K-Means memberikan dampak pada proses pemisahan karakteristik instans.
- Penentuan jumlah K dengan bantuan metode *elbow* mendapatkan K=3 yang merupakan jumlah kluster efektif untuk memperoleh kluster [12], [13] .

V. ACKNOWLEDMENT

Terimakasih kami sampaikan kepada seluruh mahasiswa data mining Universitas Komputer Indonesia Angkatan 2020 yang membantu dalam melakukan pengumpulan data, cleaning data, dan memperbaiki dataset agar data diproses oleh mesin pembelajaran.

VI. REFERENCES

[1] A. M. Siregar, "Pengelompokan Bidang Laju Pertumbuhan Ekonomi Indonesia Menggunakan Algoritma K-Means," Jurnal Accounting Information System (Aims), Vol. 2, No. 2, Pp. 140–151, 2019.

[2] F. A. I. S. Aji, S. Achmadi, And F. X. Ariwibisono, "Penerapan Metode Clustering Pada Analisis Realisasi Pendapatan Asli Daerah Dengan Algoritma K-Means,"

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Jati (Jurnal Mahasiswa Teknik Informatika), Vol. 5, No. 2, Pp. 443–451, 2021.

[3] R. D. Bekti, R. N. Zulfahmi, M. K. Daul, W. J. Pradnyaana, And E. Sutanta, "Sistem Informasi Berbasis Website Untuk Pemetaan Wilayah Berdasarkan Clustering Kerentanan Kriminalitas," Jurnal Informatika Teknologi Dan Sains (Jinteks), Vol. 6, No. 3, Pp. 620–626, 2024.

[4] J. Li Et Al., "Feature Selection: A Data Perspective," Acm Computing Surveys (Csur), Vol. 50, No. 6, Pp. 1–45, 2017.

[5] B. Venkatesh And J. Anuradha, "A Review Of Feature Selection And Its Methods," Cybernetics And Information Technologies, Vol. 19, No. 1, Pp. 3–26, 2019.

[6] F. Juliawati, R. Buaton, And R. Saragih, "Pengelompokan Data Mining Penerimaan Bantuan Pangan Non Tunai (Bpnt) Menggunakan Metode Clustering (Studi Kasus: Kantor Desa Payabakung Hamparan Perak)," Explorer (Hayward), Vol. 3, No. 2, Pp. 69–76, 2023.

[7] S. S. Helma, M. Mustakim, E. Normala, And Others, "Analisis Cluster Menggunakan Algoritma K-Means Pada Data Fasilitas Pelayanan Kesehatan Kota Pekanbaru," In Seminar Nasional Teknologi Informasi Komunikasi Dan Industri, Pp. 131–137.

[8] J. Li Et Al., "Feature Selection: A Data Perspective," Acm Computing Surveys (Csur), Vol. 50, No. 6, Pp. 1–45, 2017.

[9] A. Ahmad And L. Dey, "A Feature Selection Technique For Classificatory Analysis," Pattern Recognit Lett, Vol. 26, No. 1, Pp. 43–56, 2005.

[10] B. Venkatesh And J. Anuradha, "A Review Of Feature Selection And Its Methods," Cybernetics And Information Technologies, Vol. 19, No. 1, Pp. 3–26, 2019.

[11] A. M. Siregar, "Pengelompokan Bidang Laju Pertumbuhan Ekonomi Indonesia Menggunakan Algoritma K-Means," Jurnal Accounting Information System (Aims), Vol. 2, No. 2, Pp. 140–151, 2019.

[12] F. A. I. S. Aji, S. Achmadi, And F. X. Ariwibisono, "Penerapan Metode Clustering Pada Analisis Realisasi Pendapatan Asli Daerah Dengan Algoritma K-Means," Jati (Jurnal Mahasiswa Teknik Informatika), Vol. 5, No. 2, Pp. 443–451, 2021.

[13] B. Ruhiman, A. Ramdan, And C. Juliane, "Algorithm K-Means Clustering Algorithm To Classify The Level Of Legal Information Service Objectives In West Java Province: K-Means Clustering Algorithm To Classify The Level Of Legal Information Service Objectives In West Java Province," Jurnal Komputer Terapan, Vol. 8, No. 1, Pp. 178–185, 2022.

[14] T. Jelita, R. Buaton, And M. Simanjuntak, "Pengelompokan Bidang Usaha Terhadap Bantuan Produktif Usaha Mikro (Bpum) Berdasarkan Wilayah Deli Serdang Menggunakan Metode Clustering K-Means (Studi Kasus: Dinas Koperasi Dan Umkm Kabupaten Deli Serdang)," Explorer (Hayward), Vol. 3, No. 2, Pp. 50–57, 2023.

[15] H. E. Fischer, W. J. Boone, And K. Neumann, "Quantitative Research Designs And Approaches," In Handbook Of Research On Science Education, Routledge, 2023, Pp. 28–59.

[16] L. Bode Et Al., "Study Designs For Quantitative Social Science Research Using Social Media," 2020.

[17] L. J. Duckett, "Quantitative Research Excellence: Study Design And Reliable And Valid Measurement Of Variables," Journal Of Human Lactation, Vol. 37, No. 3, Pp. 456–463, 2021.

[18] J. Bloomfield And M. J. Fisher, "Quantitative Research Design," Journal Of The Australasian Rehabilitation Nurses Association, Vol. 22, No. 2, Pp. 27–30, 2019.

[19] P. D. Morrell And J. B. Carroll, "Quantitative Study Designs," In Conducting Educational Research, Brill, 2010, Pp. 175–186.

[20] N. L. Anggreini And Others, "Teknik Clustering Dengan Algoritma K-Medoids Untuk Menangani Strategi Promosi Di Politeknik Tedc Bandung," Jurnal Teknologi Informasi Dan Pendidikan, Vol. 12, No. 2, Pp. 1–7, 2019.

[21] B. S. Shedthi, S. Shetty, And M. Siddappa, "Implementation And Comparison Of K-Means And Fuzzy C-Means Algorithms For Agricultural Data," In 2017 International Conference On Inventive Communication And Computational Technologies (Icicct), 2017, Pp. 105–108.

[22] D. Deng, "Dbscan Clustering Algorithm Based On Density," In 2020 7th International Forum On Electrical Engineering And Automation (Ifeea), 2020, Pp. 949–953.

[23] A. Latifi-Pakdehi And N. Daneshpour, "Dbhc: A Dbscan-Based Hierarchical Clustering Algorithm," Data Knowl Eng, Vol. 135, P. 101922, 2021.

[24] R. D. Bekti, R. N. Zulfahmi, M. K. Daul, W. J. Pradnyaana, And E. Sutanta, "Sistem Informasi Berbasis Website Untuk Pemetaan Wilayah Berdasarkan Clustering Kerentanan Kriminalitas," Jurnal Informatika Teknologi Dan Sains (Jinteks), Vol. 6, No. 3, Pp. 620–626, 2024.

[25] T.-H. Huang, M. L. Huang, And K. Zhang, "An Interactive Scatter Plot Metrics Visualization For Decision Trend Analysis," In 2012 11th International Conference On Machine Learning And Applications, 2012, Pp. 258–264.

[26] R. Ratra And P. Gulia, "Experimental Evaluation Of Open Source Data Mining Tools (Weka And Orange)," International Journal Of Engineering Trends And Technology, Vol. 68, No. 8, Pp. 30–35, 2020.

[27] Z. R. Mohi, "Orange Data Mining As A Tool To Compare Classification Algorithms," Dijlah Journal Of Sciences And Engineering, Vol. 3, No. 3, Pp. 13–23, 2020.

Received: July 15th, 2024| Published: September 15th, 2024

DOI 10.30656/jsii.v11i2.9276
p-ISSN: 2406-7768
e-ISSN: 2581-2181

[28] F. Juliawati, R. Buaton, And R. Saragih, "Pengelompokan Data Mining Penerimaan Bantuan Pangan Non Tunai (Bpnt) Menggunakan Metode Clustering (Studi Kasus: Kantor Desa Payabakung Hamparan Perak)," Explorer (Hayward), Vol. 3, No. 2, Pp. 69–76, 2023.

[29] E. Mardiani Et Al., "Membandingkan Algoritma Data Mining Dengan Tools Orange Untuk Social Economy," Digital Transformation Technology, Vol. 3, No. 2, Pp. 686–693, 2023.

[30] M. Heydarian, T. E. Doyle, And R. Samavi, "Mlcm: Multi-Label Confusion Matrix," Ieee Access, Vol. 10, Pp. 19083–19095, 2022.

[31] D. Krstinić, M. Braović, L. Šerić, And D. Božić-Štulić, "Multi-Label Classifier Performance Evaluation with Confusion Matrix," Computer Science & Information Technology, Vol. 1, Pp. 1–14, 2020.

[32] J. Liang, "Confusion Matrix: Machine Learning," Pogil Activity Clearinghouse, Vol. 3, No. 4, 2022.

[33] K. Zhou and S. Yang, "Effect Of Cluster Size Distribution On Clustering: A Comparative Study Of K-Means And Fuzzy C-Means Clustering," Pattern Analysis And Applications, Vol. 23, No. 1, Pp. 455–466, 2020.

 Received: July 15th, 2024| Published: September 15th, 2024