IMTechno: Journal of Industrial Management and Technology
Volume 5 No. 2 Juli 2024
E-ISSN: 2774-342X

Komparasi Algoritma *Decision Tree, Random Forest* dan SVM untuk Prognosis COVID-19

Agung Wibowo1, Indarti2, Dewi Laraswati3

[1,2,3]Sistem Informasi

[1,2,3]Universitas Bina Sarana Informatika
e-mail: [1]agung.awo@bsi.ac.id, [2]indarti.ini@bsi.ac.id, [3]dewi.dwl@bsi.ac.id

Diterima	Direvisi	Disetujui
19-10-2023	18-05-2024	22-07-2024

Abstrak - Virus corona merupakan virus yang dapat menginfeksi manusia sehingga menyebabkan infeksi saluran pernafasan parah seperti *Middle East Respiratory Syndrome* (MERS) dan *Severe Acute Respiratory Syndrome* (SARS). Virus corona baru, SARSCoV-2, adalah virus yang menyerang sistem pernapasan dan menyebabkan gejala parah seperti batuk, bersin, dan ruam. Ini dapat menyerang siapa saja, termasuk anak-anak, orang dewasa, orang tua, anak-anak, dan orang dewasa. Pada tahun 2020, virus corona berubah menjadi berbagai varian, antara lain *Alpha, Beta, Gamma, Delta, Lambda*, dan *Kappa*. Varian-varian tersebut memberikan dampak signifikan terhadap kesehatan masyarakat, khususnya di Indonesia. Penelitian ini bertujuan untuk mengidentifikasi gelaja Covid-19 dengan cara mengkomparasi dan mengklasifikasi gejala Covid-19 menggunakan algoritma pembelajaran mesin, khususnya *decision tree, random forest*, dan *Support Vector Machine* (SVM). Hasil penelitian menunjukkan bahwa *decision tree* memiliki tingkat akurasi yang lebih tinggi dibandingkan *random forest* dan SVM, dengan skor F1 yang lebih tinggi mendekati 1,0 yaitu 0,97. Penelitian ini juga menemukan bahwa Algoritma *decision tree* memiliki nilai confusion matrix lebih baik dibandingkan dua algoritma lainnya.
Kata Kunci: Covid-19, *Decision Tree, Klasifikasi*

Abstract *– The Corona virus is a virus that can infect humans, causing severe respiratory tract infections such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The new coronavirus, SARSCoV-2, is a virus that attacks the respiratory system and causes severe symptoms such as coughing, sneezing, and rashes. It can affect anyone, including children, adults, and the elderly. In 2020, the Corona virus changed into various variants, including Alpha, Beta, Gamma, Delta, Lambda, and Kappa. These variants have a significant impact on public health, especially in Indonesia. This research aims to identify COVID-19 symptoms by comparing and classifying them using machine learning algorithms, specifically decision trees, random forests, and support vector machines (SVM). The research results show that the decision tree has a higher level of accuracy than random forest and SVM, with a higher F1 score approaching 1.0, namely 0.97. This research also found that the decision tree algorithm has a better confusion matrix value than the other two algorithms.*
Keywords: COVID-19, Decision Tree, Classification

PENDAHULUAN

Coronavirus merupakan suatu kelompok virus yang dapat menyerang pada hewan atau manusia. Beberapa diantaranya jenis virus ini dapat menyebabkan infeksi saluran pernafasan pada manusia mulai dari batuk pilek hingga yang lebih serius seperti *Middle East Respiratory Syndrome* (MERS) dan *Severe Acute Respiratory Syndrome* (SARS). *Coronavirus* jenis baru yang ditemukan menyebabkan penyakit COVID-19 (World Health Organization, n.d.). Virus corona atau *severe acute respiratory syndrome coronavirus 2* (SARSCoV-2) adalah virus yang menyerang sistem pernapasan. Penyakit karena infeksi virus ini disebut Covid 19. Virus corona bisa menyebabkan gangguan ringan pada sistem pernapasan, infeksi paru- paru yang berat, hingga kematian. *Severe acute respiratory syndrome corona virus 2* (SARS- CoV-2) yang lebih dikenal dengan nama virus corona adalah jenis baru dari corona virus yang menular ke manusia. Virus ini bisa menyerang siapa saja, baik bayi, anak-anak, orang dewasa, lansia, ibu hamil, maupun ibu menyusui (Kemenkes, 2020).

Corona virus adalah kumpulan virus yang bisa menginfeksi sistem pernapasan (Kemenkes, 2020).

Pada akhir tahun 2020, *corona virus* bermutasi menjadi berbagai macam jenis varian diantaranya varian *Alfa, Beta, Gamma, Delta, Lambda* dan *Kappa*. Virus varian ini menyebar keseluruh belahan dunia termasuk di Indonesia (Adrian, n.d.). Kemunculan beragam varian Covid-19 ini terjadi diantaranya karena adanya respon yang diberikan oleh virus terhadap perubahan lingkungan yang ada. Dengan adanya mutasi virus maka menyebabkan timbulnya varian baru dari covid -19. Dari banyaknya varian yang muncul, beberapa varian memiliki tingkatan kategori masing-masing diantaranya varian dalam kategori menjadi perhatian utama (*variant of concern*) dimana tingkat penularan dan dampak yang dihasilkan dalam masyarakat cukup signifikan karena tingkat penularannya dan dampaknya cukup signifikan. Ada juga kategori *variant of interest*, yaitu varian covid-19 yang diperkirakan bisa berdampak pada masyarakat. Dan kategori lainnya yaitu variant under monitoring, yaitu varian covid-19 yang belum diketahui bagaimana penyebaran dan dampaknya bagi manusia (Alya Eka Putri, n.d.).

Dalam proses pengklasifikasian terdapat beberapa algoritma yang umum digunakan, yaitu algoritma *decision tree, algoritma random forest* dan algoritma *support vector machine* (SVM). Setiap metode memiliki hasil dan cara pemprosesan data yang berbeda-beda. Oleh karena itu, diperlukan suatu algoritma yang tepat untuk pengklasifikasian Covid-19. Metode klasifikasi tersebut merupakan algoritma berbasis machine learning. Sedangkan tujuan dari penelitian ini yaitu untuk membandingkan algoritma *decision tree, algoritma random forest* dan algoritma *support vector machine* (SVM), sehingga dapat diketahui algoritma yang terbaik untuk klasifikasi Covid-19.

METODE PENELITIAN

Paper ini menggunakan data sekunder yang diambil dari *https://www.kaggle.com/datasets/maulanazhahran/dataset-covid* dengan dimensi 21 kolom dan 5434 baris tipe isi data bernilai ordinal (*Yes, No*), dengan variabel COVID-19 sebagai *class*-nya. Sebaran data dapat dilihat pada pada Tabel 1. Sebaran *Dataset* Klasifikasi COVID-19.

Tabel 1. Sebaran Dataset Klasifikasi COVID-19

	count	*unique*	*top*	*freq*
Breathing Problem	5434	2	Yes	3620
Fever	5434	2	Yes	4273
Dry Cough	5434	2	Yes	4307
Sore throat	5434	2	Yes	3953
Breathing Problem	5434	2	Yes	3620
Fever	5434	2	Yes	4273
Dry Cough	5434	2	Yes	4307
Running Nose	5434	2	Yes	2952
Asthma	5434	2	No	2920
Chronic Lung Disease	5434	2	No	2869
Headache	5434	2	Yes	2736
Heart Disease	5434	2	No	2911
Diabetes	5434	2	No	2846
Hyper Tension	5434	2	No	2771
Fatigue	5434	2	Yes	2821
Gastrointestinal	5434	2	No	2883
Abroad travel	5434	2	No	2983
Contact with COVID Patient	5434	2	Yes	2726
Attended Large Gathering	5434	2	No	2924
Visited Public Exposed Places	5434	2	Yes	2820
Family working in Public Exposed Places	5434	2	No	3172
Wearing Masks	5434	1	No	5434
Sanitization from Market	5434	1	No	5434
COVID-19	5434	2	Yes	4383

Sumber: diolah oleh peneliti (2023)

dan paper ini mengikuti alur penelitian pada gambar 1. Alur penelitian berikut:

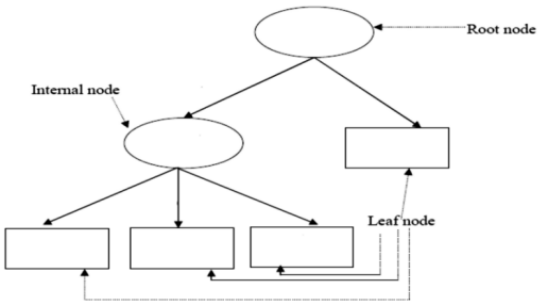
Sumber: diolah oleh peneliti (2023)
Gambar 1. Alur penelitian

Pada tahap *Data Preparation*, kami menghapus variabel yang tidak digunakan dan *encode* data. Proses *encode* data ini lakukan agar dataset mudah diproses secara matematis dan statistika sehingga hasil nya pun *reliable*. Pada tahap ini variabel data

yang dihapus adalah '*Wearing Masks*' dan '*Sanitization from Market*'. Tahap selanjutnya kami melakukan uji *Classification Algorithm*, sebelumnya dataset dibagi menjadi data *training* sejumlah 3622 baris *record* dan data testing sejumlah 1812 baris record. Kami mengujikan *dataset* menggunakan tiga algoritma klasifikasi populer menurut peneliti sebelumnya yaitu: "*Random Forests*" (Breiman, 2001) , "*Support Vector Machines*" (Cortes & Vapnik, 1995) , dan "*Decision Trees*" (Kotz & Johnson, 1992). Ketiga algoritma tersebut dapat dilihat hasil komparasi akurasinya sehingga dapat dipilih algoritma terbaik untuk mengklasifikasi. Berikut ini penjelasan singkat dari tiga algoritma yang kami gunakan dalam penelitian ini.

1. *Decision Tree*

Decision Tree adalah salah satu cara pemprosesan data untuk memprediksi masa depan dengan cara membangun klasifikasi atau regresi model dalam bentuk struktur pohon, lihat gambar 2. Hasil dari pemprosesan menggunakan *decision tree* yaitu berupa pohon dengan node keputusan dan node daun. Salah satu alasan menggunakan metode *decision tree* yaitu metode ini dapat mengeliminasi perhitungan atau data-data yang tidak diperlukan. Karena sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu. Namun pada metode ini ada kalanya pohon keputusan yang dihasilkan memungkinkan terjadinya tumpang tindih, terutama jika kelas dan kriteria yang digunakan sangat sering dapat meningkatkan waktu pengambilan keputusan sesuai dengan kapasitas memori yang diperlukan (Ramadhan, n.d.).

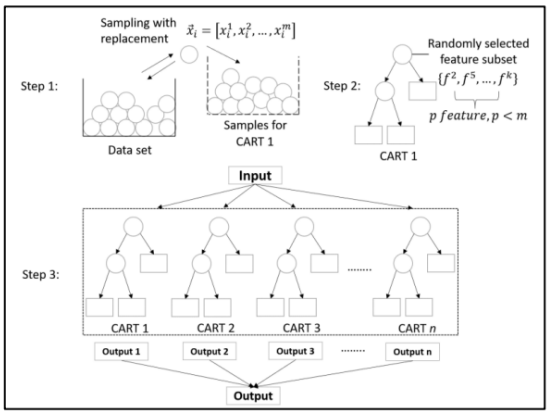


Sumber (Supriyadi et al., 2020)

Gambar 2. Struktur Decision Tree

2. *Random Forest*

Random forest merupakan suatu algoritma dalam melakukan klasifikasi yang termasuk *ensamble learning* (Fadillah et al., 2022). Pada algoritma ini terdapat k pohon dengan vektor random yang *independent* dengan vektor-vektor *random* sebelumnya, tetapi memiliki distribusi yang identik (Breiman, 2001). juga menyatakan bahwa metode ini memanfaatkan algoritma *decision tree* dalam melakukan klasifikasi. Kemudian dibentuk sebuah model dengan menerapkan metode *bootstrap aggregating* ketika membentuk sebuah sampel *training set*, dan setiap tree yang dibentuk menggunakan metode yang sama untuk membangun CART (*Classification and Regression Tree*). Berikut adalah proses dari algoritma *random forest* yang terlihat pada gamabr 3.



Sumber :(Fadillah et al., 2022)

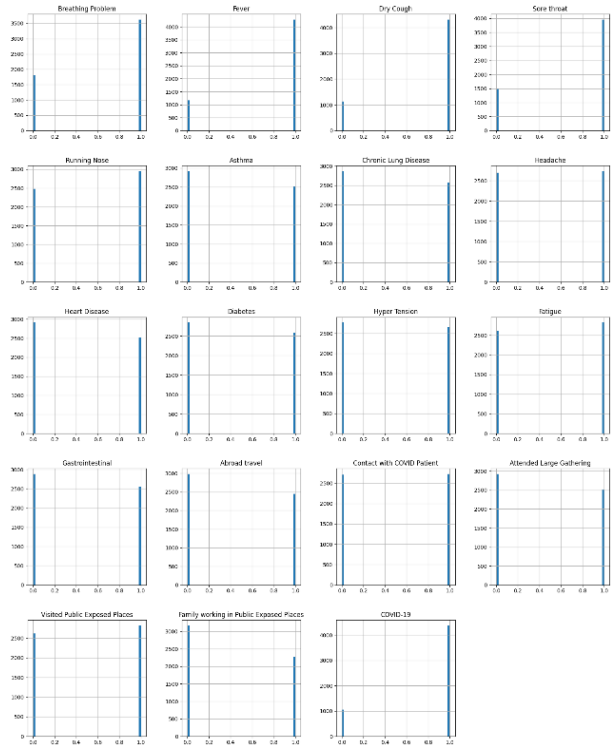
Gambar 3. Proses Algoritma *Random Forest*

3. SVM (*Support Vector Machine*)

SVM (*Support Vector Machine*) adalah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. SVM mencari *hyperplane* terbaik yang memisahkan dua kelas data dengan margin terbesar. *Hyperplane* ini ditemukan dengan memaksimalkan jarak antara *hyperplane* dan titik-titik terdekat dari kedua kelas data. SVM juga dapat digunakan untuk regresi, yaitu memprediksi nilai numerik berdasarkan fitur-fitur yang diberikan (Khikmanto Supribadi, 2014).

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan bahasa pemrograman python 3.9.13. Pada tahap pertama, proses data *preparation*, dataset dianalisa dengan cara memeriksa data yang kosong, variasi isi data, tipe isi data, menghapus variabel yang tidak digunakan dan *encode* data. Sebaran data dari *dataset* ini dapat dilihat pada gambar 4. Sebaran data *dataset* COVID-19



Sumber: diolah oleh peneliti

Gambar 3. Sebaran data dataset COVID-19

Hasil pengujian akurasi dari ketiga algoritma pada saat *traning* menunjukkan bahwa nilai akurasi nya sama yaitu 98%, tetapi ketika model hasil *training* dilakukan *testing*, nilai akurasi algoritma *decision tree* lebih baik dari dua algoritma lainnya. Nilai akurasi dari setiap algoritma yang dikomparasi dapat dilihat pada tabel 2. Nilai Komparasi Akurasi.

Tabel 2. Nilai Komparasi Akurasi

	Decision Tree	*Random Forest*	SVM
Training	0.984538928 7686361	0.984538928 7686361	0.984538928 7686361
Testing	0.97**9028697 5717439**	0.976269315 6732892	0.976269315 6732892

Sumber: diolah oleh peneliti

Peningkatan akurasi klasifikasi data dapat dilihat berdasarkan nilai dari *confusion matrix, Confusion Matrix* adalah tabel dengan kombinasi nilai prediksi dan nilai aktual, representasi hasil proses klasifikasi nilai *True Positif, True Negatif, False Positif,* dan *False Negatif.* Ilustrasi dari *confusion matrix* dapat dilihat pada gambar 4. *Confusion Matrix.*

Sumber: diolah oleh peneliti

Gambar 4. *Confusion Matrix*

Berdasarkan nilai *confusion matrix* dari ketiga algoritma yang dikomparasi dalam paper ini (lihat tabel 3. Hasil Komparasi *Confusion Matrix*), menunjukkan bahwa algoritma *decision tree* memiliki nilai *False Positif* lebih rendah dibandingkan kedua algoritma lainnya. Nilai *False Positif* dari algoritma *decision tree* memiliki selisih 3 nilai dibandingkan dengan *random forest* dan selish 8 nilai lebih baik dibandingkan dengan SVM yang artinya kesalahan mengklasifikasi nilai yang seharusnya masuk kedalam kelompok benar menggunakan algoritm *decision tree* lebih kecil dibandingkan algoritma lainnya.

Tabel 3. Hasil Komparasi *Confusion Matrix*

Decision Tree	***Random Forest***	**SVM**
[[352 **18**] [20 1422]]	[[349 21] [20 1422]]	[[344 26] [17 1425]]

Sumber: diolah oleh peneliti

Hasil pengujian secara statistik kami menggunakan *Precision, Recall, f1-Score. Recall,* adalah perbandingan antara *True Positive* (TP) dengan banyaknya data yang sebenarnya positif. *f1-Score* adalah *harmonic mean* dari *precision* dan *recall*, secara matematik ditulis sebagai berikut:

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right)$$

Nilai terbaik *F1-Score* adalah 1.0 dan nilai terburuknya adalah 0. Secara representasi, jika *F1-Score* punya skor yang baik mengindikasikan bahwa model klasifikasi kita punya *precision* dan *recall* yang baik. *Macro average* adalah rata-rata presisi, *recall* atau *f1-score* dari semua kelas. Nilai statitistik dari pengujian menunjukkan bahwa nilai *Recall, f1-score* dan *Macro average* dari algoritma *decision tree* lebih besar dibandingkan dengan kedua algoritma lainnya yang dikomparasi. Algoritma SVM memiliki nilai *recall* dan *macro average* terendah. Secara lengkap nilai hasil uji dari setiap algoritma dapat dilihat pada tabel 4, tabel 5 dan tabel 6.

Tabel 4. Hasil Komparasi *Classification report Decision Tree*.

Decision Tree				
	precision	*recall*	*f1-score*	*support*
0	0.95	**0.95**	**0.95**	370
1	0.99	0.99	0.99	1442
accuracy			0.98	1812
macro avg	0.97	**0.97**	**0.97**	1812
weighted avg	0.98	0.98	0.98	1812

Sumber: diolah oleh peneliti

Tabel 5. Hasil Komparasi *Classification report Random Forest*.

Random Forest				
	precision	*recall*	*f1-score*	*support*
0	0.95	0.94	0.94	370
1	0.99	0.99	0.99	1442
accuracy			0.98	1812
macro avg	0.97	**0.96**	0.97	1812
weighted avg	0.98	0.98	0.98	1812

Sumber: diolah oleh peneliti

Ttabel 6. Hasil Komparasi *Classification report SVM*.

SVM				
	precision	*recall*	*f1-score*	*support*
0	0.95	**0.93**	0.94	370
1	0.99	0.99	0.99	1442
accuracy			0.98	1812
macro avg	0.97	0.96	**0.96**	1812
weighted avg	0.98	0.98	0.98	1812

Sumber: diolah oleh peneliti

KESIMPULAN

Hasil pengujian akurasi dari ketiga algoritma pada saat traning menunjukkan bahwa nilai akurasi nya sama yaitu 98%, tetapi ketika model hasil *training* dilakukan *testing*, nilai akurasi algoritma *decision tree* lebih baik dari dua algoritma lainnya.

Nilai statitistik dari pengujian menunjukkan bahwa nilai *Recall, f1-score* dan *Macro average* dari algoritma *decision tree* lebih besar dibandingkan dengan kedua algoritma lainnya yang dikomparasi. Algoritma SVM memiliki nilai *recall* dan *macro average* terendah.

Berdasarkan nilai *confusion matrix* dari ketiga algoritma yang dikomparasi dalam paper ini, menunjukkan bahwa algoritma *decision tree* memiliki nilai *False Positif* lebih rendah dibandingkan kedua algoritma lainnya. Nilai *False Positif* dari algoritma *decision tree* memiliki selisih 3 nilai dibandingkan dengan *random forest* dan selisih 8 nilai lebih baik dibandingkan dengan SVM yang artinya kesalahan mengklasifikasi nilai yang seharusnya masuk kedalam kelompok benar menggunakan algoritm *decision tree* lebih kecil dibandingkan algoritma lainnya.

REFERENSI

Alya Eka Putri, S. (n.d.). *Covid-19*. Retrieved October 19, 2023, from https://corona.jakarta.go.id/id/artikel/varian-varian-covid-19-apa-perbedaannya

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Fadillah, I. J., Fadila, L. M. A., & Darundiye, L. M. W. (2022). Perbandingan Hot-deck, SVM, dan Random Forest dalam Mengidentifikasi Industri Mikro dan Kecil Terdampak Covid-19 Tahun 2020. *Seminar Nasional Official Statistics*, *2022*(1), 147–154.

Kemenkes, R. I. (2020). *Pedoman Pencegahan dan Pengendalian Serta Definisi Coronavirus Disease (COVID-19). Germas*, *11–45*.

Adrian, dr. K. (n.d.). *Kenali Perbedaan COVID-19 Varian Alfa, Beta, Gamma, Delta, Lambda, dan Kappa - Alodokter*. Retrieved October 19, 2023, from https://www.alodokter.com/kenali-perbedaan-covid-19-varian-alfa-beta-gamma-dan-delta

Khikmanto Supribadi, S. T. (2014). *Analisis metode support vector machine (SVM) untuk klasifikasi penggunaan lahan berbasis penutup lahan pada citra ALOS AVNIR-2*. Universitas Gadjah Mada.

Kotz, S., & Johnson, N. L. (1992). *Breakthroughs in Statistics* (S. Kotz & N. L. Johnson, Eds.).

Springer New York. https://doi.org/10.1007/978-1-4612-4380-9

Ramadhan, A. , S. (n.d.). *DECISION TREE ALGORITMA BESERTA CONTOHNYA PADA DATA MINING – School of Information Systems*. Retrieved October 19, 2023, from https://sis.binus.ac.id/2022/01/21/decision-tree-algoritma-beserta-contohnya-pada-data-mining/

Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis: Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 67–75.

World Health Organization. (n.d.). *Pertanyaan jawaban terkait COVID-19 untuk publik*. Retrieved October 19, 2023, from https://www.who.int/indonesia/news/novel-coronavirus/qa/qa-for-public