Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 1, No. 2, November 2023, hlm 97-105

Analisis Prediksi Kematian Pasien Covid-19 di Meksiko Menggunakan Algoritma Random Forest

Dennis Fitri Salsabilla Arianti [a,1,*], Latifah Arum [a,2], Mohamad Burhanudin [a,3]

[a] Program Studi Sistem Informasi Universitas Jenderal Achmad Yani Yogyakarta, Jl. Siliwangi Ringroad Barat, Sleman 55293, Indonesia

[1]denissalsa59@gmail.com*; [2]latifaharums123@gmail.com; [3]mohamadburhan151@gmail.com

* corresponding author

ABSTRACT

This research aims to analyze and predict the deaths of Covid-19 patients in Mexico using the Random Forest algorithm. The data used in this study is sourced from official sources, including the number of cases, symptoms, risk factors, and Covid-19 patient mortality data. The first stage of this research is data preprocessing, where the acquired data is collected, cleaned, and prepared for analysis. Subsequently, data exploration is conducted to understand the characteristics and patterns within the dataset. Then, the Random Forest model is developed to predict the deaths of Covid-19 patients based on relevant factors. Model evaluation is performed using accuracy, precision, recall, and F1-score metrics. The results of this research indicate that the random forest model can provide good predictions for Covid-19 patient deaths in Mexico. The evaluation results show a high level of accuracy and satisfactory performance for the model. These findings can be used as guidance in decision-making and strategic planning to address the Covid-19 pandemic in Mexico. This research contributes significantly to the field of predictive analysis and provides valuable insights in the efforts to manage the Covid-19 pandemic.

ARTICLE INFO

Article history
Received: 12 September 2023
Revised: 19 November 2023
Accepted: 23 November 2023

Keywords

Covid-19; Mexico; Random Forest

1. Pendahuluan

Covid-19 atau *Corona Virus Disease* 2019 merupakan penyakit yang muncul pada tahun 2019 dan dapat menyebabkan gangguan pernapasan [1]. Penyebab penyakit ini adalah *Severe Acute Respiratory Syndrome Coronavirus* 2 (SARS-CoV-2) yang dapat menyebabkan kematian [2]. *World Health Organization* (WHO) mengumumkan korban yang meninggal akibat Covid-19 di Meksiko mencapai 333.960 orang dari jumlah kasus terkonfirmasi 7.595.574 orang [3]. Studi ini dilakukan untuk memberikan informasi hasil prediksi kematian pasien Covid-19 di Meksiko berdasarkan gejala yang dialami pasien tersebut.

Pasien dengan tingkat gejala yang lebih kompleks memerlukan prioritas penanganan dibanding dengan pasien dengan gejala ringan atau tanpa gejala [4]. Tenaga medis memerlukan bantuan untuk memprediksi secara otomatis apakah pasien dapat sembuh atau tidak berdasarkan gejala yang dialaminya. Hal ini untuk meminimalisir resiko penanganan yang terlambat terhadap pasien. Oleh karena itu dibutuhkan solusi teknologi berbasis data secara otomatis yang dapat membantu mengklasifikasikan apakah pasien dapat sembuh atau tidak berdasarkan data gejala pasien.

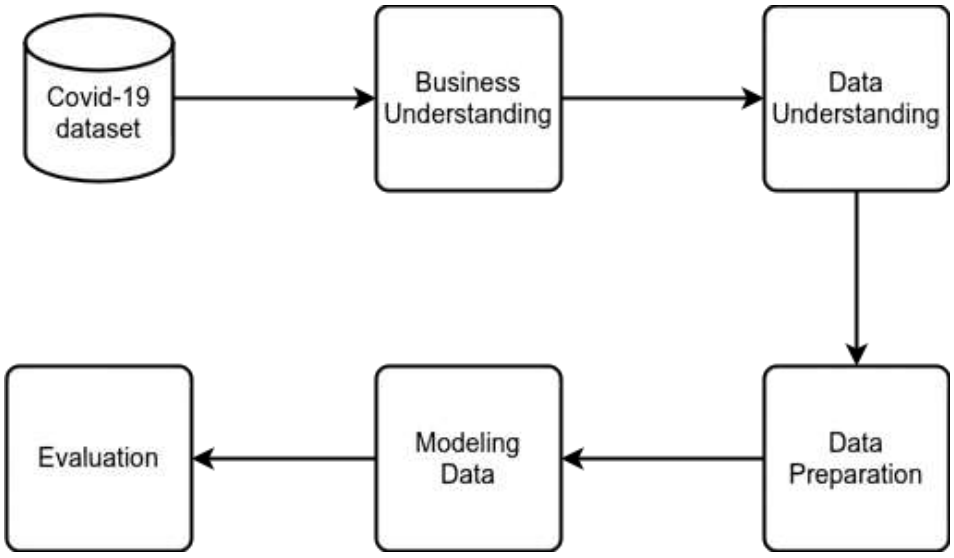
Machine Learning adalah salah satu metodologi cerdas yang telah menunjukan hasil dalam klasifikasi dan prediksi. Tugas *Machine Learning* secara umum melibatkan memprediksi variabel target dalam data yang sebelumnya atau disebut klasifikasi. Tujuan klasifikasi adalah untuk memprediksi variabel target (kelas) dengan membangun model klasifikasi. Tujuan klasifikasi adalah untuk memprediksi variabel target (kelas) dengan membangun model klasifikasi berdasarkan dataset pelatihan, dan kemudian menggunakan model untuk memprediksi nilai kelas data uji (Bunker & Thatbah).

Model proses yang digunakan dalam penelitian ini menggunakan *CRISP-DM (Cross-Industry Standard Process for Data Mining).* Model proses ini menyediakan kerangka kerja yang terstruktur untuk mengelola proyek data mining dari awal hingga akhir. Dalam mengklasifikasikan pasien yang terkena *Covid-19* mati atau tidak dengan menggunakan metode *Random Forest*. Metode ini umum digunakan dalam prediksi karena metode ini memberikan estimasi yang baik dalam prediksi. Dengan adanya penelitian ini diharapkan mampu membantu pemerintah Meksiko menentukan prioritas penanganan pasien *Covid-19*.

2. Metode

Penelitian ini menggunakan metode proses *CRISP-DM (Cross-Industry Standard Process for Data Mining)* yaitu metode atau proses standar untuk melakukan penggalian data atau data mining yang digunakan untuk membantu memahami dan menyelesaikan masalah bisnis dengan menggunakan teknik-teknik analisis data. Metode ini memiliki lima tahapan yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling,* dan *Evaluation.*

Gambar 1 merupakan proses untuk memprediksi kematian pasien Covid-19 di Meksiko pada Covid-19 dataset. Proses pertama adalah mengambil Covid-19 *dataset* dari *Kaggle*, dilanjutkan dengan business understanding untuk mengidentifikasi masalah bisnis yang spesifik dan mengklarifikasi tujuan yang ingin dicapai. *Data preparation* membersihkan, mentransformasi dan mengintegrasikan data agar siap untuk analisis lebih lanjut. Pada data *modelling*, model prediksi atau algoritma dikembangkan dan dilatih menggunakan data yang telah disiapkan, dengan tujuan menghasilkan prediksi yang akurat. Tahap *evaluation* mengukur kinerja model dan menganalisis hasilnya, supaya memungkikan penentuan model terbaik untuk mencapai tujuan bisnis yang diinginkan.



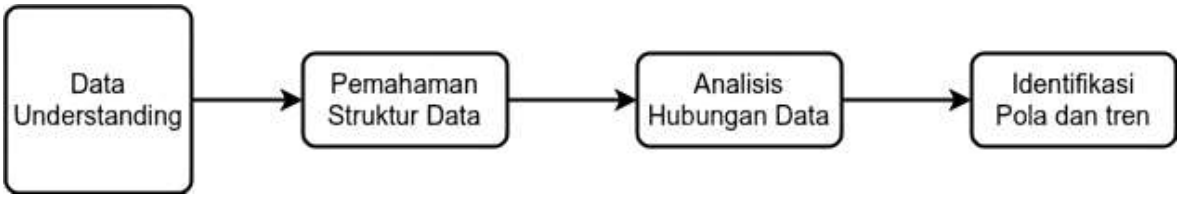
Gambar 1 Tahapan *CRISP-DM*

2.1. *Business Understanding*

Tahap *business understanding* bertujuan untuk mendapatkan pemahaman bisnis yang harus dipecahkan, tujuan penelitian yang akan dicapai, serta *stakeholder* yang terlibat. Hal ini akan membantu dalam merumuskan rencana proyek supaya sesuai dengan tujuan yang spesifik, yaitu prediksi kematian *Covid-19* di Meksiko.

2.2. *Data Understanding*

Pada tahap ini, peneliti melakukan analisis mendalam terhadap karakteristik data dan pola - pola yang mungkin terdapat didalamnya. Berikut ini langkah - langkah data understanding:



Gambar 2 Proses Data Understanding

Data yang digunakan pada penelitian ini adalah *Covid-19 Dataset* yang dapat diakses di Kaggle [6]. *Dataset* ini memiliki 1.048.576 *rows*, dan 21 *attributes* yang dijelaskan di Tabel 1.

Tabel 1 Parameter *Dataset Covid-19*

Parameters	Details	Value
USMR	Menunjukkan apakah pasien dirawat unit medis tingkat pertama, kedua atau ketiga (1 = Iya, 2 = Tidak)	int64
MEDICAL_UNIT	Jenis lembaga Sistem Kesehatan Nasional yang menyediakan perawatan (1 - 13)	int64
SEX	Jenis kelamin (1 = Perempuan, 2 = Laki - laki)	int64
PATIENT_TYPE	Jenis perawatan pasien yang diterima di unit (1 = di rumah, 2 = rawat inap)	int64
DATE_DIED	Jika pasien meninggal, maka akan menunjukkan tanggal kematian dan 9999-99-99 untuk yang tidak meninggal	object
INTUBED	Apakah pasien terhubung ke ventilator (1 = Iya, 2 = Tidak)	int64

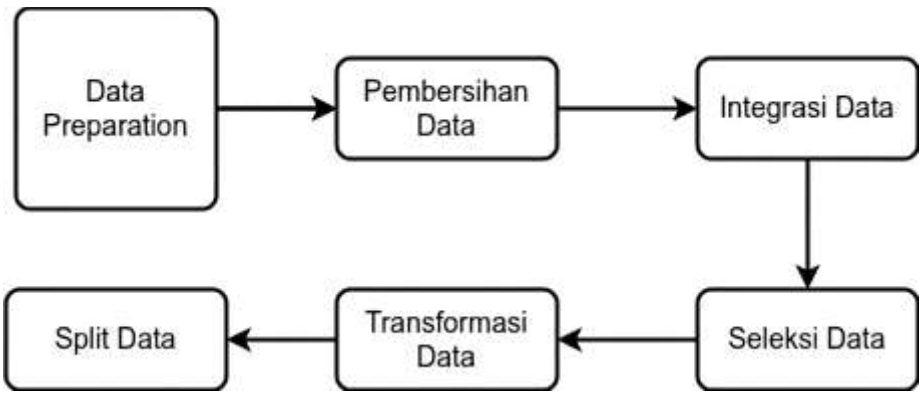
Parameters	Details	Value
USMR	Menunjukkan apakah pasien dirawat unit medis tingkat pertama, kedua atau ketiga (1 = Iya, 2 = Tidak)	int64
PNEUMONIA	Apakah pasien pernah mengalami penyakit Pneumonia (1 = Iya, 2 = Tidak)	int64
AGE	Umur pasien	int64
PREGNANT	Apakah pasien hamil (1 = Iya, 2 = Tidak)	int64
DIABETES	Apakah pasien menderita penyakit diabetes (1 = Iya, 2 = Tidak)	int64
COPD	Apakah pasien mengidap penyakit paru - paru obstruktif (1 = Iya, 2 = Tidak)	int64
ASTHMA	Apakah pasien menderita penyakit asma (1 = Iya, 2 = Tidak)	int64
INMSUPR	Apakah pasien mengalami *imunosupresi* (1 = Iya, 2 = Tidak)	int64
HIPERTENSION	Apakah pasien menderita penyakit hipertensi(1 = Iya, 2 = Tidak)	int64
OTHER_DISEASE	Apakah pasien memiliki penyakit lain (1 = Iya, 2 = Tidak)	int64
CARDIOVASCULAR	Apakah pasien memiliki penyakit terkait jantung atau pembuluh darah (1 = Iya, 2 = Tidak)	int64
OBESITY	Apakah pasien mengalami (1 = Iya, 2 = Tidak)	int64

Parameters	Details	Value
USMR	Menunjukkan apakah pasien dirawat unit medis tingkat pertama, kedua atau ketiga (1 = Iya, 2 = Tidak)	int64
RENAL_CHRONIC	Apakah pasien memiliki penyakit ginjal kronis (1 = Iya, 2 = Tidak)	int64
TOBACCO	Apakah pasien seorang perokok(1 = Iya, 2 = Tidak)	int64
CLASIFFICATION_FINAL	Temuan tes covid. Nilai 1 - 3 berarti pasien didiagnosis dengan covid dalam derajat yang berbeda. Nilai 4 atau lebih berarti pasien bukan pembawa covid atau tesnya tidak meyakinkan	int64
ICU	Apakah pasien dirawat di Unit Perawatan Intensif 1 = Iya, 2 = Tidak)	int64

Data pada tabel 1 diperoleh dari kaggle yang sudah memberikan keterangan dataser pada setiap baris. Pada tabel tersebut dilakukan pemahaman struktur data, dapat diketahui terdapat 21 kolom dalam dataset. Pada step 2 menganalisis hubungan antar kolom yaitu kolom *Diabetes*, *COPD*, *Asthma*, *Inmsupr*, *Hipertension*, *Other_disease*, *Cardiovascular*, *Obesity*, *Renal_chronic*, *dan Tobacco* adalah atribut yang nilainya paling tinggi diantara atribut lainnya, artinya naik turunnya memiliki hubungan yang signifikan dengan diagnosis kasus penyebab kematian pasien. Pada step 3 identifikasi pola dan tren data. Kolom *Patient_type* tidak memiliki hubungan yang signifikan dengan hasil diagnosis ICU dan *Intubed*, sehingga Kolom *ICU* dan *Intubed* dapat diabaikan.

2.3. *Data Preparation*

Pada tahap ini melibatkan persiapan data yang akan digunakan untuk pemodelan dan evaluasi prediksi kematian pasien *Covid-19* di Meksiko menggunakan metode CRISP-DM.



Gambar 3 Data Preparation

Pembersihan data, pada tahap ini peneliti melakukan pembersihan data dengan mengidentifikasi dan mengatasi nilai yang bernilai *null* atau nilai yang hilang. Jika terdapat multiple dataset, peneliti melakukan integrasi data dengan melibatkan penggabungan data. Selanjutnya peneliti menghilangkan variabel yang tidak relevan atau memiliki dampak yang kecil terhadap prediksi. Setelah itu peneliti melakukan transformasi data yang mencakup normalisasi, penggantian nilai yang hilang untuk memastikan bahwa data memenuhi persyaratan pemodelan yang akan digunakan. Terakhir melakukan pembagian dataset menjadi 80% subset pelatihan (training set) dan 20% subset pengujian (*testing set*).

2.4. *Modeling Data*

Tahap ini melibatkan pengembangan model data mining yang dapat digunakan untuk menggali informasi yang berguna dari data dengan memilih teknik data mining yang tepat untuk digunakan dan melakukan analisis data untuk mengembangkan model yang akurat.

Metode *Random Forest* adalah pengembangan dari metode *Classification and Regression Tree* (CART), yaitu dengan menerapkan metode *bootstrap aggregating* (bagging) dan *random feature selection* [7]. *Random Forest* merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi. Metode ini dapat meningkatkan hasil akurasi, dengan cara membangkitkan simpul anak untuk setiap node (simpul di atasnya) dan dilakukan pemilihan secara acak, Kemudian hasil klasifikasi dari setiap pohon diakumulasikan dan dipilih hasil klasifikasi yang paling banyak muncul [8]. Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1, sedangkan nilai information gain menggunakan persamaan 2 [9].

$$Entropy\ (Y) = -\Sigma ip(c|Y)\log^{2}p(c|Y)....(1)$$

Keterangan :
Y = Himpunan kasus
P(c|Y) = Proporsi nilai Y terhadap kelas c.
Information Gain (Y,a) =

$$\text{Entropy (Y)} - \Sigma_{V}\varepsilon \text{Values}(\alpha)_{X}).....(2)$$

Keterangan :
Values(a) = Nilai yang mungkin dalam himpunan kasus a.
Y_v = Subkelas dari Y dengan kelas v yang berhubungan dengan kelas a.
Y_a = Semua nilai yang sesuai dengan a.

2.5. *Evaluation*

Pada tahap evaluasi, peneliti dapat menilai kinerja model prediksi yang dikembangkan untuk memprediksi kematian pasien *Covid-19* di Meksiko. Skor yang baik menunjukkan keefektifan model dan keandalannya dalam peramalan yang akurat. Hasil evaluasi akan membantu peneliti untuk memvalidasi model, membuat keputusan tentang penggunaan model dalam konteks praktis, dan lebih memahami prediksi kematian pasien *Covid-19* di Meksiko.

2.5.1 *Feature Importance*

Feature importance adalah suatu proses menghapus features yang berlebihan dan tidak relevan dari dataset yang sebenarnya. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena features yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif.

2.5.2 *Confusion Matrix*

Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Dimana evaluasi confusion matrix adalah sebuah matrik dari prediksi yang

akan melakukan pengujian untuk memperkirakan obyek yang benar dan salah agar menghasilkan nilai akurasi, presisi dan recall. Presisi atau confidence adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Recall atau sensitivity adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar [10].

3. Hasil dan Pembahasan

3.1. Pembentukan Proses Model

Pada bagian pembentukan model penelitian ini menggunakan metode *Random Forest* untuk membentuk model. Berikut ini adalah tahap-tahap dan penjelasan dalam pembentukan model.

Tahap 1: Unduh data dari Kaggle yaitu *Covid 19 da*taset kemudian *import* data ke *google collaboratory* setelah itu memberi label data terhadap *dataset.*

Tahap 2: Membaca *dataset* dengan menampilkan data info, didalamnya terdapat nama kolom, jumlah kolom, jumlah data, dan tipe data.

Tahap 3: *Import library* yang akan digunakan untuk memproses data dan visualisasi.

Tahap 4: Preprocessing data dengan replace missing values yang sudah dijelaskan di bab 2, dan menghapus kolom yang tidak relevan.

Tahap 5: *Split* data atau membagi data sebesar masing-masing 80% dan 20% untuk *training* dan *testing* data

Tahap 6: Setelah itu membuat model dengan independent test.

Tahap 7: Evaluasi untuk mengetahui atribut mana yang paling berpengaruh terhadap kematian pasien *Covid-19* dalam dataset.

Tahap 8: Evaluasi menggunakan Confussion Matrix untuk memperkirakan obyek yang benar dan salah agar menghasilkan nilai akurasi, presisi dan recall.

3.2. Hasil *Feature Importance*

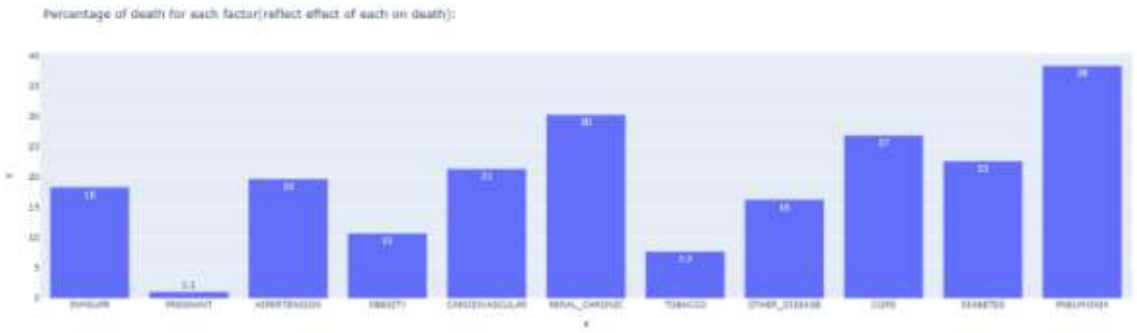
Hasil dari tabel 2 merupakan *Feature Importance* dari Covid19 dataset dimana INMSUPR (imunosupresi), kehamilan, hipertensi, obesitas, cardiovascular (penyakit jantung atau pembuluh darah), ginjal kronis, tobacco (pengguna tembakau), COPD (penyakit paru obstruktif kronik), diabetes, pneumonia dan penyakit lainnya sangat berpengaruh terhadap penyebab kematian pasien Covid19 di Meksiko. Berikut adalah tabel persentase pengaruh kematian pasien:

Tabel 2 Hasil *Feature Importance*

Faktor Kematian	Persentase
Inmsupr	18,35%
Pregnant	1,07%
Hipertension	19,71%
Obesity	10,74%
Cardiovascular	21,36%

Renal Chronic	30,28%
Tobacco	7,75%
Other Disease	16,31%
COPD	26,87%
Diabetes	22,61%
Pneumonia	38,67%

Data tabel 2 diperoleh dari hasil *Feature Importance* berdasarkan dataset. Bagi pasien yang mempunyai penyakit yang ada di dalam tabel sangat mempengaruhi hasil kematian pasien, karena pasien yang meninggal pasti mempunyai salah satu atau lebih penyakit yang ada di daftar sehingga dapat memprediksi kematian pasien. Sedangkan atribut lain sama sekali tidak mempengaruhi dalam kematian pasien.



Gambar 4. Grafik Hasil Persentase Faktor Kematian

Tabel 3 Hasil *Performance Evaluation Cross Validation*

	Persentase Hasil *Cross Validation*
Akurasi	90,67%
Recall	91%
Precision	91%

Data pada tabel 3 diperoleh dari hasil pengujian menggunakan pemrograman *Cross Validation*. Dari hasil pengujian hasil klasifikasi ditunjukkan dengan 80% data *training* dan 20% jika

menggunakan algoritma *Random Forest* dan hasil yang didapatkan *accuracy* sebesar 90.67%, recall 91%, dan *precision* 91%

Tabel 4 *Confusion Matrix Cross Validation Evaluation*

	True Rendah	*True* Tinggi
Prediksi Rendah	13896	1760
Prediksi Tinggi	1028	13202

Data pada tabel 4 diperoleh dari hasil *Confution Matrix Cross Validation Evaluation* berdasarkan model yang telah dibuat. Dapat dilihat bahwa nilai dari true positive mendapatkan nilai dari prediksi rendah ***13896*** dan *true negative* mendapatkan nilai ***1760*** yang merupakan hasil dari data *training* yang telah dilakukan. Berikutnya nilai dari *true positif* dari prediksi tinggi mendapatkan nilai ***1028*** dan *true negative* ***13202***

4.Kesimpulan

Pada penelitian ini dikembangkan model Data Mining untuk prediksi kematian status pasien Covid-19 di Meksiko menggunakan algoritma *Random Forest*. Algoritma ini digunakan karena kehandalannya dalam mengklasifikasi data berdasarkan atribut-atribut yang dimiliki, baik berupa nilai numerik maupun kategorik. Model dibangun menggunakan dataset pasien Covid-19 di Meksiko yang didapat dari website www.kaggle.com dan diimplementasikan menggunakan metode proses *CRISP-DM*. Hasil dari model *Random Forest* untuk prediksi kematian pasien Covid-19 di Meksiko memberikan hasil yang diukur dalam nilai akurasi, recall, dan presisi, berturut-turut nilainya 90,67%, 91%, dan 91%. Nilai akurasi yang tinggi menunjukkan kinerja *Random Forest* yang baik dalam mengklasifikasikan kematian pasien, yaitu hidup atau tidak. Hasil dari penelitian ini bermanfaat untuk diterapkan pada situasi nyata, untuk membantu tenaga medis menentukan tindakan. Ke depan, jumlah dataset nyata dan berukuran besar dengan proporsi nilai setiap kelas yang seimbang sangat baik untuk mendapatkan akurasi prediksi yang lebih tinggi.

3.3. Daftar Pustaka

[1] C. Long et al., “Diagnosis of the Coronavirus disease (Covid-19): rRT-PCR or CT?,” Eur. J. Radiol., vol. 126, p. 108961, May 2020, doi: 10.1016/j.ejrad.2020.108961.

[2] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke, “High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2,” Emerg. Infect. Dis. J., vol. 26, no. 7, 2020, doi: 10.3201/eid2607.200282.

[3] *“Meksiko Situation”* [Online]. Available: https://covid19.who.int/region/amro/country/mx.

[4] M. Abed Alah, S. Abdeen, and V. Kehyayan, “The first few cases and fatalities of Corona Virus Disease 2019 (Covid-19) in the Eastern Mediterranean Region of the World Health Organization: A rapid review,” J. Infect. Public Health, vol. 13, no. 10, pp. 1367–1372, 2020, doi: 10.1016/j.jiph.2020.06.009

[5] M. Habibi, “Analisis Konten Jejaring Sosial Twitter dalam Kasus Pemilihan Gubernur DKI 2017,” Teknomatika, vol. 11, no. 1, pp. 31–40, 2018.

[6] *“Covid-19 Dataset”* [Online]. Available: https://www.kaggle.com/datasets/meirnizri/covid19-dataset.

[7] L. Breiman, ‘Random Forests’, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] F. A. Kurniawan and A. P. Kurniati, ‘Analisis Dan Implementasi Random Forest Dan Classification dan Regression Tree (Cart) Untuk Klasifikasi Pada Misuse Intrusion Detection System’

[9] Y. S. Nugroho, ‘Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest’, vol. 9, no. 1, p. 6, 2017.

[10] Herdiawan, ‘Analisis Sentimen Terhadap Telkom Indihome Berdasarkan Opini Publik Menggunakan Metode Improved K-Nearest Neighbor’, Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)