DOI : 10.36499/psnst.v12i1.7294
e-ISSN: 2964-2531
p-ISSN: 2964-5131

PENERAPAN ALGORITMA K-MEANS PADA KLASTERISASI DATA KAWALCOVID19.ID

Akhmad Pandhu Wijaya[1*], Agyztia Premana[2] da Nur Ariesanto Ramdhan[2]
[1] Jurusan Teknik Informatika, Fakultas Teknik, Universitas Wahid Hasyim
Jl. Menoreh Tengah X/22, Sampangan, Semarang 50236.
[2] Jurusan Teknik Informatika, Fakultas Teknik, Universitas Muhadi Setiabudi
Jl. Pangeran Diponegoro No.KM2, Wanasari, Brebes 52212.
[*]Email: [1] pandhudsn@unwahas.ac.id, [2] a.premana@umus.ac.id, [3] ariesantoramdhan@gmail.com

Abstrak

Indonesia merupakan Negara dengan jumlah pasien Covid-19 cukup besar mengingat kepadatan penduduknya yang meningkat dari masa ke masa, virus Corona Virus Disease 2019 atau yang disebut juga dengan Covid-19 muncul pertama kali pada Provinsi Hubei kota Wuhan-Tiongkok pada akhir Desember 2019. Pemerintah menerapkan berbagai cara untuk mendisiplinkan masyarakat Indonesia dalam beraktivitas dan menjaga diri dengan protokol kesehatan serta membatasi interaksi antar sesama, sehingga masyarakat kita mampu melewati gelombang pandemi. Perean Teknologi pada masa pandemi sangatlah besar terutama dalam menyebar luasan informasi secara digital yang mudah diakses dari perangkat dengan koneksi internet, hal ini menjadi doronga untuk diterapkannya teknik klasterisasi menggunakan K - Means untuk mengetahui angka peningkatan dan penurunan pada kasus Covid-19 serta membangun kewaspadaan bagi masyarakat. Data kasus tersebut diperoleh dari website kawalcovid19.id, dengan jumlah yang banyak tidak akan mudah dalam menganalisa dan mengambil informasi, maka dibutuhkan teknik untuk klasterisasi dengan tujuan untuk mendapatkan intisari dari kumpulan data tesebut. C4.5 menjadi algoritma yang dipilih untuk proses klasterisasi pada data ini, dengan dilakukannya proses kastering maka didapatkan sebanyak 28 Provinsi (82%) menjadi Provinsi dengan tingkat penularan rendah, dan 4Provinsi (12%) dengan tingkat penularan sedang, serta 2 Provinsi (6%) memiliki tingkat penularan tinggi dengan total 34 Provinsi di indonesia dengan kurun waktu sampling yang dibatasi.

***Kata kunci**: covid19 ,K-Means, kawalcovid19, klasterisasi*

1. PENDAHULUAN

Awal tahun 2020 menjadi saat yang diwaspadai oleh hampir seluruh negara di Dunia dengan munculnya virus covid19, *World Health Organization* (WHO) menetapkan covid-19 sebagai pandemi bertepatan pada tanggal 11 Maret 2020, maka diterapkanlah kebijakan secara global untuk menurunkan resiko penularan dengan membatasi berbagai kegiatan ditengah masyarakat (Dimas Bayu Febriyanto, Lovi Handoko, Wahyuli, Hanif Aisyah, 2021). Jumlah kasus di Indonesia pada saat ini 6.216.621 terkonfirmasi 6.010.545 dinyatakan sembuh data tersebut update pada tanggal 2 Agustus 2022 pada portal informasi kawalcovid19.id.

Perkembangan teknologi saat ini membawa manfaat pada berbagai bidang khususnya pada bidang kesehatan, hal tersebut sejalan dengan hadirnya kemudahan bagi masyarakat luas mengakses informasi dari portal online dengan berbagai bidang. Informasi yang didapatkan dari media online tidak selalu dapat terserap dengan baik dikarenakan kualitas informasi yang belum bisa dipastikan. Kemampuan dari manusia untuk menyaring informasi dari data yang banyak sangatlah terbatas, sehingga butuh proses klasterisasi yang dilakukan oleh mesin komputasi sehingga informasi dapat diserap secara optimal (Dimas Bayu Febriyanto, Lovi Handoko, Wahyuli, Hanif Aisyah, 2021). Penggalian informasi yang dilakukan dari kumpulan data dalam skala besar dilakukan dengan teknik Data Mining, yang bertujuan menjadi otomatisasi pencarian informasi dan selanjutnya dapat diambil keputusan terhadap setiap data yang dihasilkan. Data Mining sendiri berfokus pada pencarian informasi terhadap data yang tidak diketahui sebelumnya (Azwanti, 2018).

Penelitian mengenai klasterisasi data penyebaran covid-19 menggunakan algoritma C4.5 sebelumnya dilakukan oleh (Putra, 2021) untuk mengatasi tumpukan data covid-19 pada kota Pagar Alam yang diakumulasikan dan dioptimalisasi dengan klasterisasi tersebut, dari hasil pengujian

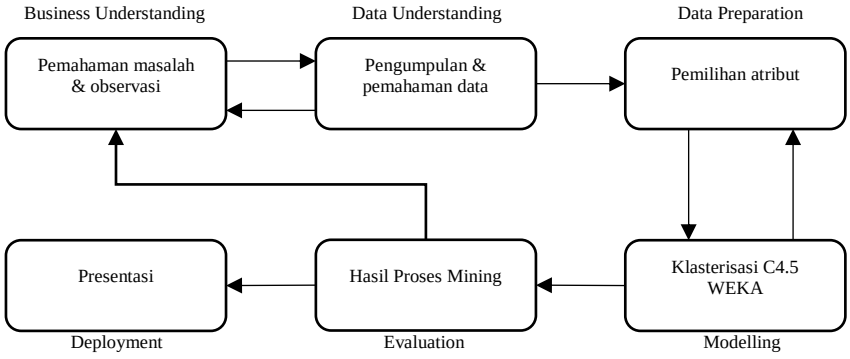
diperoleh hasil akurasi sebesar 86.67%. Sedangkan penelitian terhadap data pasien pada rumah sakit dengan pemanfaatan teknologi sebagai proses klasterisasi telah dilakukan oleh (Turnip and Siltionga, 2018), proses tersebut menggunakan metode C4.5 dan berfokus pada penyakit tuberkolosis dengan hasil akurasi yang sangat baik. Dari ulasan tersebut dapat diasumsikan bahwa Data Mining merupakan solusi dalam mengatasi permasalahan pada data khususnya data kesehatan untuk memberikan informasi yang akurat serta efisien, untuk menjadi bahan pertimbangan bagi pemangku kepentingan terhadap data tersebut. Secara umum proses Data Mininig melibatkan ilmu matematika, statistik, kecerdasan buatan, serta machine learning untuk melakukan ekstraksi informasi yang bermanfaat (Turnip and Siltionga, 2018).

Kemudahan dalam memahami hasil keluaran dari klasterisasi menggunakan metode C4.5 diutarakan oleh (Yuli Mardi, 2019) pada penelitiannya, disampaikan bahwa tahap data mining ini merupakan sebuah proses untuk mengetahui *Knowledge Discovery in Database* (KDD) yang dapat dilakukan dengan berbagai model seperti prediksi, perkiraan, atau klasterisasi seperti pada penelitian yang dilakukan oleh penulis. Tujuan dari rangkaian proses yang dilakukan pada data mining adalah memperoleh pengetahuan daripada database, dengan asumsi bermagai jenis tabel – tabel pada database yang tidak diketahui informasi didalamnya dan saling berelasi satu dengan yang lain. Penelitian ini menggunakan dataset dengan cakupan 34 provinsi, sedangkan beberapa penelitian tentang data covid-19 yang menjadi rujukan menggunakan data daerah kabupaten/kota, metode C.45 dipilih dikarenakan metode tersebut telah terbuktu efektif untuk mengolah berbagai macam data pada penelitian sebelumnya. Hasil yang diperoleh dari proses tersebut adalah sebuah pengetahuan *(knowledge base)* yang dipenuhi dalam rangka pengembilan keputusan.

2. METODOLOGI

2.1. CRISP-DM

Dalam pelaksanaan penelitian terdapat alur yang dapat dijelaskan secara sistematis dimulai dari identifikasi data sampai dengan presentasi, yaitu CRISP-DM (*Cross-Industry Standard Prodess for Data Mining*). Proses tersebut melibatkan beberapa hal seperti *Business Understanding, Data understanding, Data Preparation, Modeling, Evaluation, dan Deployment.* Alur tersebut dapat dijelaskan pada Gambar 1 (Purwanto, Primajaya and Voutama, 2020).



Gambar 1. Proses Data Mining CRISP-DM

a. *Business Understanding* / Proses pemahaman masalah : Pada tahap ini dilakukan observasi terhadap data yang menjadi tren pada masa kini dan memiliki potensi untuk diteliti, terpilihlah data covid-19 yang dirilis oleh kawalcovid19.id dalam skala nasional, diharapkan dapat memperkaya pembelajaran terhadap data terkini.
b. *Data Understanding* / Pemahaman data : Tahap ini melakukan pembelajaran pada dataset yang diunduh dan menyesuaikan dengan format yang dibutuhkan untuk proses selanjutnya. Pada tahap ini perlu adanya pemahaman atribut pada dataset.
c. *Data Preparation* / Pemilihan atribut : Pada tahap ini dilakukan pemilihan atribut yang dipergunakan sebagai meodeling dari daya yang ada, untuk kemudian dilakukan format data dari excel menjadi CSV sebagai format yang umum digunakan pada teknik pengolahan data.

e-ISSN: 2964-2531
p-ISSN: 2964-5131

d. Modeling → Klasterisasi data : Proses klasterisasi dilakukan menggunakan algoritma K-Means yang diuji menggunakan tools WEKA.
e. Evaluation → Hasil proses mining : Intepretasi terhadap data dilakukan pada preoses ini terhadap hasil pada proses sebelumnya, evaluasi secara mendalam dilakukan sebaik mungkin sehingga menghasilkan kesesuaian model yang didapat agar didapatkan kesesuaian terhadap hasil yang diharapkan. Proses evaluasi dilakukan menggunakan *Confusion Matrix* dengan nilai *accuracy, precision,* dan *recall.*
f. *Deployment* : Tahap ini dipaparkan laporan dari pengetahuan atau informasi yang diperoleh dari pengolahan data agar dapat dibaca dan diketauhi oleh khalayak umum. Hasil penelitian diharapkan berupa informasi klasterisasi penyebaran covid-19.

2.2. Data *Mining*

Cara untuk menguraikan dan mencari sebuah penemuan yang terkandung dalam *database* disebut juga dengan Data Mining, penemuan tersebut akan menjadi bermanfaat dan memberikan informasi yang akurat jika dilakukan dengan baik. Data Mining merupakan sebuah proses yang melibatkan ilmu matematika, statistik, kecerdasan buatan dan *machine learning* (Silitonga Irene Sri, 2017). Salah satu kegunaan dari Data Mining adalah klasterisasi yang berfungsi untuk melakukan pengelompokan terhadap jenis yang berbeda, maka klasterisasi membutuhkan paremeter yang telah ditentukan sebelumnya. Dalam pembahasan lain klasterisasi dipaparkan sebagai bentuk analisis data dengan fungsi ekstrak model untuk menggambarkan kelas data (Novandya, 2017). Pada literatur lain disebutkan bahwa Data Mining adalah proses pencarian yang dilakukan secara otomatis pada media penyimpanan dalam jumlah besar, serta dapat diistilahkan pula sebagai analisa untuk menyimpulkan data yang sebelumnya tidak pernah diketahui dan menemukan integrasi dengan data terbaru agar dapat dipahami bagi pengguna.

2.3. *Clustering*

Proses permodelan pada data yang tidak memiliki supervisi (unsupervised) yang sangat populer adalah *clustering*, pemrosesan dilakukan dengnan cara mengelompokkan data dengan sistem partisi (Sinaga, Hardinata and Fauzan, 2021). *Clustering* adalah proses untuk pengelompokan data berdasarkan sebuah pendekatan penting yang berperan mencari kesamaan dalam data dan mengelompokkan data sesuai dengan kesamaannya. Analoginya adalah sebuah cluster merupakan kumpulan dari benda – benda dengan kemiripan yang sama dan memiliki perbedaan terhadap benda pada *cluster* lainnya. Clustering juga dapat ditemukan pada beberapa aplikasi pengelompokan seperti *machine learning*, *data mining*, pengenalan pola dan lainnya (Putu *et al.*, 2015). Clustering juga memiliki prosedur dalam mempertimbangkan sebuah pendekatan yang penting untuk mencari kesamaan(Sinaga, Hardinata and Fauzan, 2021).

2.4. Algoritma *K-Means*

Algoritma K- Means merupakan salah satu algoritma dengan kemampuan melakukan partisi pada klaster yang diinginkan dengan melakukan pendefinisian centroid terlebih dahulu (Putu *et al.*, 2015). Digunakannya metode ini dengan alasan bahwa pengelompokan dataset harus dilakukan dengan membentuk kelompok – kelompok atau kelas – kelas taksonomi atau batryologi (Priyatman, Sajid and Haldivany, 2019).

Algoritma K – Means :

a. Tentukan k sebagai jumlah cluster yang akan dibentuk.
b. Tentukan k Centroid awal secara acak. (Centroid adalah titik pusat)

$v \frac{\sum_{i=1}^{n} x_i}{n}$; i = 1,2,3,...,n (1)

Dimana v :
centroid pada cluster
x_i : objek ke – i
n : banyaknya objek pada anggota cluster

c. Menghitung jarak setiap objek terhadap masing – masing centroid pada setiap cluster

$d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$; i = 1,2,3,…,n (2)

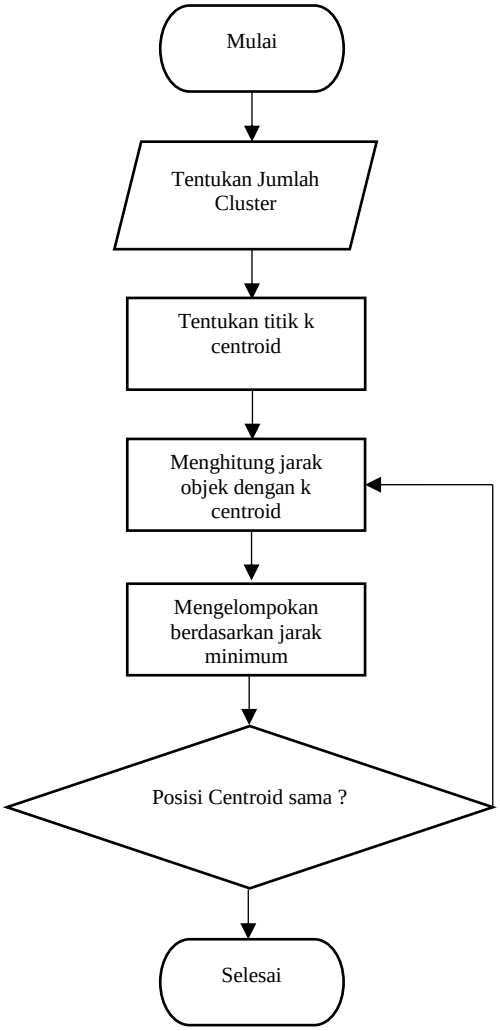
Keterangan : x_i

adalah objek *x* ke *i*

y_i adalah daya *y* ke *I*

n adalah banyaknya objek

d. Mengalokasikan setiap masing – masing objek pada centroid yang paling dekat.
e. Melakukan iterasi, dan menentukan posisi centroid baru menggunakan persamaan 1.
f. Mengulangi langkah 3 apabila centroid baru tidak memiliki kesamaan (Putu *et al.*, 2015)



Gambar 2. Flowchart Algoritma K – Means

3. HASIL DAN PEMBAHASAN

Bagian ini menjelaskan mengenai step dan runtutan penelitian dalam memperoleh hasil klaster dari data teridentifikasi positif covid-19 yang dirilis oleh kawalcovid19.id dan diklasterisasi dengan menggunakan metode C4.5, dan ditetapkan sejumlah 3 kelas. Tahap awal yang dilakukan adalah persiapan dataset yang ada dengan menentukan rule dan *classifier*. Dataset berisi tentang angka positif covid-19 pada periode Maret sampai dengna Oktober tahun 2020, dengan rincian sebagai berikut :

DOI : 10.36499/psnst.v12i1.7294
e-ISSN: 2964-2531
p-ISSN: 2964-5131

Tabel 1. Atribut Dataset

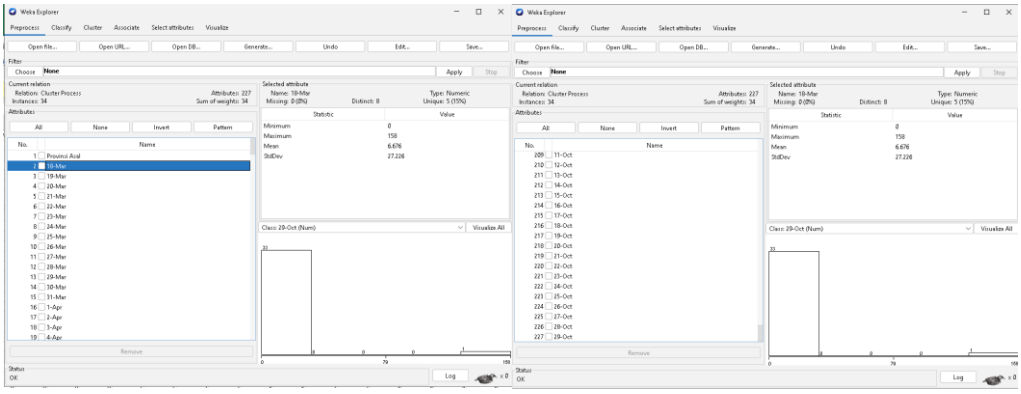
No	Nama Atribut	Jumlah Atribut
1	Periode penularan Maret sampai Oktober 2020	226
2	Seluruh Provinsi di Indonesia	34

Tabel 1 menggambarkan mengenai jumlah atribut yang ada pada data berdasarkan periode yang telah ditentukan, data tesebut diambil *sample* sejak tanggal 18 Maret hingga 29 Oktober 2020 yaitu 226 hari. Pada perioden tersebut sebanyak 34 provinsi melaporkan jumlah kasus covid-19 sehingga didapatkan data kasus covid pada periode tersebut taerhadap 34 Provinsi se-Indonesia.

Tabel 2. Dataset Penularan Covid-19

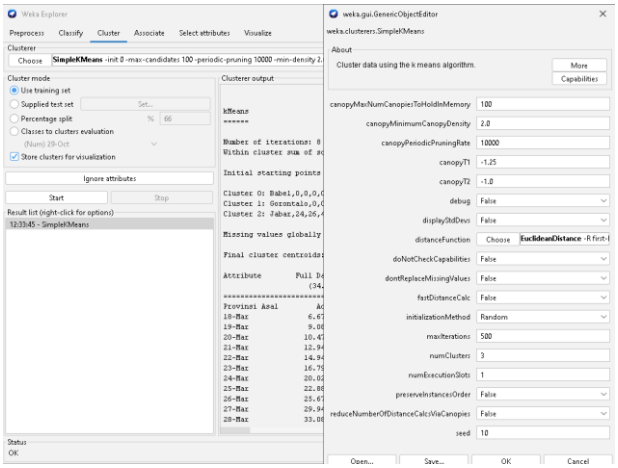
Provinsi / TGL	18-Mar	19-Mar	20-Mar	21-Mar	22-Mar	…	29-Oct
Aceh	0	0	0	0	0	…	7373
Bali	1	1	4	3	3	…	11647
Banten	17	27	37	43	47	…	9299
Babel	0	0	0	0	0	…	578
Bengkulu	0	0	0	0	0	…	1049
DIY	3	5	4	5	5	…	3744
Jakarta	158	210	215	267	307	…	104235
Jambi	0	0	0	0	0	…	1219
Jabar	24	26	41	55	59	…	35607
Jateng	8	12	12	14	15	…	33100
…	…	…	…	…	…	…	29-Oct
NTT	0	0	0	0	0	…	7373
Gorontalo	0	0	0	0	0	…	11647

Sampling dataset ditunjukkan pada tabel 2 memaparkan varian dari dataset dengan total sebanyak 34 Provinsi di Indonesia, dan sampling data covid dibatasi selama 226 hari sejak tanggal 18 Marer 2020 hingga 29 Oktober 2020.



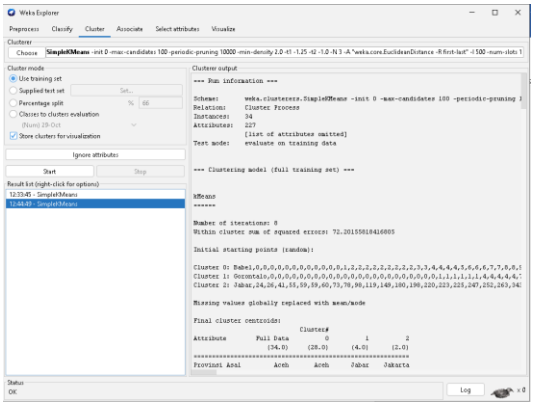
Gambar 3. Atribut Dataset

Pada percobaan *cluster process* terdapat 34 data *instance* dan 227 atribut yang akan dilakukan proses klasterisasi hingga mendapatkan hasil yang diharapkan, seperti pada gambar 3. Proses ini akan dilakukan berdasarkan jumlah data real yang didapatkan dari data pebyebaran virus covid-19 yang dirilis oleh kawalcovid19.id.



Gambar 4. Pemilihan klaster dan numclass

Algoritma yang dipilih pada tools adalah K-Means seperti pada gambar 4, dengan pembagian kelas menjadi 3 kelompok. Dengan asumsi kelompok ditempati oleh kelas dengan penularan rendah, sedang , dan tinggi.



Gambar 5. Iterasi cluster

Gambar 5 menjelaskan mengenai analisa metode tarhadap data yang diuji, dari jumlah data dan atribut yang ada telah terklaster menjadi 3 kelompok yang masing – masing memiliki nilai tertentu sebagai indikator kedekatan terhadap centroid pada metode K-Means. Angka tersebut yang akan menjadi parameter bagi data untuk dikelompokkan sesuai kelasnya.

3.1. **Hasil Simulasi**

```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      28 ( 82%)
1       4 ( 12%)
2       2 (  6%)
```

Gambar 6. Hasil cluster

DOI : 10.36499/psnst.v12i1.7294
e-ISSN: 2964-2531
p-ISSN: 2964-5131

Setelah proses klasterisasi berhasil maka didapatkan hasil seperti pada gambar 6 yaitu 3 buah kelas dengan jumlah provinsi yang berbeda, sebanyak 28 provinsi terdeteksi sebagai wilayah dengan tingkat penularan yang rendah antara lain provinsi : Aceh, Bali, Banten, Babel, Bengkulu, DIY, Jambi, Kalbar, Kaltim, Kalteng, Kaltara, 'Kep Riau', NTB, Sumsel, Sumbar, Sulut, Sumut, Sultra, Sulteng, Lampung, Riau, Malut, Maluku, Papbar, Papua, Sulbar, NTT, Gorontalo.

Sedangkan 4 provinsi dikelompokkan menjadi provinsi dengan penularan sedang / menengah antara lain : Jabar, Jateng, Kalsel, Sulsel. Serta 2 provinsi lain yang memiliki tingkat penularan paling tinggi adalah provinsi : DKI Jakarta dan Jawa Timur. Salah satu tujuan uji data pada penelitian ini adalah menjadi model latih dalam mendukung kebijakan diberlakukannya PPKM atau aturan – aturan lainnya yang mendukung upaya pemerintah dalam menekan angka penularan covid 19. Pemilihan tiga klister dianggap cukup dalam membantu mengupayakan sebuah keputusan yang tepat dengan kualitas data yang cukup baik (K-means, Solichin and Khairunnisa, 2020).

Secara tabulasi dapat disajikan seperti pada gambar 7 yang mengelompokkan secara visual julmah provinsi yang memiliki prioritas penanganan berdasarkan urgensi yang ada.

Gambar 7. Komposisi cluster

4. KESIMPULAN

Berdasarkan pengujian yang dilakukan dengan menerapkan metode K-Means pada dataset yang diperoleh dari situs kawalcovid19.id, dengan melakukan klasterisasi pada 34 provinsi di Indonesia. Maka dapat disimpulkan bahwa metode *clustering* K-Means dapat digunakan dengan baik dengan sebaran sebanyak 28 Provinsi (82%) menjadi Provinsi dengan tingkat penularan rendah, dan 4Provinsi (12%) dengan tingkat penularan sedang, serta 2 Provinsi (6%) memiliki tingkat penularan tinggi dengan total 34 Provinsi di indonesia dengan kurun waktu sampling yang telah ditentukan, yaitu dimulai dari tanggal 18 Maret 2020 sampai dengan 29 Oktober 2020 atau selama 226 hari. Pada penelitian selanjutnya proses klasterisasi dapat dilakukan dengan metode *clustering* lainnya.

DAFTAR PUSTAKA

Azwanti, N. (2018) 'Algoritma C4.5 Untuk Memprediksi Mahasiswa Yang Mengulang Mata Kuliah (Studi Kasus Di Amik Labuhan Batu)', *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 9(1), pp. 11–22. doi: 10.24176/simet.v9i1.1627.

Dimas Bayu Febriyanto, Lovi Handoko, Wahyuli, Hanif Aisyah, R. (2021) 'Klasifikasi Penyebaran Covid-19 Menggunakan Pendahuluan Studi Literatur', 4, pp. 23–35.

K-means, M. M., Solichin, A. and Khairunnisa, K. (2020) 'Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta', 5(2).

Novandya, A. (2017) 'Penerapan Algoritma Klasifikasi Data Mining C4.5 pada Dataset Cuaca Wilayah Bekasi', *KNiST*, 6, pp. 368–372.

Priyatman, H., Sajid, F. and Haldivany, D. (2019) 'Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan', 5(1), pp. 62–66.

Purwanto, A., Primajaya, A. and Voutama, A. (2020) 'Penerapan Algoritma C4.5 Dalam Prediksi

Potensi Tingkat Kasus Pneumonia Di Kabupaten Karawang', *Jurnal Sistem dan Teknologi Informasi (Justin)*, 8(4), p. 390. doi: 10.26418/justin.v8i4.41959.

Putra, R. S. (2021) 'Klasifikasi Penyebaran Covid-19 Menggunakan Algoritma C4.5 Kota Pagar Alam', *Jukomika*, 4(1), pp. 23–35.

Putu, N. *et al.* (2015) 'ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS', pp. 978–979.

Silitonga Irene Sri, P. M. (2017) 'Klusterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Dengan Menggunakan K-Means Clustering', *Jurnal TIMES*, VI(Vol 6, No 2 (2017)), pp. 22–25. Available at: http://ejournal.stmik-time.ac.id/index.php/jurnalTIMES/article/view/584.

Sinaga, S. M., Hardinata, J. T. and Fauzan, M. (2021) 'Implementasi Data Mining Clustering Tingkat Kepuasan', 2(2), pp. 118–124.

Turnip, S. and Siltionga, P. (2018) 'Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5', *Jurnal Teknik Informatika Unika St. Thomas*, 03(479), pp. 3–7.

Yuli Mardi (2019) 'Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika', *Jurnal Edik Informatika*, 2(2), pp. 213–219.