KOMPARASI ALGORITMA UNTUK KLASIFIKASI DATASET COVID-19 SURVILLANCE

Ivandari, M. Adib Al Karomi
Program Studi Sistem Informasi STMIK Widya Pratama Pekalongan
Email: ivandarialkaromi@gmail.com; adib.comp@gmail.com

Abstrak

Pandemi covid-19 yang terjadi di dunia saat ini merupakan yang terbesar selama lebih dari satu decade. Banyaknya penderita, kurangnya kesadaran masyarakat, serta adanya mutasi baru dari virus merupakan beberapa faktor pendukung kenaikan kasus covid-19. Di semua bidang dilakukan penelitian guna menurunkan tingkat keparahan penyakit, menghentikan penyebaran, serta mengobati pasien terdampak virus. Di bidang data mining banyak dilakukan penelitian dengan menggunakan dataset yang ada untuk memperoleh pengetahuan baru. Pengetahuan inilah yang nantinya dapat digunakan sebagai dasar untuk menentukan kebijakan institusi, perusahaan bahkan untuk pemerintahan. Dalam penelitian ini membandingkan algoritma KNN, Neural Network, Bayes, serta Decission Tree untuk klasifikasi dataset Covid-19 survillance. Dataset yang digunakan adalah data dari Kementrian Kesehatan Republik Indonesia. Data ini diambil dari portal penyedia data ternama yaitu uci repository. Dari hasil perhitungan membuktikan bahwa algoritma decision tree merupakan model terbaik untuk klasifikasi dataset Covid-19 survillance. Akan tetapi tingkat akurasi yang diperoleh decision tree dalam klasifikasi ini hanyalah 65% yang masih tergolong dalam tingkatan yang belum memuaskan.

Kata Kunci : covid-19 *survillance*, akurasi, decision tree

1. PENDAHULUAN

1.1 Latar Belakang

Covid-19 merupakan sebuah pandemi baru yang dialami sebagian besar Negara di dunia. *Coronavirus Diseases* atau sering lebih dikenal dengan covid-19 adalah penyakit yang mudah menular pada manusia. Penyebaran virus yang sangat cepat serta banyaknya mutasi virus baru membuat penyakit ini menjadi objek penelitian yang popular. Kasus Covid-19 pertama kali ditemukan di Indonesia pada Maret tahun 2020. Penyebaran virus ini tergolong sangat cepat. Di Indonesia sendiri telah terimbas 3 kali gelombang penyakit ini. Mutasi terakhir yang tercatat dari virus ini adalah varian omicron yang dikabarkan terdampak di banyak Negara pada awal 2022.

Beberapa kebijakan diambil oleh pemerintah guna menanggulangi meluasnya omicron ini. Dari semua bidang melakukan hal terbaik serta banyak menghasilkan penemuan baru untuk menunjang pemulihan pasca pandemi. Di bidang informatika terdapat data mining yaitu bidang ilmu yang dapat mengolah data untuk menjadi pengetahuan baru.

(Witten et al., 2011). Di dalam data mining sebenarnya terdapat lagi berbagai sub disiplin perhitungan, seperti klasifikasi, asosiasi, klustering, prediksi dan estimasi. Berbagai model dalam perhitungan klasifikasi dilakukan peneliti untuk mendapatkan hasil terbaik. Klasifikasi sebelumnya pernah dilakukan dengan menggunakan algoritma KNN (Ivandari & Al Karomi, 2021a), serta menggunakan algoritma Decission tree (Ivandari & Al Karomi, 2021b).

Uci Machine Learning Repository adalah basis data yang banyak digunakan guna melakukan uji coba metode atau algoritma. Data *Uci Machine Learning Repository* ini adalah data yang dapat diakses semua orang, serta telah teruji sebelumnya oleh tim dari uci. Dari banyak bidang, terdapat pula banyak macam data yang

dapat digunakan untuk uji ciba algoritma. Dalam penelitian ini digunakan dataset penderita Covid-19 serta berbagai gejalanya yang diambil ari Kementrian Kesehatan Republik Indonesia. Link data dapat diambil pada: (https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance).

Penelitian ini membandingkan performa algoritma klasifikasi KNN, Neural Network, Bayes, serta Decission Tree. Hasil klasifikasi membuktikan bahwa Decission tree memiliki tingkat akurasi terbaik yaitu 65%. Diikuti dengan Bayes dengan tingkat akurasi 60%, KNN dengan tingkat akurasi 55% dan Neural Network memiliki tingkat akurasi terkecil yaitu 45%.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan menggunakan metode eksperimen. Yaitu dengan menggunakan dataset yang ada dan dilakukan perhitungan menggunakan masing masing algoritma klasifikasi yang ada. Proses perhitungan dilakukan dengan bantuan rapid miner.

2.1. Pengumpulan Data

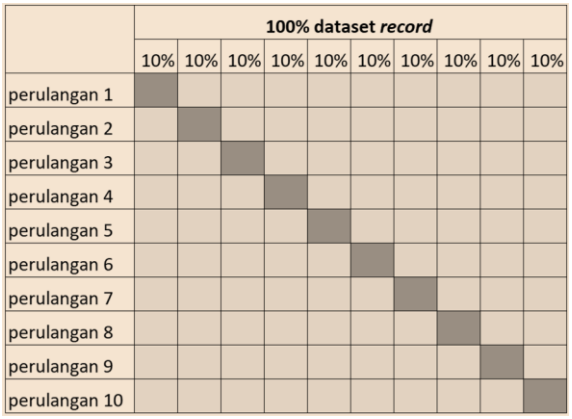
Pengumpulan data dilakukan dengan mengambil data dari data publik. Data ini dapat diambil di: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance. Dalam data ini terdapat 7 atribut regular dan 1 atribut label.

2.2. Perhitungan Algoritma

Proses perhitungan dilakukan dengan menggunakan rapid miner. Dalam proses perhitungan algoritma klasifikasi ini terdapat dua tahapan yaitu proses validasi dan proses evaluasi hasil.

2.3. *10 folds cross validation*

Tahapan ini merupakan proses validasi dengan menggunakan software rapid miner. Proses ini dilakukan dengan membagi dataset menjadi sepuluh bagian, dengan satu bagian dijadikan sebagai data uji dan sembilan bagian lainnya sebagai data pembanding. Proses ini diulang sepuluh kali dengan bagian lain diubah menjadi data uji pada perulangan berikutnya. Proses perulangan akan berakhir bila semua *record* sudah mendapat bagian satu kali menjadi data uji. Representasi *10 folds cross validation* dapat dilihat pada gambar 1.



Gambar 1. Representasi *10 folds cross validation*

2.4. *Confussion matrix*

Matrix kebingungan atau *confussion matrix* adalah salah satu model evaluasi yang banyak digunakan pada penelitian data mining, terutama di bidang klasifikasi dan prediksi. Proses ini sebenarnya membandingkan data prediksi atau klasifikasi dengan data asli. Perhitungan sederhananya adalah jumlah data prediksi atau klasifikasi yang sesuai dengan kondisi asli dibagi dengan keseluruhan total perhitungan yang dilakukan. Hasil akhir dari perhitungan akan didapatkan output presentase kebenaran (Ivandari & Al Karomi, 2021a). Gambar 2 merupakan representasi proses perhitungan matrix kebingungan.

Classification		*Predicted class*	
		Class: YES	*Class*: NO
Observed class	*Class*YES	***a*** ***True Positive (TP)***	***b*** ***False Negative (FN)***
	*Class*NO	***c*** ***False Positive (FP)***	***d*** ***True Negative (TN)***

Gambar 2. *Confussion matrix* (Gorunescu, 2011)

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data dan Analisa Data

Tahapan pertama dari penelitian ini adalah pengumpulan data. Dataset dari penelitian ini bersumber dari Kementrian Kesehatan Republik Indonesia yang sebelumnya telah dijadikan data public di uci repository. database Covid-19 survillance ini dapat diunduh di url: https://archive.ics.uci.edu/ml/datasets/COVID-19+Surveillance. Pada dataset tersebut terdapat 7 atribut regular serta 1 atribut label atau atribut tujuan. Tabel 1 merupakan dataset yang digunakan dalam penelitian ini.

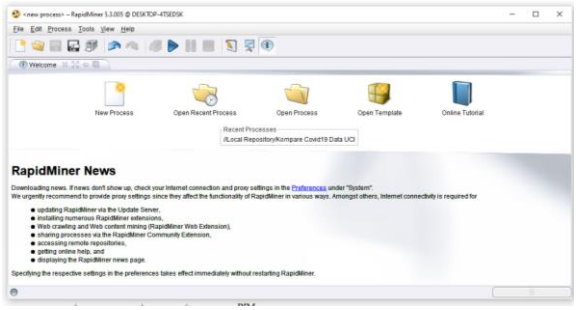
Tabel 1. Dataset Covid-19 survillance

A01	A02	A03	A04	A05	A06	A07	Categories
+	+	+	+	+	-	-	PUS
+	+	-	+	+	-	-	PUS
+	+	+	+	-	+	-	PUS
+	+	-	+	-	+	-	PUS
+	-	-	-	-	-	+	PUS
+	+	+	-	-	-	+	PUS
+	+	-	-	-	-	+	PUS
+	+	+	+	-	-	-	PUS
+	-	-	+	+	-	-	PIM
-	+	-	+	+	-	-	PIM
+	-	-	+	-	+	-	PIM
-	+	-	+	-	+	-	PIM
-	+	-	-	-	-	+	PIM
-	-	-	-	-	-	+	PWS

Dari tabel 1 diketahui bahwa semua atribut data yang ada menggunakan tipe nominal. Tipe data nominal ini tidak memungkinkan untuk dibandingkan satu nilai dengan nilai lainnya. Beberapa metode yang dapat digunakan untuk klasifikasi data bertipe nominal antara lain KNN, Bayes, decision tree.

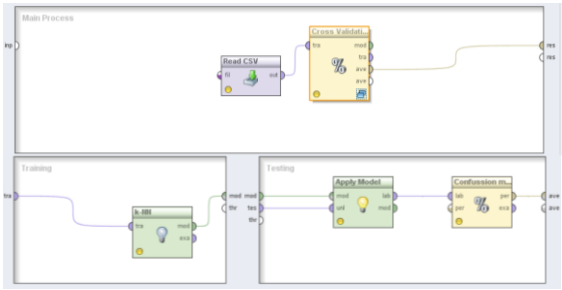
3.2. Perhitungan Algoritma

Perhitungan algoritma dilakukan dengan bantuan alat atau aplikasi rapid miner. Aplikasi rapid miner ini banyak digunakan dan terbukti dapat dengan singkat melakukan komparasi perhitungan algoritmik (Alkaromi, 2014). Gambar 3 dibawah adalah halaman utama atau halaman pembuka aplikasi rapid miner.



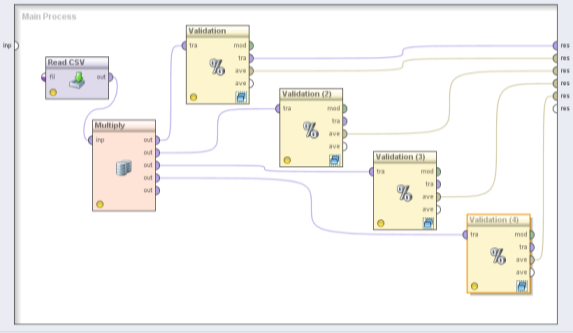
Gambar 3. Halaman utama rapid miner

Proses perhitungan dilakukan dengan melakukan validasi menggunakan *10 folds cross validation*, serta evaluasi menggunakan *confussion matrix.* Proses validasi dengan menggunakan rapid miner dilakukan sesuai dengan gambar 4.



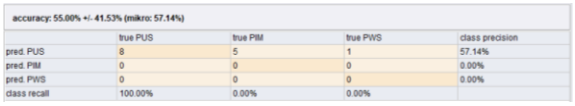
Gambar 4. Proses validasi untuk KNN

Lalu untuk algoritma lainnya dilakukan dengan cara yang sama pada gambar 4. Perbedaannya hanya pada tools metode/algoritma yang digunakan. Untuk melakukan proses dengan bersamaan dapat dilakukan penggabungan sebagaimana digambarkan pada gambar 5.

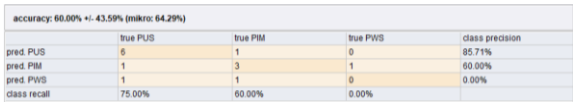


Gambar 5. Proses Perhitungan

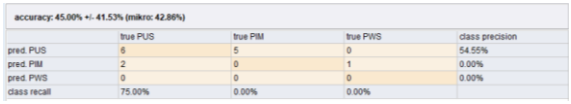
Proses tersebut berjalan dan menghasilkan *confussion matrix*. Matrix ini merupakan proporsi atau prosentasi akurasi dari setiap perhitungan model seluruh algoritma yang digunakan. Gambar 6 merupakan hasil *confussion matrix* dari algoritma KNN. Dilanjutkan gambar 7 untuk Bayes, gambar 8 untuk Neural Network dan gambar 9 untuk decision tree.



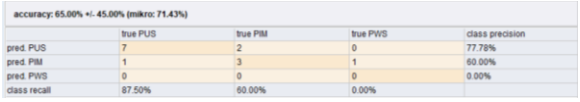
Gambar 6. Hasil akurasi KNN



Gambar 7. Hasil akurasi bayes



Gambar 8. Hasil akurasi neural network



Gambar 9. Hasil akurasi decision tree

Tingkat akurasi tertinggi diperoleh decision tree dengan 65%. Sedangkan tingkat akurasi terendah didapatkan neural network dengan hanya 45%

3.3. Pembahasan

Nilai akurasi dari keseluruhan algoritma dalam klasifikasi data covid-19 survillance masih tergolong rendah.

Ini karena di dalam dataset Covid-19 survillance hanya memiliki 14 *record*. Beberapa algoritma sangat peka terhadap data. Banyaknya *record* dapat mempengaruhi proses pembelajaran pada algoritma K-NN yang hasilnya juga dapat mempengaruhi tingkat akurasi algoritma K-NN (Indrayanti et al., 2017).

Selain sedikitnya *record* yang ada, rendahnya tingkat akurasi juga dipengaruhi oleh kurangnya varian dari atribut label. Pada dataset Covid-19 survillance hanya terdapat 3 varian label dengan rincian: 8 *record* (PUS), 5 *record* (PIM), serta 1 *record* (PWS). Ketimpangan ini jelas mempengaruhi tingkat akurasi karena beberapa label jauh lebih dominan dibandingkan atribut label yang lainnya.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Hasil komparasi dari ke empat metode terpapar dalam tabel 2 berikut:

Tabel 2. Dataset Covid-19 survillance

Algoritma	Akurasi	mikro
KNN	55%	57,14%
Neural Network	45%	42,86%
Bayes	60%	64,29%
Decission Tree	65%	71,43%

Dari pembuktian yang ada pada tabel 2 diatas dapat diketahui bahwa decision tree merupakan metode terbaik untuk klasifikasi dataset covid-19 survillance.

4.2. Saran

Dataset dalam penelitian ini tergolong kecil dengan hanya 14 *record*. Penelitian berikutnya dapat menggunakan dataset dengan jumlah *record* yang lebih banyak.

DAFTAR PUSTAKA

Alkaromi, M. A. (2014). Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner. *IC Tech*, *XI*(2).

Gorunescu, F. (2011). *Data Mining: Concepts; Models and Techniques*. Springer.

Indrayanti, I., Devi, S., & Al Karomi, M. A. (2017). Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus. *IC-TECH*, *XIII*(2), 1–6.

Ivandari, & Al Karomi, M. A. (2021a). Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance. *IC Tech*, *16*(1), 12–15. https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137

Ivandari, & Al Karomi, M. A. (2021b). Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm. *Jaict*, *6*(1), 44–49. https://jurnal.polines.ac.id/index.php/jaict/article/view/2896

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition* (Vol. 40, Issue 6). Elsevier. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C