Juminten
Jurnal Manajemen Industri dan Teknologi

e-ISSN: 2721-4079
Vol. 04 No. 01 Tahun 2023
Hal. 1-12

Analisis Persebaran Penyakit Diare di Jawa Barat Menggunakan *Data Mining* dengan Algoritma*K-Means Clustering*

Ivan Alvino Ryansyah Putra Pratama✉, Dira Ernawati
Program Studi Teknik Industri, Fakultas Teknik
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya Surabaya 60294
e-mail: vinopratama8715@gmail.com✉, dira.ti@upnjatim.ac.id

ABSTRAK

Penyakit diare banyak ditemui di negara berkembang. Penyakit ini sering menyebabkan kasus kematian dan permasalahan kesehatan di dunia. Provinsi di Indonesia yang memiliki tingkat kasus Diare terbanyak adalah Jawa Barat dimana jumlah kasus yang terjadi mengalami peningkatan dan penurunan dalam kurun waktu enam tahun terakhir. Maka dari itu, Pemerintah Provinsi Jawa Barat harus memberikan perhatian lebih dalam hal penanganan penyakit Diare ini. Penelitian ini dilakukan dengan melakukan klasterisasi kota terhadap tingkat penyakit Diare di Jawa Barat pada tahun 2016-2021. Tujuan dari penelitian ini yaitu untuk mengetahui besar tingkat risiko penyakit diare setiap kota di Jawa Barat guna membantu Pemerintah dalam pemilihan strategi untuk penanganan kasus ini. Metode Knowledge Discovery in Database merupakan metode yang diterapkan dalam penelitian ini. Sedangkan algoritma yang dipakai adalah k-means clustering dengan metode elbow untuk menentukan jumlah clusternya yaitu sebanyak 3 cluster. Didapatkan bahwa Cluster 0 dengan kategori rendah terdapat 12 wilayah, cluster 1 dengan kategori sedang terdapat 14 wilayah, dan cluster 2 dengan kategori tinggi terdapat 1 wilayah. Evaluasi koefisien siluet bernilai 0,48 dengan kriteria struktur kurang baik. Dari hasil tersebut dapat disimpulkan bahwa sebaiknya Pemerintah Jawa Barat diharapkan lebih menekankan penanggulanagan kasus penyakit Diare pada cluster 2 dengan kategori tinggi yaitu pada Kabupaten Bogor.

***Kata Kunci**: Analisis Data, Data Mining, K-Means Clustering, Python*

Data Analysis of Diarrhea in West Java Using Data Mining with the K-Means Clustering Algorithm

ABSTRACT

Diarrheal disease is common in developing countries. This disease often causes cases of death and health problems in the world. The province in Indonesia that has the highest rate of diarrhea cases is West Java where the number of cases has increased and decreased in the last six years. Therefore, the Provincial Government of West Java must pay more attention in terms of handling this Diarrhea disease. This research was conducted by city clustering the level of diarrheal disease in West Java in 2016-2021. The purpose of this study is to determine the level of risk of diarrheal disease in each city in West Java in order to assist the Government in choosing a strategy for handling this case. The Knowledge Discovery in Database method is the method applied in this study. While the algorithm used is k-means clustering with the elbow method to determine the number of clusters, namely as many as 3 clusters. It was found that Cluster 0 with the low category had 12 regions, cluster 1 with the medium category had 14 regions, and cluster 2 with the high category had 1 region. Evaluation of the silhouette coefficient is 0.48 with unfavorable structural criteria. From these results it can be concluded that the West Java Government should be expected to put more emphasis on the management of Diarrhea cases in cluster 2 with a high category, namely in Bogor Regency.

***Keywords**: Data Analysis, Data Mining, K-Means Clustering, Python*

I. PENDAHULUAN

Negara Indonesia merupakan negara yang saat ini masih memiliki banyak sekali macam persebaran penyakit yang mematikan yang salah satunya yaitu diare. Angka penyakit diare di Indonesia masih tergolong tinggi untuk beberapa tahun terakhit. Menurut WHO penyakit diare menyebab kematian terbanyak nomor dua padakasus penyakit anak-anak khususnya yang meiliki umur 5 tahun kebawah, serta angka kematian yang didapatkan sekitar 525.000 setiap tahun. Provinsi di Indonesia yang memiliki kasus penyakit diare yang banyak yaitu Jawa Barat. Situs *Open Data* Jawa Barat mengungkapkan bahwa terjadi fluktuasi tingkat kasus diare pada Provinsi Jawa Barat dari rentang tahun 2016 sampai 2021.

Dengan fakta tersebut, dapat dinyatakan bahwa penanganan kasus diare pada Provinsi Jawa Barat kurang bisa dikendalikan dengan baik. Hal tersebut bisa disebabkan karena strategi penanganan yang dilakukan Dinas Kesehatan Provinsi Jawa Barat kurang tepat. Salah satu strategi yang dapat digunakan yaitu dengan sistem prioritas pada tingkat risiko kasus penyakit yang tertinggi berdasarkan kota pada Provinsi tersebut. Dinas Kesehatan Provinsi Jawa Barat perlu mencari langkah yang tepat dalam menangani kasus ini. Penelitian ini bertujuan untuk membantu Dinas Kesehatan Provinsi Jawa Barat untuk melakukan segmentasi dan pengelompokan Kabupaten/Kota di Provinsi Jawa Barat berdasarkan jumlah dan kerentanan terhadap kasus penyakit diare tersebut. Dengan dilakukannya penelitian ini, harapannya Pemerintah Provinsi Jawa Barat dapat mengambil hasil dari penelitian ini untuk bahan pertimbangan kebijakan penanganan wabah penyakit diare pada waktu selanjutnya dengan cara yang lebih efektif melalui sistem prioritas tingkat kasus ataupun metode lainnya. Pengelompokan penyakit diare di Jawa Barat ini menggunakan teknik metode *data mining* dengan menggunakan algoritma *k-means clustering*.

K-means clustering sendiri adalah algoritma yang beroperasi dengan cara membagi data data kedalam *cluster- cluster* sehingga didapatkan kelompok *cluster* yang memiliki karakteristik yang mirip. Dengan metode ini diharapkan Dinas Kesehatan Jawa Barat dapat mengetahui Kota mana yang harus didahulukan penanganannya sesuai dengan urutan rentan penyakit yang dihasilkan, sehingga kedepannya dapat memprioritaskan wilayah yang banyak memiliki kasus diare, agar angka tingginya penyakit diare di daerah itu bisa berkurang. Cara untuk menentukan jumlah *cluster* bisa dilakukan secara random atau acak, namun pada penelitian ini peneliti memilih metode *elbow*. Metode ini dipilih agar nilai yang dihasilkan dapat dikontrol dan tidak menutunkan hasil evaluasi karena hal tersebut sangat mungkin untuk terjadi akhibat dari pemilihan jumlah *cluster* yang kurang tepat.

Dengan hal tersebut, peneliti akan mengelompokkan data penyakit diare di wilayah Provinsi Jawa Barat menggunakan klastering *k-means* statistik dengan metode *elbow* untuk pemilihan jumlah clusternya. Serta penelitian ini dibuat dengan tujuan untuk dapat memberikan suatu data pengelompokan Kabupaten/Kota berdasarkan jumlah kasus diare yang ada agar dapat membantu Dinas Kesehatan Jawa Barat dalam hal penyaranan prioritas penanganan dari penyakit tersebut.

II. TINJAUAN PUSTAKA

A. *Diare*

Diare merupakan penyakit yang banyak ditemui di negara berkembang seperti Indonesia. Penyakit ini juga sering menyebabkan kasus kematian (Trianto, 2018). Diare diindikasikan dengan seringnya buang air besar beberapa kali dalam sehari. Penyakit ini juga memiliki ciri yaitu pada konsistensi tinja yang tidak padat dan terkadang didapatkan lender ataupun darah. Penyakit ini dapat menular melalui media makanan dan minuman, dan madia yang lain. Keadaan lingkungan tidak sehat yang tinggi menyebabkan penyakit diare ini masih merupakan suatu masalah yang banyak ditemukan di negara berkembang. Tidak hanya pada negara berkembang saja, bahkan diduniapun penyakit ini masih merupakan penyakit yang memiliki tingkat persebaran yang tinggi. Menurut *United Nations Children's Fund*

(UNICEF), dampak wabah penyakit Diare pada tahun 2015 menyebabkan banyak dari anak yang berusia dibawah lima tahun mengalami kematian. Lebih tepatnya adalah satu dari sepuluh anak meninggal dunia diakibatkan penyakit diare ini (Ria Manurung dkk, 2020). Penyakit ini diakibatkan karena adanya infeksi sistem saluran pencernaan seperti lambung dan usus baik usus halus, yang diakibatkan oleh adanya virus, bakteri, dan parasit. Penyebaran dari organisme ini dapat melalui makanan, minuman atupun benda benda lain yang masuk kedalam mulut. Lingkungan dan juga orang yang mempunyai kebiasaan yang kurang bersih juga merupakan faktor utama yang menyebabkan penyakit ini cepat untuk menular. Kematian yang disebabkan penyakit ini kebanyakan diakibatkan oleh faktor lain, seperti kurangnya cairan dari tubuh akibat seringnya buang air besar yang konsistensi tinjanya cair, ataupun kurangnya gizi dan juga adanya infeksi yang parah didalam area pencernaan pengidap (J dkk., 2019). Terdapat tiga faktor terjadinya diare. Yang pertama yaitu faktor lingkungan, faktor lingkungan yang kumuh dan buruk dapat berdamapak pada banyaknya organisme seperti virus dan bakteri yang dapat menyebabkan berbagai penyakit. Faktor kedua yaitu faktor individu, individu yang kurang menjaga diri dari kebersihan dan juga tidak menjaga kondisi serta nutrisi yang ada pada dalam tubuhnya akan menambah tingkat kerawanan individu tersebut untuk tertular penyakit diare ini. Faktor yang terakhi adalah faktor perilaku seperti kondisi fasilitas pembuangan kototran yang kurang baik serta kehigenisan pada saat makan, minum, ataupun saat beraktivitas lainnya (Hutasoit, 2020).

B. *Data Mining*

Banyak penelitian dan pengembangan telah dilakukan pada *data mining* dalam beberapa tahun terakhir. Nama *data mining* sudah ada sejak tahun 1990-an, ketika pekerjaan data mining menjadi terkenal di berbagai bidang mulai dari akademisi hingga kedokteran. (Siregar, 2018). Penambangan data merupakan kegiatan *looping* dan interaktif untuk menemukan suatu struktur tertentu dan sebuah pola yang valid (lengkap), yang didapatkan dari proses *extract* data yang sangat banyak (*massive database*). Penambangan data merupakan kegiatan menemukan tren yang diinginkan dan bertujuan menemukan wawasan atau *insight* yang didapatkan dari *database* yang *massive* untuk keperluan pengambilan keputusan dan evaluasi di masa yang akan datang (Syahdan & Sindar, 2018). Ada juga yang berpendapat bahwa penambangan data adalah sebuah analisis untuk melihat berbagai data yang ada guna menhubungkan suatu struktur yang ditemui serta dapat meringkas hasil temuan tersebut sehingga didapatkan suatu gagasan yang mudah dipahami dan bermanfaat untuk tujuan pemilik data. Penambangan data adalah gabungan dari ilmu permesinan, ,statistik, basis data, dan teknik visualisasi untuk mengabil intisari dari basis data besar. (Utomo & Mesran, 2020). Ada istilah lain untuk *data mining* itu sendiri, yaitu penemuan pengetahuan dan pengenalan pola. *Data mining* atau penambangan data itu sendiri memiliki fungsi yang utama untuk mengekstraksi wawasan yang masih belum terlihat di dalam blok data, sehingga didapatkan penemuan pengetahuan yang sangat berguna. Di sisi lain, istilah *pattern recognition* atau pengenalan juga merupakan tujuan *data mining* karena digunakan untuk menemukan pola tersembunyi di blok data (Zulfa dkk., 2021).

C. *Clustering*

Metode dalam penambangan data yang mempunyai sifat *unsupervised* adalah *clustering* ini (Kambey dkk, 2020). Proses pengklasteran ini dilakukan dengan mengelompokkan suatu titik data bedasarkan kemiripan antara titik satu dan titik lain. Sehingga nantinya terdapat beberapa kelompok yang didalamnya beranggotakan titik-titik yang mempunyai karakteristik mirip. Kemiripan ini ditentukan berdasarkan informasi yang tersedia didalam data (Herlinda & Darwis, 2021). Algoritma hierarkis pada analisis pengklasteran ini bekerja dengan mencari *cluster* secara berurutan, dengan *cluster* ditentukan terlebih dahulu, sedangkan algoritma parsial bekerja dengan menentukan semua *cluster* yang didasarkan pada waktu. (R. A. Indraputra & R. Fitriana, 2020). Metode ini mempunyai

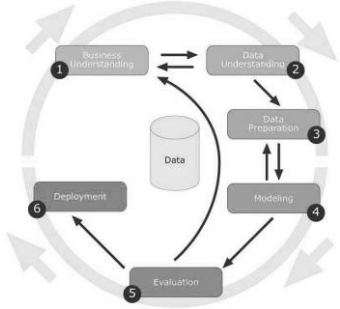
tujuan khusus yaitu untuk mengelompokkan sekumpulan data/objek yang mirip kedalam sebuah kelompok. Selain itu, *clustering* juga bertujuan untuk memberikan sela antar *cluster* seluas-luasnya. (Bahauddin dkk., 2021). Jadi dapat disimpulkan bahwa analisis *cluster* adalah alat pemisahan objek menjadi beberapa kelompok dan objek dalam kelompok yang sama-sama mempunyai kemiripan yang identik (Muningsih dkk., 2021).

D. *Algoritma K-Means*

K-means clustering disebut *non-hierarchical clustering* yang bekerja dengan membagi data agar dapat menjadi *cluster-cluster* sehingga data dalam *cluster* yang sama dikelompokkan jika mempunyai ciri khas yang identik, sedangkan data dengan yang tidak mirip akan dikelompokkan dalam *cluster* lain. (Amelia dkk., 2022). Algoritma *K-Means* merupakan algoritma segmentasi karena nilai *centroid* awal untuk menentukan jumlah kelompok awal. Algoritma *K-Means* menjadikan suatu kelompok data-data pada sebuah *dataset* ke dalam *cluster* berdasarkan jarak terdekatnya ke nilai *centroid* pertama yang cara pemilihan *centroid* tersebut adalah *random* (*centroid* awal). Semua jarak data dihitung menggunakan rumus jarak *Euclidean*. Data yang dekat dengan *centroid* membentuk *cluster*. Proses ini berlanjut sampai setiap kelompok tidak ada perubahan (Priati, 2019). Langkah awal dari algoritma ini adalah dipilihnya K secara *random*, setelah itu adalah menetapkan nilai sementara ke pusat *cluster* atau biasa dikenal dengan *centroid*. Hitung jarak dari setiap datum yang diberikan ke setiap *centroid* menggunakan rumus *Euclidean* sampai ditemukan jarak terdekat dari setiap datum ke *centroid*. Klasifikasikan setiap titik data berdasarkan seberapa dekat dengan pusat massa. Lakukan langkah tersebut hingga nilai titik berat tidak berubah (stabil) (Aswir & Misbah, 2018). *K-means clustering* adalah metode analisis *cluster* yang bertujuan untuk membagi objek menjadi k *cluster* kemudian mengamati dimana setiap objek yang dicluster didapatkan rata-rata terdekatnya. Algoritma *K-Means* merupakan algoritma evolusioner dimana operasinya memiliki arti yang sama dengan nama algoritmanya. Algoritma ini mengelompokkan pengamatan ke dalam kelompok k, dimana k adalah parameter masukan. Semua data kemudian ditempatkan pada setiap observasi dalam *cluster* berdasarkan kedekatan observasi dengan ukuran massa (Kamila dkk., 2019).

III. METODE PENELITIAN

Metodologi CRISP-DM yang dilengkapi dengan metode *clustering* menggunakan algbuat pada sekitr tahun 1996 oleh seorang analis dari beberapa industry yang sangat terkenal pada masa itu. CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan suatu standarisasi dari penambangan data yang telah dikembangkan sedemikian rupa agar data yang berhasil melewati setiap langkah tersebut akan menjadi terstruktur dan lebih efisien. (Hasanah dkk., 2021). Terdapat juga pengertian bahwa CRISP-DM adalah suatu proses penambangan data yang didalamnya dilakukan pen-standardisasian guna memecahkan suatu masalah bisnis (Feblian & Daihani, 2018).



Gambar. 1. Tahapan Metode CRISP-DM
Sumber: Hasanah (2021)

A. *Business Understanding*

Fase pertama yang dilakukan dalam metode CRISP-DM adalah analisis permasalahan yang akan diteliti pada penelitian ini. Pada *step* pertama memiliki empat proses yang harus dilakukan secara runtut yaitu:

a. *Determine Business Objectives.* Proses yang pertama berupa menentukan tujuan bisnis. Pada proses ini merupakan proses penetapan suatu tujuan dari penelitian agar dapat menghasilkan hasil yang sesuai dengan keinginan. Penetapan dilakukan dengan meng-*explore* data tentang dampak banyaknya penyakit diare untuk mengetahui wawasan apa yang bisa di temukan oleh peneliti untuk bahan rekomendasi kepada Dinas Kesehatan Jawa Barat.
b. *Asses Situation*, proses ini adalah proses dimana menganalisis kejadian yang telah terjadi, dengan mencari tahu langkah apa saja yang sudah dilakukan oleh Dinas Kesehatan Jawa Barat serta apakah sudah melakukan pemetaan tentang penyakit diare di Wilayah Jawa Barat atau belum.
c. *Determine Data Mining Goals*, *step* yang ketiga adalah ditentukannya alat atau suatu teknik pengolahan dan pengambilan data. Pada kasus penelitian ini digunakan teknik penambangan data sebagai alat untuk mencapai tujuan dari penelitian ini.
d. *Plan Activities*, *step* yang keempat adalah menyimpulkan *plan* per *step-step* nya agar sebuah tujuan yang sudah ditetapkan dapat sepenuhnya tercapai.

B.*Data Understanding*

Pemahaman tentang data didapat merupakan *step* yang penting untuk dilakukan. Pengumpulan, pendeskripsian dan penggambaran data dilakukan untuk kemudian dilanjutkan dengan menjelajahi data atau yang disebut *data exploratory* yang berguna dan bermanfaat untuk proses penelitian. Setelah proses tersebut selesai, selanjutnya dilakukan pengidentifikasian suatu problem dengan mengaitkan data-data yang telah dikumpulkan. Langkah terakhir yaitu dengan mempelajari semua data yang didapat agar peneliti dapat secara penuh mengambil intisari atau *insight* yang tepat dari data tersebut. Pada *website Open* Data Jabar didapatkan sebuah open data dengan judul "Jumlah Kasus Penyakit Diare Berdasarkan Kabupaten/Kota di Jawa Barat" pada tahun 2016 sampai 2021.

C.*Data Preparation*

Pada tahap ketiga yaitu persiapan data. Tahap persiapan ini dilakukan dengan memilah data yang benar-benar dipergunakan dalam proses penelitian. Pada tahap ini dilakukan sebuah proses yang banyak dikenal dengan istilah *preprocessing* dimana proses tersebut memiliki tiga tahap yang dilakukan yaitu:

a. *Data Selection.* Pada dataset memiliki kolom dan atribut yang sangat banyak. Dalam melakukan penelitian tidak semua atribut tersebut terpakaai. Maka dati itu digunakan seleksi data untuk mengambil atribut data yang akan dijadikan bahan untuk pengolahan pada penelitian ini yang tujuan akhirnya dapat sesuai dengan yang diharapkan.
b. *Data Preprocessing*, proses yang kedua ini biasa dikenal dengan *data cleansing*. *Data preprocessing* merupakan tahap menghilangkan *missing value*, data yang anomali, dan juga *noise data*. *Data Cleaning* sangat penting dilakukan karena tahap ini merupakan tahap yang krusial Jika sumber data penuh dengan nilai yang kosong dan tidak normal maka juga akan berpengaruh dengan hasilnya juga. Istilah yang sering digunakan dalam hal *data cleaning* adalah "*Garbage In, Garbage Out*" yaitu jika data yang masuk penuh dengan sampah, maka data yang dihasilkanpun juga akan seperti sampah.
c. *Data Transformation*. Tidak semua struktur data sesuai dengan metode yang akan digunakan dalam penelitian. Terkadang peneliti harus mengelompokkan dan mengubah struktur data tersebut agar lebih proper saat diolah pada step modelling. Proses transformasi ini dapat berupa proses *transpose* kolom menjadi baris ataupun proses pengubahan lainnya.

D. *Modelling*

Pada fase yang keempat, pemodelan pada data dilakukan dengan menggunakan *query* python dengan teknik penambangan data dengan algoritma *k-means*. Pada tahap ini digunakan bahasa *python* dengan bantuan *tools* yaitu*Google Collaboratory*. Pada bagian pemodelan data akan diolah data mentah menjadi sebuah data yang siap disajikan. Hasil tahap *modelling* ini selanjutnya akan divisualisasikan dengan menggunakan *tools Google Data Studio* yang akan menggambarkan titik-titik lokasi berdasarkan *cluster* yang terbentuk.

E.*Evaluation*

Fase kelima dalam metodologi ini yaitu tahap *evaluation*. Evaluasi merupakan tahap yang juga penting untuk dilakukan. Untuk tahap evaluasi dalam metode ini adalah dilakukannya pengujian kualitas dari *cluster* yang terbentuk sehingga nantinya diketahui bahwa *cluster* yang dihasilkan telah sesuai dengan yang telah ditetapkan sebelumnya atau belum. Metode indeks siluet atau yang biasa disebut *silhouette index* adalah ukuran validasi berbasis kriteria internal. *Silhouette index* bekerja dengan membandingkan jarak rata-rata antar cluster yang tebentuk (Nahdliyah dkk., 2019).

IV. HASIL DAN PEMBAHASAN

Penganalisisan tentang persebaran dan pemetaan merupakan hasil dari penelitian ini. Pemetaan dilakukan pada kota yang banyak terkena penyakit diare di Provinsi Jawa Barat dengan metode klastering bertipe *k-means.* Selanjutnya hasil dari analisis *clastering* tersebut diubah menjadi visualisasi gambar menggunakan *tools Google Data Studio.*

A. *Business Understanding*

Pada *step* ini dilakukan penentuan tentang latar belakang dan tujuan. Peneliti harus memahami akan pentingnya mengetahui tingkat penyakit diare dari kota yang berada di Provinsi Jawa Barat agar mempermudah Dinas Kesehatan Jawa Barat dalam proses penanganan dan pencegahan penyakit diare. Tidak hanya itu juga dampak buruk penyakit diare yang dapat berdampak ke banyak sektor juga menjadi sebuah penjelasan masalah dalam penelitian ini.

a. *Determine Business Objectives.* Pada proses yang pertama dilakuakn proses riset data melalui internet mengenai data Dinas Kesehatan di Jawa Barat. Pada tahap ini didapatkan bahwa hanya terdapat *dataset* total penyakit diare dari berbagai kota di Jawa Barat dan belum di klasterisasi menurut tingkat tingginya kasus penyakit diare pada berbagai kota tersebut. Penelitian ini berfokus pada pemetaan penyakit diare di Jawa Barat.
b. *Asses Situation*, proses analisis fakta situasi terkini dilakukan dengan melakukan riset melalui situs *Open Data Jabar* serta situs Dinas Kesehatan Jawa Barat. Hasil ditemukan bahwa Dinas Kesehatan Jawa Barat belum melakukan pemetaan tentang penyakit diare di Wilayah Jawa Barat. Selain itu berdasarkan data dapat dilihat bahwa angka kasus penyakit diare belum juga menurun pada tahun sebelumnya.
c. *Determine Data Mining Goals*. Berdasarkan hasil *business understanding* yang sebelumnya, selanjutnya dilakukan penentuan proses secara teknik. Dalam penelitian ini digunakan metode *data mining* dalam pengolahan datanya dikarenakan metode tersebut sesuai dengan tujuan dan sumber data yang didapatkan.
d. *Plan Activities*. Perencanaan dilakukan dengan menetapkan *tools* yang akan digunakan. *Tools* tersebut yaitu *Google Colaboratory* yang digunakan dalam pengolahan data dengan bahasa pemrograman *Python* dan juga *Google Data Studio* yang digunakan dalam hal visualisasi pemetaan *clustering* dari setiap kota di Provinsi Jawa Barat.

B.*Data Understanding*

Tahapan awal sebelum mengolah data yaitu dengan memahami data mentah yang didapatkan. Tahapan awal dari fase ini adalah tahap *data collection* atau pengumpulan data.

Dataset yang akan diolah dikumpulkan dan didapat dari situs resmi *Open Data Jabar* dengan judul "Jumlah Kasus Penyakit Diare Berdasarkan Kabupaten/Kota di Jawa Barat". Data yang didapat mempunyai rentang waktu mulai tahun 2016 sampai dengan tahun 2021. Berikut merupakan tabel *dataset* awal.

Tabel 1
Dataset Awal

id	kode_provinsi	nama_provinsi	kode_kabu-paten_kota	nama-kabu-paten_kota	jumlah_ka-sus	satuan	ta-hun
1	32	JAWA BARAT	3201	KABUPATEN BOGOR	159405	ORANG	2016
2	32	JAWA BARAT	3202	KABUPATEN SUKABUMI	37369	ORANG	2016
3	32	JAWA BARAT	3203	KABUPATEN CIANJUR	41709	ORANG	2016
4	32	JAWA BARAT	3204	KABUPATEN BANDUNG	90337	ORANG	2016
5	32	JAWA BARAT	3205	KABUPATEN GARUT	96111	ORANG	2016
…	…	…	…	…	…	…	…
158	32	JAWA BARAT	3275	KOTA BEKASI	9980	ORANG	2021
159	32	JAWA BARAT	3276	KOTA DEPOK	10170	ORANG	2021
160	32	JAWA BARAT	3277	KOTA CIMAHI	1115	ORANG	2021
161	32	JAWA BARAT	3278	KOTA TASIKMALAYA	9123	ORANG	2021
162	32	JAWA BARAT	3279	KOTA BANJAR	2053	ORANG	2021

Sumber: Data Primer Diolah, 2022

Dataset awal didapatkan dengan *format* .csv yang dapat langsung diolah kedalam *tools Google Colaboratory*. Pada *dataset* awal terdapat 162 baris yang merupakan data Kota yang ada di Provinsi Jawa Barat serta 8 kolom yang merupakan indikator dari setiap data kota. Selanjutnya dilakukan peng-*explor*an data lebih lanjutt untuk dapat mengetahui struktur dan juga atribut apa saja yang terdapat pada *dataset* awal.

C.*Data Preparation*

a. *Data Selection.* Pada tahapan ini dilakukan proses pemilihan kolom yang hanya diperlukan untuk pegolahan data dalam penelitian ini. Dataset yang diambil merupakan data angka kasus diare dari wilayah kabupaten/kota di Provinsi Jawa Barat mulai 2016 sampai 2021. Atribut yang dipilih dari *dataset* ini terdapat hanya 3 atribut saja. Atribut atau kolom tersebut yaitu kolom nama_kabupaten_kota, kolom jumlah_kasus, dan kolom tahun. Kolom (atribut) tersebut dilipih karena relevan sesuai kebutuhan pengolahan data sesuai dengan tujuan penelitian. Proses penyeleksian dilakukan dengan dan didapatkan hasil pada Tabel 2.

Tabel 2
Hasil Data Selection

nama-kabupaten_kota	jumlah_kasus	tahun
KABUPATEN BOGOR	159405	2016
KABUPATEN SUKABUMI	37369	2016
KABUPATEN CIANJUR	41709	2016
KABUPATEN BANDUNG	90337	2016
KABUPATEN GARUT	96111	2016
…	…	…
KOTA BEKASI	9980	2021
KOTA DEPOK	10170	2021
KOTA CIMAHI	1115	2021
KOTA TASIKMALAYA	9123	2021
KOTA BANJAR	2053	2021

Sumber: Data Primer Diolah, 2022

b. *Data Preprocessing*, Pada tahap ini akan dilakukan proses pengecekan nilai yang kosong atau *missing value*. Hasil yang didapat ternyata *dataset* yang digunakan tidak memiliki *missing value*. Proses melihat data yang hilang dilakukan dengan menuliskan perintah *library python* (*.isna.sum*)dan dihasilkan nilai 0 pada setiap kolom yang berarti tidak ada data yang kosong sehingga proses *data preparation* dapat dilanjutkan ke fase selanjutnya.

c. *Data Transformation.* Transformasi pada data yang akan diolah dilakukan dalam penelitian ini. Hal ini disebabkan karena struktur data yang didapat tidak sesuai dengan struktur data yang akan diolah. Pada tabel awal, nilai tahun berada pada satu kolom yang sama, dan kolom jumlah berada pada kolom yang berbeda. Hal itu dapat menyulitkan pengolahan data sealnjutnya. Maka dari itu perlu diubah struktur data tersebut menjadi satu kesatuan kolom tahun yang berisi jumlah kasus penyakit diare tersebut. Tak hanya itu, kolom tahun juga dipisah berdasarkan tahun yang ada. Dilakukan juga proses *dropping* dari kolom tahun yang lama agar tidak mengganggu saat proses pengolahan. *Dataset* yang telah diubah strukturnya dapat dilihat pada Tabel 3.

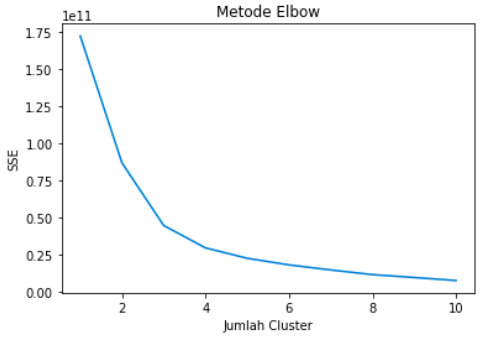
Tabel 3
Hasil Transformasi Data

nama-kabupaten_kota	2016	2017	2018	2019	2020	2021
KABUPATEN BOGOR	159405	122301	157705	161066	164382	91434
KABUPATEN SUKABUMI	37369	52505	74577	66589	67159	62891
KABUPATEN CIANJUR	41709	48291	80948	61103	61137	18179
KABUPATEN BANDUNG	90337	78273	59804	101933	64599	12893
KABUPATEN GARUT	96111	55401	103223	70805	107123	28764
…	…	…	…	…	…	…
KOTA BEKASI	22626	61196	79480	81106	68679	9980
KOTA DEPOK	37690	48247	62919	64984	67073	10170
KOTA CIMAHI	17795	12864	11464	16586	16751	1115
KOTA TASIKMALAYA	16808	14154	17646	17915	18121	9123
KOTA BANJAR	3186	3903	3327	4944	4949	2053

Sumber: Data Primer Diolah, 2022

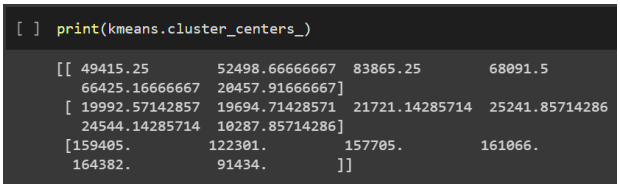
D. *Modelling*

Pada proses *modelling* digunakan teknik yang digunakan yaitu penambangan data atau *data mining*. Selanjutnya statistika klasteirng dengan algoritma *k-means* juga digunakan sebagai model metode yang dipilih. Pada algoritma ini, jumlah *cluster* harus ditentukan terlebih dahulu sebelum proses pembentukan nilai *k-means*. Salah satu cara menentukan jumlah *cluster* adalah dengan dilakukannya pengidentifikasian sebuah *elbow*. Dengan menggunakan metode yang diesbut *elbow method. Elbow method* dapat mengetahui nilai dari klaster yang paling relevan dan optimal. Nilai ini dapat diketahui dari bentuk *elbow* yang tergambar dalam grafik. Pembuatan grafik *elbow* dilakukan dengan mengetikkan *query* pada bahasa pemrograman python. Dan hasil grafik metode *elbow* dapat dilihat pada Gambar 2.



Gambar. 2. Hasil Grafik *Elbow*
Sumber: Data Primer Diolah (2022)

Menurut metode *elbow*, jumlah klaster yang relevan dan optimal terdapat pada angka 3. Hal tersebut dikarenakan pada titik nomor 3 telah membentuk sudut siku yang mulai melandai atau lurus atau dapat dikatakan sebagai ujung dari siku, sehingga nilai jumlah *cluster* yang optimal adalah 3. Setelah didapatkan klaster yang oprimal adalah 3, selanjutnya akan diolah menggunakan python programming untuk mengetahui hasil nilai *cluster* dari tiap kabupaten/kota yang terdapat pada data. Titik *centroid* juga perlu dilihat apakah nilai *centroid* tersebut tidak berubah lagi jika dibandingkan dari nilai sebelumnya. Jika tidak berubah, maka dapat dipastikan bahwa *centroid* tersebut metipakan *centroid* akhir.



Gambar. 3. *Centroid* Akhir *K-means*
Sumber: Data Primer Diolah (2022)

Setelah hasil *centorid* akhir sudah diketahui, pemodelan *cluster* sudah dapat dibentuk dan data telah masuk kedalam *cluster*. Berikut merupakan *query* dan hasil *cluster* dari tiap Kota dengan penyakit diare di Provinsi Jawa Barat.

Tabel 4
Hasil *Cluster* setiap Kabupaten/Kota

Nama-Kabupaten_Kota	2016	2017	2018	2019	2020	2021	Cluster
KABUPATEN BOGOR	159405	122301	157705	161066	164382	91434	2
KABUPATEN SUKABUMI	37369	52505	74577	66589	67159	62891	0
KABUPATEN CIANJUR	41709	48291	80948	61103	61137	18179	0
KABUPATEN BANDUNG	90337	78273	59804	101933	64599	12893	0
KABUPATEN GARUT	96111	55401	103223	70805	107123	28764	0
KABUPATEN TASIKMALAYA	25629	37393	171373	47361	47404	9686	0
KABUPATEN CIAMIS	31254	25294	30142	32270	32445	20250	1
KABUPATEN KUNINGAN	26554	22860	29012	29182	14971	12455	1
KABUPATEN CIREBON	74674	46215	60635	59208	59660	30706	0
KABUPATEN MAJALENGKA	19976	25546	32381	32536	34880	14495	1
KABUPATEN SUMEDANG	22718	24534	114991	31115	31170	13268	0
KABUPATEN INDRAMAYU	48287	36594	0	46669	46916	14437	1
KABUPATEN SUBANG	25	33438	42633	43087	43540	17415	1
KABUPATEN PURWAKARTA	20625	20187	25744	25998	26242	19472	1
KABUPATEN KARAWANG	61444	49573	60804	63556	64003	16777	0
KABUPATEN BEKASI	25251	74900	56509	101625	77768	15001	0
KABUPATEN BANDUNG BARAT	28045	35663	45897	45897	46305	11184	1
KABUPATEN PANGANDARAN	10074	8455	11789	10781	10840	2613	1
KOTA BOGOR	25345	23134	29614	30026	30026	5391	1
KOTA SUKABUMI	12849	6929	9019	8874	8929	5464	1
KOTA BANDUNG	57425	53456	81120	67713	81327	17180	0
KOTA CIREBON	19073	6705	15428	8621	8703	8563	1
KOTA BEKASI	22626	61196	79480	81106	68679	9980	0
KOTA DEPOK	37690	48247	62919	64984	67073	10170	0
KOTA CIMAHI	17795	12864	11464	16586	16751	1115	1
KOTA TASIKMALAYA	16808	14154	17646	17915	18121	9123	1
KOTA BANJAR	3186	3903	3327	4944	4949	2053	1

Sumber: Data Primer Diolah, 2022

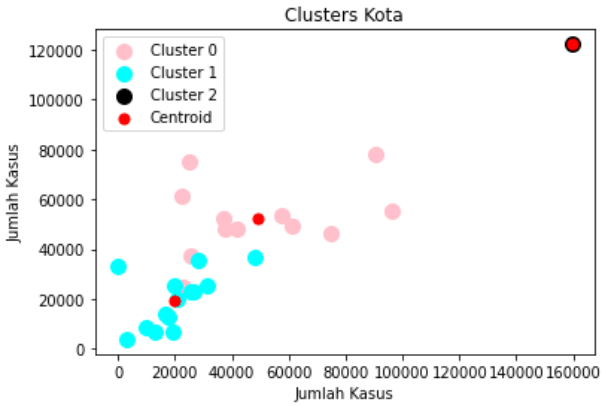
Dari hasil pengolahan data yang didapatkan, diketahui bahwa ada tiga *cluster* yaitu *cluster* 0 yang dikategorikan dengan tingkat rendah, *cluster* 1 yang dikategorikan tingkat sedang, dan *cluster* 2 yang dikategorikan dengan tingkat tinggi. Hasil data tersebut kemudian direkap dalam Tabel 5.

Tabel 5
Hasil Rekapitulasi Data Kabupaten/Kota Berdasarkan *Cluster*

Cluster	Nama Daerah	Kategori
0	Kabupaten Sukabumi Kabupaten Cianjur Kabupaten Bandung Kabupaten Garut Kabupaten Tasimalaya Kabupaten Cirebon Kabupaten Sumedang Kabupaten Karawang Kabupaten Bekasi Kota Bandung Kota Bekasi Kota Depok	Rendah
1	Kabupaten Ciamis Kabupaten Kuningan Kabupaten Majalengka Kabupaten Indramayu Kabupaten Subang Kabupaten Purwakarta Kabupaten Bandung Barat Kabupaten Panganda-ran Kota Bogor Kota Sukabumi Kota Cirebon Kota Cimahi Kota Tasimalaya Kota Banjar	Sedang
2	Kabupaten Bogor	Tinggi

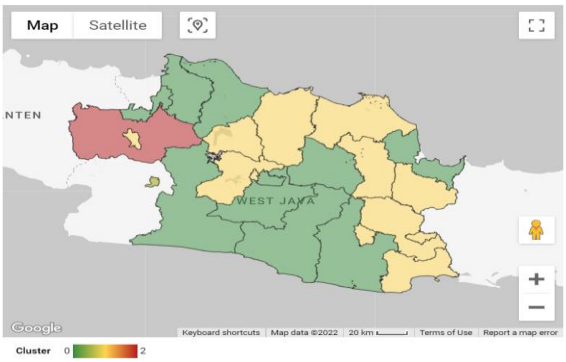
Sumber: Data Primer Diolah, 2022

Pada tabel hasil pengolahan data klastering didapatkan 1 kabupaten/kota masuk kedalam klaster 2 atau kategori tinggi, selanjutnya terdapat 5 kabupaten/kota yang tergolong dalam kalster 1 atau dengan kategori sedang, dan terdapat 22 kabupaten/kota yang tergolong dalam klaster 0 atau kategori rendah. Selanjutnya dilakukan tahap visualisasi *clustering* dengan *scatter chart*. Pada Gambar 4. merupakan hasil dari *scatter plot* persebaran algoritma *k-means*.



Gambar. 4. *Query Scatter Plot Diagram*
Sumber: Data Primer Diolah (2022)

Hasil pengelompokan persebaran yang dibuat dapat dijadikan referensi untuk Pemerintahan Provinsi Jawa Barat dalam penyusunan strategi penanganan kasus penyakit diare ini. Selanjutnya untuk lebih dapat menggambarkan dari geografik setiap kabupaten dan kota menurut *Cluster*-nya, digunakan *tools Google Data Studio* dan didapatkan hasil seperti Gambar 5.



Gambar. 5. *Geo Chart* Hasil Pengelompokan *Cluster* Kota
Sumber: Data Primer Diolah (2022)

E.*Evaluation*

Setelah dilakukan tahap *modelling* dan didapatkan hasil nilai *k-means clustering*, tahap terakhir adalah tahap evaluasi. Tahap evaluasi ini dilakukan untuk dapat melihat kualitas dari *cluster* yang tebentuk. Pada tahap evaluasi ini metode *silhouette coefficient* diambil untuk mengecek apakah jumlah *cluster* yang terbentuk sudah tepat. Metode ini mengadopsi tipe persebaran antar objek dan seberapa jauh *custer* tersebut terpisah. Nilai akan bernilai positif jika nendekati 1 dan akan bernilai negatif saat mendekati 0, maka struktur dari klaster dapat dikategorikan klaster baik. Jika nilai *silhouette coefficient* = 0 maka akan didapatkan kesimpulan bahwa tidak ditemukannya struktur pada data tersebut. Seliain itu jika nilainya = -1 maka struktur klaster dikategorikasn debagai klaster yang *overlap*.

Tabel 6
Nilai *Silhoutte Coefficient* Berdasarkan Kaufman Dan Roussew

Nilai SC	Kualitas	Interpretasi
0,71 – 1,00	*Strong*	Klaster terbaik sudah ditemukan
0,51 – 0,70	*Medium*	Penempatan klaster yang wajar
0,26 – 0,50	*Weak*	Strukturnya lemah
<= 0,25	*No Structure*	Tidak ada struktur yang ditemukan

Sumber: Fahmi, 2021

Berdasarkan hasil pengujian pada python dihasilkan nilai *index silhouette coefficient* sebesar 0.48 . Jadi dapat disimpulkan bahwa hasil klaster untuk data penyakit diare di Jawa Barat mempunyai kualitas yang rendah dengan interpretasi penempatan *cluster* yang kurang wajar. Atau dapat juga dikatakan bahwa data data yang tersebar kedalam kluster belum menempati posisi yang semestinya. Hal ini disebabkan karena struktur dari nilai data yang ada kurang baik. Maka dari itu sebagai saran untuk penelitian selanjutnya bisa ditambahkan metode lainnya untuk dapat menentukan *cluster* dengan struktur yang kuat.

V. KESIMPULAN

Kesimpulan yang didapatkan dalam penelitian ini adalah terdapat 3 klaster yang terbentuk yang masing masing klaster tersebut memiliki anggota wilayah kabupaten/kota. Klaster tersebut dikategorikan menjadi rendah, sedang, dan tinggi. Kabupaten/kota yang masuk kedalam kategori rendah berjumlah 22 kabupaten/kota. Sedangkan, kabupaten/kota yang tergolong kedalam kategori sedang berjumlah 5 kabupaten/kota. Yang terakhir adalah kabupaten/kota yang tergolong dalam kategori 2 atau dengan kategori tinggi yaitu terdapat 1 kabupaten/kota. Selanjutnya didapatkan bahwa algoritma *k-means clustering* yang dibuat dan menghasilkan 3 *cluster* ini dinyatakan kurang optimal dalam pengelompokanya. Hal tersebut diindikasikan oleh nilai *silhouette coefficient* menghasilkan nilai dibawah 0,5 yaitu sebesar 0,48. Dari hasil tersebut mengindikasikan bahwa kualitas *cluster* tergolong rendah dan data belum masuk kedalam *cluster* yang tepat. Nilai *silhouette coefficient* yang rendah dapat disebabkan karena data yang kurang bervariasi dan kurang menunjukkan struktur data.

Saran untuk penelitian lebih lanjut agar dapat menggunakan metode selain *elbow* untuk menentukan besar nilai *cluster* yang ada, dikarenakan pada data penelitian ini metode *elbow* kurang begitu terlihat dalam pembentukan sudut siku. Hal tersebut dapat berpengaruh terhadap pemilihan jumlah *cluster* yang tepat. Maka dari itu untuk peneliti selanjutnya dapat menggunakan metode lain agar jumlah *cluster* dapat sesuai dan menghasilkan *Silhoutte Score* yang tinggi sehingga struktur *cluster* akan dapat dikategorikan sebagai *cluster* kuat dan hasil yang didapatkan lebih berkualitas.

PUSTAKA

Amelia, D., Padilah, T. N., & Jamaludin, A. (2022). Optimasi Algoritma K-Means Menggunakan Metode Elbow dalam Pengelompokan Penyakit Demam Berdarah Dengue (DBD) di Jawa Barat. *Jurnal Ilmiah Wahana Pendidikan*, *8*(11), 207–215. Diambil dari http://jurnal.peneliti.net/index.php/JIWP/article/view/1907

Aswir, & Misbah, H. (2018). No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. *Photosynthetica*, *2*(1), 1–13. Diambil dari http://link.springer.com/10.1007/978-3-319-76887-8%0Ahttp://link.springer.com/10.1007/978-3-319-93594-2%0Ahttp://dx.doi.org/10.1016/B978-0-12-409517-5.00007-3%0Ahttp://dx.doi.org/10.1016/j.jff.2015.06.018%0Ahttp://dx.doi.org/10.1038/s41559-019-0877-3%0Aht

Bahauddin, A., Fatmawati, A., & Permata Sari, F. (2021). Analisis Clustering Provinsi Di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Algoritma K-Means. *Jurnal Manajemen Informatika dan Sistem Informasi*, *4*(1), 1. https://doi.org/10.36595/misi.v4i1.216

Feblian, D., & Daihani, D. U. (2018). Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada Pt X. *Jurnal Teknik Industri*, *6*(1), 1–12. https://doi.org/10.25105/jti.v6i1.1526

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Herlinda, V., & Darwis, D. (2021). Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means. *Darwis, Dartono*, *2*(2), 94–99. Diambil dari http://jim.teknokrat.ac.id/index.php/JTSI

Hutasoit, D. P. (2020). Pengaruh Sanitasi Makanan dan Kontaminasi Bakteri Escherichia coli Terhadap Penyakit Diare. *Jurnal Ilmiah Kesehatan Sandi Husada*, *12*(2), 779–786. https://doi.org/10.35816/jiskh.v12i2.399

J, H., Oktavidiati, E., & Astuti, D. (2019). Pengaruh Pendidikan Kesehatan Media Video dan Poster terhadap Pengetahuan dan Sikap Anak dalam Pencegahan Penyakit Diare. *Jurnal Kesmas Asclepius*, *1*(1), 75–85. https://doi.org/10.31539/jka.v1i1.747

Kambey, G. E. I., & Dkk. (2020). Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia. *Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan dokumen teks bahasa Indonesia*, *15*(2), 75–82.

Kamila, I., Khairunnisa, U., & Mustakim, M. (2019). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau. *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, *5*(1), 119. https://doi.org/10.24014/rmsi.v5i1.7381

Muningsih, E., Maryani, I., & Handayani, V. R. (2021). Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa. *Jurnal Sains dan Manajemen*, *9*(1), 95–100. Diambil dari https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/10428/4839

Nahdliyah, M. A., Widiharih, T., & Prahutama, A. (2019). METODE k-MEDOIDS CLUSTERING DENGAN VALIDASI SILHOUETTE INDEX DAN C-INDEX (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018). *Jurnal Gaussian*, *8*(2), 161–170. https://doi.org/10.14710/j.gauss.v8i2.26640

Priati, ahmad fauzi. (2019). *Penerapan Data Mining Dengan Teknik Clustering Menggunakan Algoritma K-Means Pada Data Transaksi Superst*. (September), 15–19.

R. A. Indraputra, & R. Fitriana. (2020). K-Means Clustering Data COVID-19. *Jurnal Teknik Industri*, *10*(3), 275–282. Diambil dari https://www.trijurnal.lemlit.trisakti.ac.id/index.php/tekin/article/view/8428/6033

Ria Manurung, Oskar Ika Adi Nugroho, E. A. (2020). Jurnal abdidas. *Jurnal Abdidas*, *1*(3), 131–136.

Siregar, M. H. (2018). Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan). *Jurnal Teknologi Dan Open Source*, *1*(2), 83–91. https://doi.org/10.36378/jtos.v1i2.24

Syahdan, S. Al, & Sindar, A. (2018). Data Mining Penjualan Produk Dengan Metode Apriori Pada Indomaret Galang Kota. *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, *1*(2). https://doi.org/10.32672/jnkti.v1i2.771

Trianto, J. (2018). Penerapan Metode Forward Chaining untuk Diagnosa Penyakit Diare pada Anak Usia 3-5 Tahun Berbasis Mobile Android. *Jurnal Informatika Universitas Pamulang*, *3*(2), 98. https://doi.org/10.32493/informatika.v3i2.1520

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Zulfa, N., Auliya, R. I., & Zaenal, A. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi dan Sistem Informasi (JTSI)*, *2*(2), 100. Diambil dari http://jim.teknokrat.ac.id/index.php/JTSI