Visualisasi Data dan Penerapan Machine Learning Menggunakan Decision Tree Untuk Keputusan Layanan Kesehatan COVID-19

Amin Fahri[1], Yudi Ramdhani[2]

[1] Teknologi Informasi, Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya, Bandung, Indonesia
Email: [1]aminfahri512@gmail.com, [2]yudi@ars.ac.id

Abstrak—Pada Desember 2019, virus corona baru yang sekarang dinamai SARS-CoV-2, menyebabkan serangkaian penyakit pernapasan atipikal akut di Wuhan, Provinsi Hubei, China. Penyakit yang disebabkan oleh virus ini disebut COVID-19. Virus ini dapat menular antar manusia dan telah menyebabkan pandemi di seluruh dunia. Virus yang mendasari penyakit COVID-19, SARS-CoV-2, telah menyebabkan lebih dari 120 juta kasus yang dikonfirmasi dan 1,5 juta kematian sejak April 2022. Penelitian ini menggunakan algoritma Decision Tree untuk memprediksi COVID-19 dengan validasi parameter Cross Validation, Split Validation. Cross Validation pada algoritma Decision Tree Regressor memiliki tingkat performa terbaik diantara 3 algoritma lainnya, seperti; Linear Regression, Support Vector Machine Regression dan Neural Network Regression. Algoritma Decision Tree menghasilkan nilai average 57 untuk RMSE (Root Mean Square Error). Validasi data menggunakan split validation menghasilkan nilai average 29 untuk MAE (Mean Absolute Error) , 3816 untuk MSE (Mean Square Error), 59 untuk RMSE (Root Mean Square Error) dan 0,956 untuk R2 Square. Split ratio 0,9 memiliki nilai MAE, MSE, RMSE dan R2 Square tertinggi. Artinya algoritma Decision Tree Regressor memiliki kinerja yang baik untuk meningkatkan kinerja algoritma prediksi. Berdasarkan hasil penelitian mendapatkan nilai RMSE terbaik sehingga bisa digunakan oleh tenaga medis dan peneliti dalam melakukan prediksi COVID-19 dan dapat menjadi bahan rujukan metode yang akan diimplementasikan pada saat membuat penelitian mengenai COVID-19.

Kata Kunci: COVID-19, Coronavirus, SARS-CoV-2, Decision Tree Regressor, Cross Validation, Split Validation.

Abstract—In December 2019, a new coronavirus now named SARS-CoV-2, caused a series of acute atypical respiratory illnesses in Wuhan, Hubei Province, China. The disease caused by this virus is called COVID-19. This virus can be transmitted between humans and has caused a worldwide pandemic. The virus that underlies the COVID-19 disease, SARS-CoV-2, has caused more than 120 million confirmed cases and 1.5 million deaths since April 2022. This study uses a Decision Tree algorithm to predict COVID-19 with parameter validation Cross Validation, Split Validation. Cross Validation on the Decision Tree Regressor algorithm has the best performance level among 3 other algorithms, such as; Linear Regression, Support Vector Machine Regression and Neural Network Regression. The Decision Tree algorithm produces an average value of 57 for RMSE (Root Mean Square Error). Data validation using split validation resulted in an average value of 29 for MAE (Mean Absolute Error), 3816 for MSE (Mean Square Error), 59 for RMSE (Root Mean Square Error) and 0.956 for R2 Square. The split ratio of 0.9 has the highest MAE, MSE, RMSE and R2 Square values. This means that the Decision Tree Regressor algorithm has a good performance to improve the performance of the prediction algorithm. Based on the results of the study, the best RMSE value can be obtained so that it can be used by medical personnel and researchers in predicting COVID-19 and can be used as a reference material for methods that will be implemented when conducting research on COVID-19.

Keywords: COVID-19, Coronavirus, SARS-CoV-2, Decision Tree Regressor, Cross Validation, Split Validation.

1. PENDAHULUAN

Coronavirus adalah jenis virus. Ada banyak jenis yang berbeda, dan beberapa menyebabkan penyakit. Virus corona yang diidentifikasi pada 2019, SARS-CoV-2, telah menyebabkan pandemi penyakit pernapasan, yang disebut COVID-19. Kasus pertama COVID-19 dilaporkan di kota Wuhan, Provinsi Hubei, China pada 1 Desember 2019, dan penyebabnya adalah virus corona baru yang kemudian bernama SARS-CoV-2. SARS-CoV-2 mungkin berasal dari hewan dan berubah (bermutasi) sehingga dapat menyebabkan penyakit pada manusia [1]. Seperempat pasien COVID-19 yang dirawat melaporkan adanya gangguan pada indra penciuman mereka [2]. Wabah yang sedang berlangsung telah menyebabkan gejolak social ekonomi global yang serius. Hingga 22 April 2022, tercatat lebih dari 505 juta kasus COVID-19 di total 237 negara dan wilayah yang mengakibatkan lebih dari 6,21 juta kematian. Sebagai tanggapan, banyak negara telah menerapkan langkah-langkah seperti isolasi diri dan jarak social untuk mencegah penyebaran lebih lanjut, akibatnya meratakan kurva epidemi yang terbukti penting dalam mempertahankan layanan kesehatan kepada pasien yang paling membutuhkan perawatan baik untuk COVID-19 atau untuk kondisi serius lainnya [3]. Gejala paling umum dari pasien COVID-19 meliputi sesak napas, malaise, batuk kering dan demam. Sebagian kecil pasien memiliki gejala pencernaan, seperti mual, muntah dan diare [4]. Di sisi lain, 80% pasien sembuh tanpa pengobatan. Studi yang menyebutkan pria yang lebih tua, anak-anak, dan pria yang sudah menderita penyakit kardiovaskular, obesitas, dan diabetes rentan terhadap COVID-19. Kasus fatalitas COVID-19 terbukti cukup tinggi. Lihat Gambar 1 untuk 10 negara dengan angka kematian tertinggi dan Gambar 2 untuk negara dengan jumlah kasus terinfeksi tertinggi [5]. World Health

Organization (WHO) menyatakan pada tanggal 30 Januari 2020 wabah sebagai keadaan darurat dan pandemi untuk kesehatan masyarakat [5]. Kasus COVID-19 pertama di Indonesia menimpa dua warga Depok, Jawa Barat pada 2 Maret 2020. Coronavirus menyeberluas di Indonesia dengan sangat cepat. Hal ini membuat pemerintah membuat kebijakan baru Pembatasan Sosial Bersekala Besar (PSBB) untuk mencegah penyebaran coronavirus. Dampak dari berlangsungnya kebijakan PSBB ini masyarakat menanggapi bahwa kebijakannya dapat menyebabkan sejumlah industri dan mata pencaharian menjadi terganggu [6]. Pandemic COVID-19 menjadi tantangan besar bagi sistem Pendidikan. perspektif ini memberikan panduan kepada pendidik, pemimpin laboratorium, dan staf tentang cara mengelola krisis. Bagaimana lembaga perlu mempersiapkan diri dalam waktu singkat dan bagaimana mereka memenuhi kebutuhan siswa di setiap tingkat dan mata pelajaran? Meyakinkan siswa dan orang tua merupakan elemen penting dari respon institusional. Ketika sekolah dan universitas memperluas kemampuan pembelajaran jarak jauh mereka, mereka perlu memanfaatkan pembelajaran asinkron, yang bekerja paling baik dalam bentuk digital. Selain mata pelajaran regular, kelas harus mencakup berbagai tugas dan tugas yang menempatkan COVID-19 dalam konteks global dan historis. Saat membawa silabus, desain penilaian siswa membantu guru fokus. Terakhir, perspektif ini mengusulkan cara yang fleksibel untuk memperbaiki kerusakan jalur belajar siswa setelah pandemi berakhir dan menyediakan daftar sumber daya [7].

World Health Organization (WHO) menyatakan pada tanggal 30 Januari 2020 wabah sebagai keadaan darurat dan pandemi untuk kesehatan masyarakat [5]. Coronavirus menyeberluas di Indonesia dengan sangat cepat. Hal ini membuat pemerintah membuat kebijakan baru Pembatasan Sosial Bersekala Besar (PSBB) untuk mencegah penyebaran coronavirus. Dampak dari berlangsungnya kebijakan PSBB ini masyarakat menanggapi bahwa kebijakannya dapat menyebabkan sejumlah industri dan mata pencaharian menjadi terganggu [6]. Pada bulan Desember 2019, wabah pneumonia yang tidak dapat dijelaskan telah dilaporkan di Wuhan, Provinsi Hubei, Cina. Kasus pneumonia secara epidemiologis terkait dengan pasar grosis makanan laut Cina Selatan. Ketika sampel pernapasan diinokulasi ke dalam sel epitel pernapasan manusia, garis sel Vero E6 dan Huh7, virus pernapasan baru diisolasi dan analisis genomiknya mengungkapkan bahwa itu adalah virus corona baru yang terkait dengan SARSCoV dan oleh karena itu merupakan penyakit pernapasan akut yang parah. Diklasifikasikan sebagai Coronavirus 2. Sindrom (SARS-CoV-2). SARSCoV2 adalah beta coronavirus milik subgenus Salvecovirus dengan penyebaran SARSCoV2 di seluruh dunia dan kematian ribuan orang karena penyakit coronavirus (COVID-19), World Health Organization (WHO) menyatakan pandemic pada 12 Maret 2020. Hingga saat ini, pandemic ini telah melanda dunia dari segi kehidupan, dampak ekonomi dan peningkatan kemiskinan. Ikhtisar ini memberikan informasi tentang diagnostic epidemiologi, serologis dan molekuler, asal mula SARSCoV2 dan kemampuannya untuk menginfeksi sel manusia, dan masalah keamanan. Selanjutnya, terapi yang tersedia untuk memerangi COVID-19, pengembangan vaksin, peran kecerdasan buatan dalam manajemen pandemic dan membatasi penyebaran virus, dampak epidemi COVID-19 pada gaya hidup kita, dan kemungkinan gelombang kedua. Fokus pada cara mempersiapkan diri [8].

Visualisasi data berarti menggambar tampilan grafik untuk menampilkan data. Terkadang setiap titik data di gambar, seperti dalam scatterplot, terkadang ringkasan statistik dapat ditampilkan, seperti dalam histogram. Tampilan utamanya deskriptif, berkonsentrasi pada data ‘mentah’ dan ringkasan sederhana. Seperti aspek lain dari bekerja dengan grafik, akan berguna untuk memiliki dasar konsep dan terminology yang disepakati untuk dibangun. Tujuan utamanya adalah untuk memvisualisasikan data dan statistik, menafsirkan tampilan untuk mendapatkan informasi [9]. Dampak mematikan dari COVID-19 mendorong sejumlah besar penelitian yang bertujuan untuk memahami berbagai karakteristik pandemi. Kecepatan penyebaran penyakit ke seluruh dunia menuntut solusi tangkas untuk memahami dan memperkirakan perkembangan penyakit [10]

Amerika Serikat telah memasuki fase sejarah baru dengan penyebaran cepat virus corona baru SARS-CoV-2 dan kematian akibat COVID-19. Dalam satu survei dengan 1210 peserta yang dilakukan pada Januari dan Februari 2020, 54% menilai dampak psikologis pandemi COVID-19 sebagai sedang hingga berat, 29% melaporkan gejala kecemasan sedang hingga berat, 17% melaporkan depresi sedang hingga berat. Studi ini juga menemukan perbedaan regional dalam tekanan psikologis, dengan responden dari provinsi Hubei, pusat pandemi COVID-19, melaporkan tekanan yang jauh lebih tinggi. Selain itu, orang dengan gangguan mental yang sudah ada sebelumnya dapat lebih terpengaruh oleh pandemi COVID-19, termasuk kemungkinan kekambuhan atau eksaserbasi kondisi kejiwaan. Penelitian dari China telah menunjukkan bahwa mahasiswa yang keluarganya memiliki pendapatan kurang stabil berisiko lebih tinggi mengalami tekanan mental karena COVID-19 [11]. Kami menggunakan data untuk 690.825 orang dewasa di Negara Bagian New York untuk menilai efektivitas vaksin BNT162b2, mRNA-1273 dan Ad26.COV2.S terhadap COVID-19 (yaitu, COVID-19 didiagnosis pada atau setelah masuk). Kami menilai efektivitas vaksin terhadap COVID-19 dari 1 Mei hingga 3 September 2021, dan terhadap rawat inap dengan COVID-19 dari 1 Mei hingga 31 Agustus 2021. Efektivitas terhadap rawat inap dengan COVID-19 diantara 7 orang dewasa berusia 18 hingga 64 tahun tetap hampir secara eksklusif lebih besar dari 86% tanpa tren waktu yang jelas. Efektivitas terhadap rawat inap tetap tinggi, dengan penurunan sederhana terbatas pada penerima BNT126b2 dan mRNA-1273 berusia 65 tahun atau lebih [12].

United States of America merupakan negara dengan kasus terinfeksi paling tinggi yaitu mencapai 80.006.661 kasus dan Turkey menjadi negara dengan kasus terinfeksi urutan ke-10 paling tinggi yaitu mencapai 15.010.718 kasus. Rata-rata kasus terinfeksi dari 10 negara tertinggi yaitu mencapai 29.237.896. berikut table 1 yang memperlihatkan 10 negara dengan kasus terinfeksi tertinggi.

Tabel 1. 10 Negara dengan Kasus Terinfeksi Tertinggi

No.	Country	Cumulative Cases
1.	United States of America	80.006.661
2.	India	43.052.425
3.	Brazil	30.311.969
4.	France	27.272.068
5.	Germany	24.006.254
6.	The United Kingdom	21.909.513
7.	Russian Federation	18.119.862
8.	Republic of Korea	16.755.055
9.	Italy	15.934.437
10.	Turkey	15.010.718

Data mining adalah penemuan struktur yang menarik, tidak terduga atau berharga dalam kumpulan data yang besar. Dengan demikian, ia memiliki dua aspek yang berbeda. Salah satunya menyangkut skala besar, struktur 'global', dan tujuannya adalah untuk memodelkan bentuk atau fitur dari bentuk, distribusi. Yang lainnya menyangkut struktur 'lokal' skala kecil, dan tujuannya adalah untuk mendeteksi anomaly ini dan memutuskan apakah itu kejadian nyata atau kebetulan [13]. Data mining dapat digunakan untuk klasifikasi atau prediksi. Tujuan dari penelitian ini adalah untuk mendapatkan metode prediksi terbaik yang dapat mencapai tingkat akurasi yang tinggi dalam kombinasinya dengan algoritma Decision Tree [14]. Data mining adalah teknologi yang baru dikembangkan, yang memiliki metode, prosedur dan tekniknya sendiri. Ini adalah tugas untuk menemukan pola yang berguna dari sejumlah data besar. Dalam perawatan kesehatan, data mining menjadi semakin popular dan menjadi kebutuhan. Penerapannya dalam perawatan kesehatan dapat memiliki potensi dan kegunaan yang luar biasa. Ini memainkan peran penting untuk mengungkap tren baru dalam industri perawatan kesehatan. Meskipun data mining kesehatan masih dalam masa pertumbuhan, literatur data mining perawatan kesehatan sangat kaya. Masa depan perawatan kesehatan akan bergantung pada penggunaan data mining untuk mengurangi biaya perawatan kesehatan dan meningkatkan standar perawatan pasien [15].

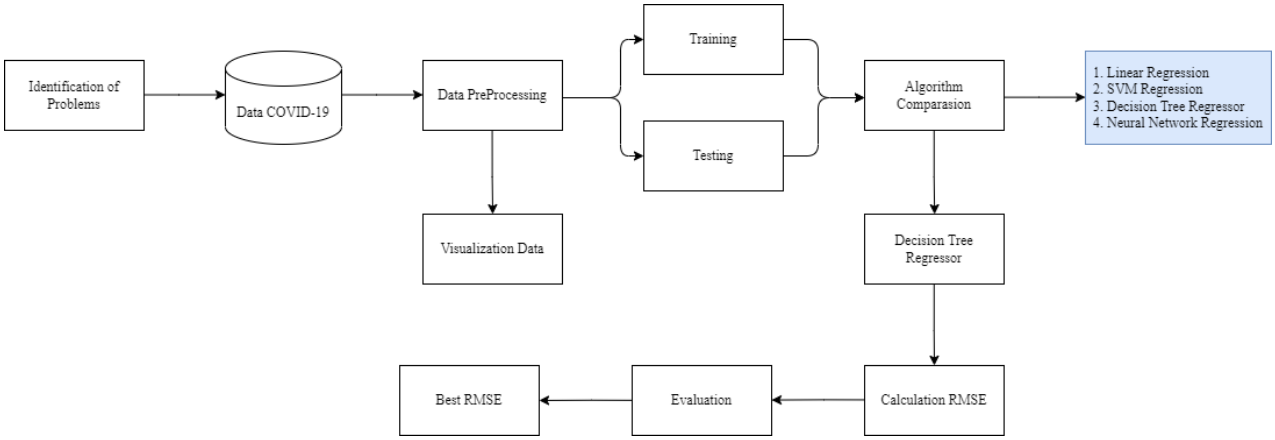
Dalam *Machine Learning*, pendekatan konvensional adalah memproses data dalam batch. Model pembelajaran batch mengasumsikan bahwa semua data tersedia sekaligus. Ketika kumpulan data baru tersedia, model ini harus dilatih ulang dari awal. Asumsi ketersediaan data merupakan kendala keras untuk penerapan *Machine Learning* di beberapa aplikasi dunia nyata di mana data dihasilkan secara terus menerus [16]. Banyak metode *Machine Learning* tersedia untuk mengembangkan model prediksi yang bermanfaat. Namun, kebanyakan dari mereka sulit untuk diinterpretasikan karena mekanisme model internal mereka [4]. Decision tree adalah metode penambangan data yang umum digunakan untuk membangun system klasifikasi berdasarkan beberapa kovariat atau untuk mengembangkan algoritma prediksi untuk variabel target. Metode ini mengklasifikasikan populasi ke dalam segmen seperti cabang yang membangun pohon terbalik dengan simpul akar, simpul internal dan simpul daun. Algoritma non-parametrik dan dapat secara efisien menangani kumpulan data yang besar dan rumit tanpa memaksakan struktur parametrik yang rumit. Karena ukuran sampel cukup besar, data studi dapat dibagi menjadi dataset pelatihan dan validasi. Menggunakan dataset pelatihan untuk memutuskan ukuran pohon yang sesuai yang diperlukan untuk mencapai model akhir yang optimal [17]. Python adalah salah satu bahasa pemograman paling popular untuk *data science* dan oleh karena itu menikmati sejumlah besar *libraries* tambahan yang berguna yang dikembangkan oleh komunitasnya yang hebat. Meskipun kinerja Bahasa yang ditafsirkan, seperti python, untuk tugas komputas intensif lebih rendah daripada bahasa pemograman tingkat rendah, *extension libraries* seperti *NumPy* dan *SciPy* telah dikembangkan yang dibangun di atas implementasi *Fortran* dan *C* lapisan bawah untuk opeasi cepat dan vector pada array multidimensi. Untuk tugas pemograman *Machine Learning*, kami

Sebagian besar akan merujuk ke *libraries sckit-learn*, yang merupakan salah satu *open source machine learning libraries* yang paling popular dan dapat diakses hingga hari ini [20].

Tujuan dari penelitian ini adalah membuat visualisasi data menjadi komponen yang menarik dan berharga. Mereka dapat memberikan struktur dan wawasan, memungkinkan individu untuk mengelola kesehatan mereka sendiri secara efektif[18] dan memilih Decision Tree Regressor sebagai model prediksi karena terlihat dapat dioperasikan secara klinis dan mudah diinterpretasikan karena system Decision Tree yang rekursif [4].

2. METODE PENELITIAN

Dalam penelitian ini menggunakan metode Decision Tree Regressor untuk prediksi. Pada penelitian ini dilakukan beberapa langkah atau tahapan penelitian, ditunjukkan pada Figure 1.



Gambar 1. Metode Penelitian

2.1 Identifikasi Masalah

Penelitian ini dilakukan untuk memvisualisasikan dan menganalisis kasus COVID-19 di Indonesia agar memudahkan para ahli mengetahui perkembangan kasus COVID-19 dengan mudah.

2.2 Pengumpulan Data

Penelitian ini mengambil data dari situs resmi World Health Organization (WHO) (https://covid19.who.int/data/). Dataset tersebut memiliki 199.317 record data dan 8 variabel yaitu, *Date_reported*, *Country_code*, *Country*, *WHO_region*, *New_cases*, *Cumulative_cases*, *New_deaths* dan *Cumulative_deaths*. Dataset COVID-19 dari tanggal 03 Januari 2020 sampai dengan 22 April 2022 dari 236 Negara ditunjukkan pada Table 2.

Tabel 2. Dataset COVID-19 Dunia

NO	Date	Country_code	Country	WHO region	New cases	Cases	New deaths	Deaths
0.	2020-01-03	AF	Afghanistan	EMRO	0	0	0	0
1.	2020-01-04	AF	Afghanistan	EMRO	0	0	0	0
2.	2020-01-05	AF	Afghanistan	EMRO	0	0	0	0
3.	2020-01-06	AF	Afghanistan	EMRO	0	0	0	0
4.	2020-01-07	AF	Afghanistan	EMRO	0	0	0	0
…	…	…	…	…	…	…	…	

199312	2022-04-18	ZW	Zimbabwe	AFRO	17	247.278	1	5.463
199313	2022-04-19	ZW	Zimbabwe	AFRO	19	247.297	1	5.464
199314	2022-04-20	ZW	Zimbabwe	AFRO	39	247.336	2	5.466
199315	2022-04-21	ZW	Zimbabwe	AFRO	47	247.383	1	5.467
199316	2022-04-22	ZW	Zimbabwe	AFRO	0	247.383	0	5.467

2.3 *Preprocessing Data*

Langkah selanjutnya setelah mengumpulkan data adalah melakukan data preprocessing. Pada tahap ini, dataset diperiksa untuk type data, dan describe data. Type data adalah suatu media pada komputer yang digunakan untuk menampung informasi. Describe data adalah digunakan untuk menghitung beberapa data statistik seperti persentil, mean dan std dari nilai numerik dataset.

2.4 Visualisasi Data

Visualisasi data adalah tentang mengembangkan wawasan dari data yang dikumpulkan menjadi sebuah diagram yang objeknya unik dan karenanya menimbulkan ketertarikan dan skeptisme.

2.5 Validasi Data

Dataset akan dibagi menjadi dua bagian, yaitu data training dan data testing. Pembagian data tersebut dilakukan menggunakan Split Validation dan Cross Validation. Pembagian data menggunakan Split Validation dilakukan untuk menguji suatu algoritma berdasarkan rasio pemisahan sedangkan Cross Validation digunakan untuk mengetahui algoritma yang memiliki performa terbaik.

2.6 Komparasi Algoritma

Menguji komparasi dataset menggunakan 4 algoritma, yaitu *Decision Tree Regressor*, *Linear Regression*, *Support Vector Machine Regression* dan *Neural Network Regression*.

2.7 Model *Decision Tree Regressor*

Pada tahap ini, model *Decision Tree Regressor* mendapatkan hasil RMSE terbaik dari tiga algoritma yang telah diuji. Model terbaik yang telah didapatkan akan diuji menggunakan teknik split validation dengan split ratio mulai dari 0,5 sampai 0,9.

2.8 Evaluasi

Pada tahap evaluasi hasil dari pengujian split validation yang telah dilakukan. Pengujian split validation tersebut menghasilkan nilai MAE, MSE, RMSE dan R2 Square. Nilai dari split ratio 0,5 sampai 0,9 akan dibandingkan sehingga didapatkan split ratio terbaik dalam prediksi COVID-19.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan untuk penelitian hanya negara Indonesia saja. Berikut table 3 menampilkan dataset COVID-19 di Indonesia dari 3 Januari 2020 sampai dengan 22 April 2022.

Tabel 3. Dataset COVID-19 Dunia

	Date_ reported	**Country_ code**	**Country**	**WHO_ region**	**New_ cases**	**Cumulative_ cases**	**New_ deaths**	**Cumulative_ deaths**
1.	2020-01-03	ID	Indonesia	SEARO	0	0	0	0
2.	2020-01-04	ID	Indonesia	SEARO	0	0	0	0
3.	2020-01-05	ID	Indonesia	SEARO	0	0	0	0
4.	2020-01-06	ID	Indonesia	SEARO	0	0	0	0

5	2020-01-07	ID	Indonesia	SEARO	0	0	0	0
…	…	…	…	…	…			
810.	2022-04-18	ID	Indonesia	SEARO	559	6.040.432	37	155.903
811.	2022-04-19	ID	Indonesia	SEARO	837	6.041.269	34	155.937
812.	2022-04-20	ID	Indonesia	SEARO	741	6.042.010	37	155.974
813.	2022-04-21	ID	Indonesia	SEARO	585	6.042.595	41	156.015
814.	2022-04-22	ID	Indonesia	SEARO	0	6.042.595	0	156.015

Setelah pengumpulan data, peneliti melakukan preprocessing data. Pada tahap ini dilakukan pengecekan data type terhadap data untuk melihat type data apa saja yang berada di dataset tersebut. Selanjutnya melakukan describe data untuk menghitung beberapa data statistic seperti persentil, mean dan std dari nilai numerik dari dataset. Setelah itu melakukan missing value untuk melihat apakah terdapat data yang tidak sesuai. Berikut ini adalah hasil dari pengecekan data type, describe type dan missing values yang telah dilakukan.

Terdapat 3 data type yaitu datatime64 untuk variabel (Date_reported), object untuk variabel (Country_code, Country dan WHO_region) dan int64 untuk variabel (New_cases, Cumulative_cases, New_deaths, Cumulative_deaths). Seperti yang ditampilkan pada gambar 1.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 841 entries, 81577 to 82417
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Date_reported      841 non-null    datetime64[ns]
 1   Country_code       841 non-null    object
 2   Country            841 non-null    object
 3   WHO_region         841 non-null    object
 4   New_cases          841 non-null    int64
 5   Cumulative_cases   841 non-null    int64
 6   New_deaths         841 non-null    int64
 7   Cumulative_deaths  841 non-null    int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 59.1+ KB
```

Gambar 1. Data Type Dataset

Metode describe data digunakan untuk menghitung beberapa data statistik seperti jumlah, mean, std (standard deviation) dan persentil pada dataset. Berikut table 4 menampilkan describe data pada dataset.

Tabel 4. Describe Data Dataset

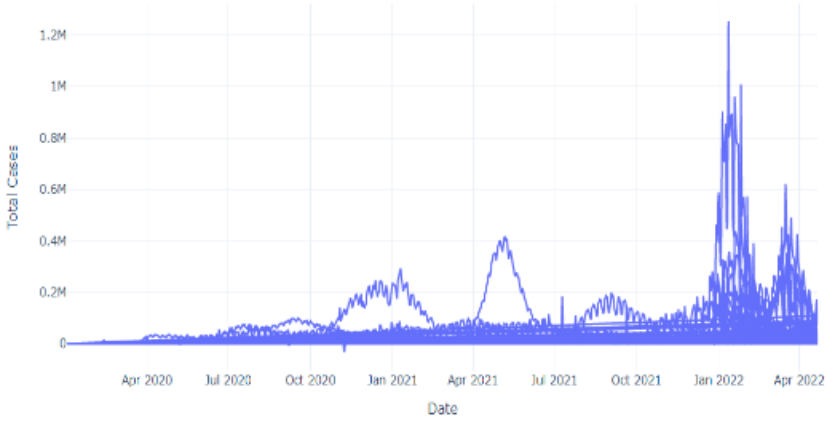
	New_cases	**Cumulative_Cases**	**New_deaths**	**Cumulative_deaths**
count	841.00	8.4100	841.00	841.00
mean	7185.01	1.943063	185.51	58827.30
std	11540.98	1.974537	325.19	59770.07
min	-4.00	0.00	0.00	0.00
25%	533.00	1.083760	21.00	5131.00
50%	3448.00	1.322866	91.00	35786.00
75%	6825.00	4.204116	184.00	141258.00
max	64718.00	6.042595	2069.00	156015.00

a. Visualisasi Data

Teknik visualisasi telah menjadi yang terdepan dalam upaya mengkomunikasikan sains seputar COVID-19 kepada khalayak yang sangat luas. Dalam penelitian ini, saya merangkum dan mengilustrasikan dengan contoh bagaimana visualisasi dapat membantu memahami berbagai aspek pandemi. Banyak contoh lain visualisasi data dalam analisis data COVID-19 dan banyak lagi yang muncul setiap hari. Kami berharap ringkasan ini menyoroti contoh-contoh menarik, memberikan petunjuk ke referensi lain dan memotivasi orang untuk mengejar aplikasi lain.

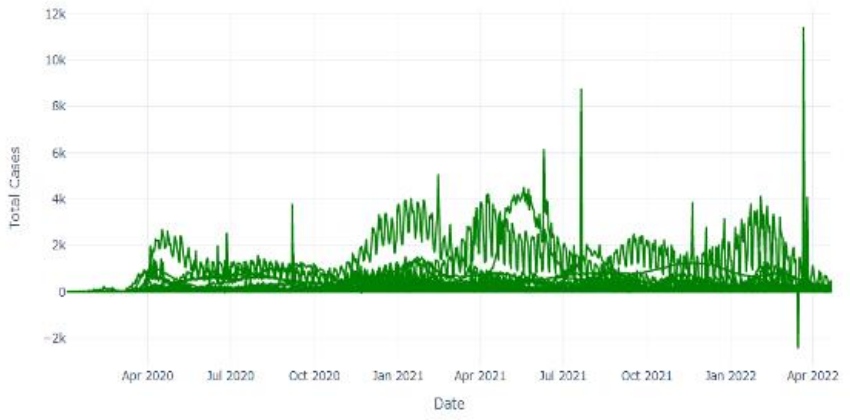
1) COVID-19 di dunia

Jumlah kasus terinfeksi di Dunia paling tinggi yaitu mencapai 1.252.717 kasus pada 12 Januari 2022. Grafiknya bisa dilihat pada Gambar 3.



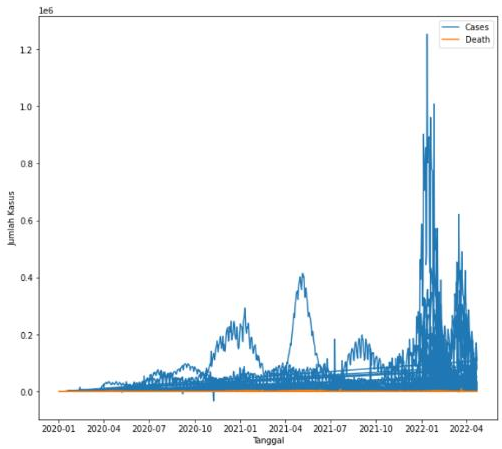
Gambar 3. Grafik Kasus Terinfeksi COVID-19 di Dunia

Jumlah kasus kematian di Dunia paling tinggi yaitu mencapai 11.447 kasus pada 22 Maret 2022. Grafiknya bisa dilihat pada Gambar 4.



Gambar 4. Grafik Kasus Kematian COVID-19 di Dunia

Pada gambar 5 menampilkan perbandingan kasus terinfeksi dan kasus kematian akibat COVID-19 di Dunia.



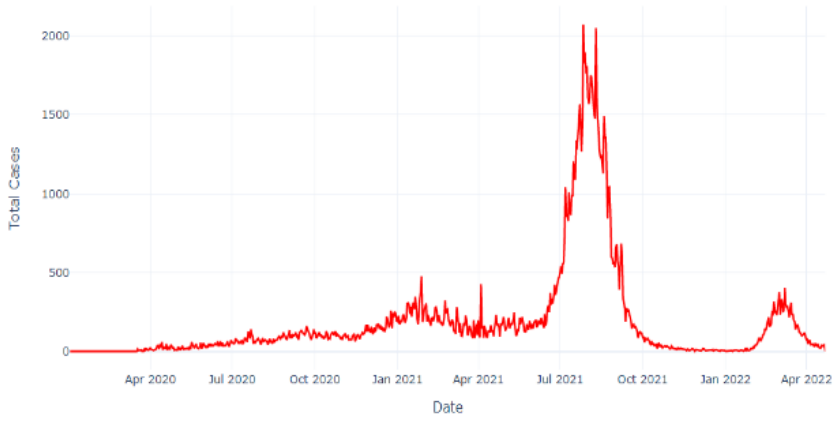
Gambar 5. Grafik Perbandingan Kasus Terinfeksi dan Kasus Kematian COVID-19 di Dunia

2) COVID-19 di Indonesia

Jumlah kasus terinfeksi di Indonesia paling tinggi yaitu mencapai 64.781 kasus pada 16 Februari 2022. Grafiknya bisa dilihat pada Gambar 6.

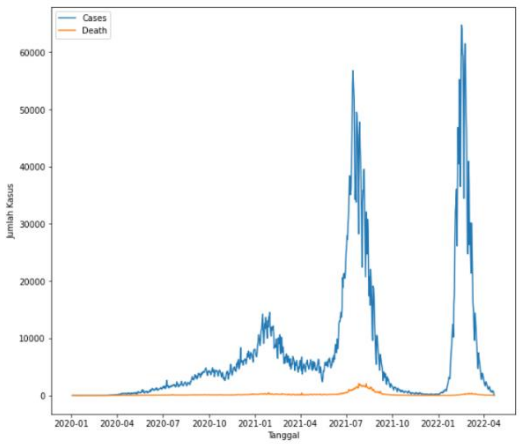
Gambar 6. Grafik Kasus Terinfeksi COVID-19 di Indonesia

Jumlah kasus kematian di Indonesia paling tinggi yaitu mencapai 2069 kasus pada 27 Juli 2021. Grafiknya bisa dilihat pada Gambar 7.



Gambar 7. Grafik Kasus Kematian COVID-19 di Indonesia

Pada gambar 8 terdapat perbandingan kasus terinfeksi dan kematian akibat COVID-19 di Indonesia.



Gambar 8. Grafik Perbandingan Kasus Terinfeksi dan Kasus Kematian COVID-19 di Indonesia

B. Hasil Penelitian

Peneliti melakukan komparasi algoritma untuk mencari algoritma mana yang terbaik. Berdasarkan table 5, diketahui bahwa algoritma Decision Tree Regression mendapatkan nilai RMSE tertinggi dari algoritma lainnya.

Tabel 5. Komparasi Algoritma

Algoritma	Validasi	RMSE
Linear Regression	Cross Validation	248
SVM Regression	Cross Validation	329
Decision Tree Regressor	Cross Validation	57
Neural Network Regression	Cross Validation	340

Setelah melakukan komparasi algoritma dapat diketahui bahwa Decision Tree Regressor merupakan algoritma dengan performa terbaik dalam prediksi terhadap kasus COVID-19 di Indonesia, selanjutnya melakukan pengujian validasi data untuk menguji performa algoritma Decision Tree Regressor menggunakan Split Validation dengan split ratio 0,5 sampai 0,9 yang ditunjukkan pada table 6.

Tabel 6. Pengujian Performa Decision Tree Regressor

Algoritma	Validasi	Ratio	MAE	MSE	RMSE	R2 Square
Decision	Split	0,5	31	3766	61	0,96
Tree	Validation	0,6	30	3503	59	0,96
Regressor		0,7	36	7450	86	0,93
		0,8	27	3236	56	0,95
		0,9	21	1125	33	0,98
Average			29	3816	59	0,956

Berdasarkan Table 6 diketahui bahwa algoritma Decision Tree Regression dengan split ratio 0,5 hingga 0,9 memiliki nilai average 29 untuk MAE (*Mean Absoulte Error*), 3816 untuk MSE (*Mean Squared Error*), 59 untuk RMSE (*Root Mean Squared Error*) dan 0,956 untuk R2 Square. Validasi dengan split ratio 0,9 memiliki hasil akhir yang tinggi dibandingkan split ratio lainnya.

Berdasarkan pengujian yang telah diperoleh pada dataset COVID-19, dapat diketahui bahwa algoritma Decision Tree Regressor memiliki rmse yang baik yaitu 57, sehingga dapat digunakan oleh para ahli di bidang kesehatan dalam memprediksi COVID-19 dan bagi seorang peneliti dapat menjadi bahan rujukan metode yang akan diimplementasikan pada saat membuat penelitian mengenai prediksi COVID-19.

4. KESIMPULAN

Berdasarkan pengujian yang telah diperoleh pada dataset COVID-19, dapat diketahui bahwa algoritma Decision Tree Regressor memiliki rmse yang baik yaitu 57, sehingga dapat digunakan oleh para ahli di bidang kesehatan dalam memprediksi COVID-19 dan bagi seorang peneliti dapat menjadi bahan rujukan metode yang akan diimplementasikan pada saat membuat penelitian mengenai prediksi COVID-19.

Penelitian ini melakukan visualisasi data dan pemodelan menggunakan algoritma Decision Tree Regressor dengan menggunakan dataset COVID-19 yang didapatkan dari website resmi WHO. Validasi data dengan Cross Validation pada Algoritma Decision Tree Regressor memiliki tingkat performa terbaik diantara 3 algoritma lainnya seperti; Linear Regression, Neural Network Regression dan Support Vector Machine Regression. Algoritma Decision Tree Regressor menghasilkan nilai RMSE 57. Untuk menguji algoritma Decision Tree Regressor menggunakan validasi data split validation dengan split ratio 0,5 hingga 0,9. Berdasarkan pengujian yang telah dilakukan, nilai average yang dihasilkan adalah 29 untuk MAE, 3816 untuk MSE, 59 untuk RMSE dan 0,956 untuk R2.

Berdasarkan hasil dari penelitian yang telah dilakukan, maka peneliti mengajukan saran untuk melakukan penelitian dengan menggunakan metode algoritma lainnya seperti Stochastic Gradient Descent.

REFERENCES

[1] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual Quant*, 2021, doi: 10.1007/s11135-021-01176-w.

[2] D. Hornuss, B. Lange, N. Schroeter, S. Rieg, W. v Kern, and D. Wagner, "Anosmia in COVID-19 patients," *Clinical Microbiology and Infection*, vol. 26, no. 10, pp. 1426–1427, 2020.

[3] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "COVID-19: a comparison of time series methods to forecast percentage of active cases per population," *Applied sciences*, vol. 10, no. 11, p. 3880, 2020.

[4] Q. Yang *et al.*, "Clinical characteristics and a decision tree model to predict death outcome in severe COVID-19 patients," *BMC Infect Dis*, vol. 21, no. 1, pp. 1–9, 2021.

[5] H. Sastypratiwi, Y. Yulianti, and H. Muhardi, "Uji Komparasi Algoritma Naïve Bayes dan Decision Tree Classification Menggunakan Covid-19 Dataset," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 8, no. 1, pp. 1–6.

[6] A. Mahmudan, "Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering," *Jurnal Matematika, Statistika dan Komputasi*, no. Vol. 17 No. 1 (2020): JMSK, SEPTEMBER, 2020, pp. 1–13, 2020.

[7] S. J. Daniel, "Education and the COVID-19 pandemic," *Prospects (Paris)*, vol. 49, no. 1, pp. 91–96, 2020.

[8] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini, "The COVID-19 pandemic," *Crit Rev Clin Lab Sci*, vol. 57, no. 6, pp. 365–388, 2020.

[9] A. Unwin, "Why is data visualization important? what is important in data visualization?," 2020.

[10] J. L. D. Comba, "Data Visualization for the Understanding of COVID-19," *Comput Sci Eng*, vol. 22, no. 6, pp. 81–86, 2020, doi: 10.1109/MCSE.2020.3019834.

[11] C. Holingue *et al.*, "Mental distress in the United States at the beginning of the COVID-19 pandemic," *Am J Public Health*, vol. 110, no. 11, pp. 1628–1634, 2020.

[12] E. S. Rosenberg *et al.*, "COVID-19 vaccine effectiveness in New York state," *New England Journal of Medicine*, vol. 386, no. 2, pp. 116–127, 2022.

[13] D. J. Hand, "Principles of data mining," *Drug Saf*, vol. 30, no. 7, pp. 621–622, 2007.

[14] T. W. Pratiwi and T. Arifin, "Optimasi Decision Tree Menggunakan Particle Swarm Optimization untuk Klasifikasi Kesuburan pada Pria," *Sistemasi: Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 13–25, 2021.

[15] M. Sadiku, K. Eze, and S. Musa, "Data Mining in Healthcare," *International Journal of Advances in Scientific Research and Engineering*, vol. 4, pp. 90–92, Jan. 2018, doi: 10.31695/IJASRE.2018.32881.

[16] J. Montiel *et al.*, "River: machine learning for streaming data in Python," 2021.

[17] Y.-Y. Song and L. U. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, p. 130, 2015.

[18] A. Polhemus *et al.*, "Data Visualization for Chronic Neurological and Mental Health Condition Self-management: Systematic Review of User Perspectives," *JMIR Ment Health*, vol. 9, no. 4, p. e25249, 2022.