Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023105935
Vol. 10, No. 1, Februari 2023, hlm. 115-122 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

ANALISIS PERFORMA ALGORITMA DECISION TREE, NAÏVE BAYES , K-NEAREST NEIGHBOR UNTUK KLASIFIKASI ZONA DAERAH RISIKO COVID-19 DI INDONESIA

Ainurrohmah[*1], Dian Tri Wiyanti[2]

[1,2] Universitas Negeri Semarang, Semarang
Email: [1]ainu.rohmah54@gmail.com, [2]diantriwiyanti@mail.unnes.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 06 Desember 2021, diterima untuk diterbitkan: 27 Februari 2023)

Abstrak

Pandemi Covid-19 terjadi di Indonesia. Pemerintah berupaya melakukan penanganan Covid-19, salah satunya dengan pembuatan peta risiko Covid-19. Peta risiko Covid-19 membagi zona berdasarkan Kabupaten/Kota. Zona risiko Covid-19 menjadi patokan pemerintah dalam mengambil kebijakan setiap daerah. Pemerintah menggunakan pembobotan dari 15 indikator untuk menentukan zona. Beberapa kali perubahan zona risiko Covid-19 pada *website* mengalami keterlambatan. Klasifikasi dapat menjadi alternatif penentuan zona risiko Covid-19, sehingga perubahan zona dapat dilakukan secara cepat dan efisien. Klasifikasi memiliki berbagai algoritma, setiap algoritma memiliki keunggulan dan kelemahan. Algoritma klasifikasi yang memiliki akurasi yang baik dengan waktu relatif cepat yaitu *Decision Tree*, *Naïve Bayes* dan *K-Nearest Neighbor.* Tujuan penelitian ini menghitung performa setiap algoritma, mendapatkan algoritma terbaik dan mendapatkan pola klasifikasi dari algoritma terbaik. Metode penelitian menggunakan 10-*fold cross validation* untuk pembagian data dan *confusion matrix* untuk menilai performa. *Software* yang digunakan yaitu Rapidminer. Hasil dari pengolahan data menunjukan semua algoritma mempunyai nilai performa yang baik yaitu diatas 70%. Semua algoritma tidak memerlukan waktu yang lama dalam pembuatan model. Nilai performa terbaik didapatkan dengan menggunakan algoritma *decision tree*. Pola klasifikasi dari algoritma terbaik menghasilkan 20 aturan yang membagi 3 zona klasifikasi yaitu rendah, sedang, dan tinggi. Atribut yang berpengaruh dalam klasifikasi zona risiko Covid-19 yaitu aktif, CR, CFR, laju insidensi, positif, dan meninggal.

Kata kunci: *Klasifikasi, Zona Risiko, Covid-19, Performa Algoritma*

PERFORMANCE ANALYSIS OF DECISION TREE, NAÏVE BAYES, K-NEAREST NEIGHBOR ALGORITHM FOR COVID-19 RISK ZONE CLASSIFICATION IN INDONESIA

Abstract

The Covid-19 pandemic occurred in Indonesia. The government is trying to handle Covid-19, one of which is by making a Covid-19 risk map. The Covid-19 risk map divides zones based on Regency/City. The Covid-19 risk zone is the government's benchmark policy for each region. The government uses a weighting of 15 indicators to determine the zone. Several times the Covid-19 risk zone change on the website has been delayed. Classification can be an alternative to determining the Covid-19 risk zone, that zone changes can be quickly and efficiently. Many algorithms can be used for classification. Several classification algorithms have good accuracy with relatively fast time are Decision Tree, K-Nearest Neighbor, and Naïve Bayes. The purpose of this study is to calculate the performance of each algorithm, get the best algorithm, and get the classification pattern from the best algorithm. The research method uses 10-fold cross validation for data sharing and confusion matrix to assess performance. The software used is Rapidminer. The results show that all algorithms have good performance values, which are above 70%. All algorithms do not require a long time in modeling. The best performance value using a Decision Tree algorithm. The classification pattern of the best algorithm produces 20 rules that divide 3 classification zones, namely low, medium, and high. Attributes that influence the classification of the Covid-19 risk zone are active, CR, CFR, incidence rate, positive, and death.

Keywords: *Classification, Risk Zone, Covid-19, Algorithm Performance*

1. PENDAHULUAN

Covid-19 menjadi pandemi yang dirasakan seluruh dunia (Ikbal, Andryana and Sari, 2021). Kebijakan pemerintah dalam penanganan Covid-19 yaitu pembuatan peta zonasi risiko Covid-19 yaitu peta zonasi daerah tingkat kabupaten/kota. Peta zona risiko Covid-19 dibagi menjadi empat kategori zona. Zona risiko Covid-19 mempengaruhi kebijakan penanganan Covid-19 di setiap daerah, seperti kebijakan PSBB atau *new normal* (Syarifudin, 2020). Pembagian zona yang penting dalam pengambilan keputusan dan klasifikasi dapat menjadi alternatif dalam penentuan zona risiko Covid-19. Melihat hal tersebut peneliti tertarik melakukan klasifikasi zona risiko Covid-19 di Indonesia.

Penelitian terdahulu yang menggunakan metode klasifikasi untuk membagi risiko Covid-19 adalah penelitian dari Qurrohman et al., (2020) dengan menggunakan algoritma klasifikasi yaitu C4.5. Data yang dipakai adalah jumlah pasien positif, ODP, PDP, jumlah pasien dirawat, jumlah pasien sembuh, dan jumlah pasien meninggal yang berasal dari *website* resmi pemerintah Jakarta. Status Covid-19 setiap kelurahan dibagi menjadi tiga label yaitu merah, hijau, dan kuning. Penelitian lain dari Bird et al., (2020) yang berjudul "*Country-Level Pandemic Risk and Preparedness Classification Based on COVID- 19 Data: A Machine Learning Approach*" yang menilai risiko kesigapan pandemi tingkat negara. Dengan membagi empat kelas yaitu rendah, sedang, menengah dan tinggi menggunakan data tingkat penularan, tingkat kematian dan tingkat ketidakmampuan dalam tes.

Penelitian ini memiliki perbedaan dari penelitian terdahulu yaitu menggunakan data risiko Covid-19 setiap kabupaten/kota di Indonesia, menggunakan data publik yaitu positif, aktif, meninggal, sembuh, CR, CFR, dan laju insidensi. Pembaharuan lain yaitu penelitian ini hanya menggunakan 3 label zona yaitu zona rendah, zona sedang, dan zona tinggi. Data publik diambil dari *website* resmi satgas Covid-19 Indonesia baik *website* satgas pusat maupun dari daerah, dengan waktu pengambilan data 5 Juli 2020 sampai 8 Agustus 2021.

Klasifikasi merupakan salah satu teknik menemukan pola yang mampu memisahkan kelas data untuk masuk sesuai dengan kategori dan atribut dari suatu kelompok (Romli and Zy, 2020). Klasifikasi mempunyai dua tugas utama yaitu membangun model sebagai prototipe, dan menggunakan model yang dihasilkan untuk memprediksi objek data lain di kelas mana objek data atau model berada (Lishania, Goejantoro and Nasution, 2019). Klasifikasi memiliki berbagai algoritma yang dapat digunakan (Azis, Tangguh Admojo and Susanti, 2020). Di antara algoritma klasifikasi yang paling populer adalah *Decision Trees, Naïve Bayes, K-Nearest Neighbor* (Novianti, 2019).

Penelitian terdahulu yang membahas perbandingan algoritma klasifikasi yang digunakan, salah satunya adalah Ünal and Dudak, (2020) yang bertujuan untuk mengolah data publik Covid-19 Meksiko dengan menggunakan machine learning dan algoritma klasifikasi Decision Tree, Random Forest, Naïve Bayes , SVM, dan KNN. Penelitian lain dari Muhammad et al., (2020) yaitu bertujuan memprediksi kesembuhan pasien Covid-19 menggunakan klasifikasi dengan algoritma klasifikasi yang dipakai adalah Decision Tree, SVM, KNN, Naïve Bayes, logistic regression dan random forest.

Pemilihan algoritma dalam penelitian ini, mempertimbangkan algoritma yang sering digunakan pada klasifikasi risiko Covid-19 serta penggunaan algoritma yang cepat dan memiliki akurasi yang cukup baik. Sehingga algoritma klasifikasi yang digunakan dan dibandingkan pada penelitian ini adalah *Decision Tree*, *Naïve Bayes* dan *K-Nearest Neighbor*.

Decision Tree merupakan salah satu algoritma klasifikasi yang paling umum digunakan dalam pengambilan keputusan (Lishania, Goejantoro and Nasution, 2019). Decision Tree adalah algoritma pohon keputusan yang mengelompokkan sampel data yang ada dengan mengurutkannya berdasarkan nilai fitur (Osisanwo et al., 2017).

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi yang mempunyai kompleksitas rendah dan dapat dilakukan tolak ukur dengan cepat (Bird et al., 2020). Algoritma K-Nearest Neighbor sangat sederhana, menentukan k tetangga terdekat berdasarkan jarak minimum dari sampel query ke sampel pelatihan (Selvakumar et al., 2017). Umumnya rumus jarak Euclidean digunakan untuk menentukan jarak antara dua objek pelatihan dan pengujian.

Naïve Bayes merupakan salah satu algoritma klasifikasi yang proses pengerjaannya sangat cepat dan mudah (Saputra, Widiyaningtyas and Wibawa, 2018). Algoritma Naïve Bayes menjadi salah satu algoritma pembelajaran yang diawasi dan metode statistik klasifikasi, algoritma pembelajaran menghasilkan fungsi untuk memprediksi nilai keluaran yaitu sistem menyediakan target nilai masukan baru setelahnya data pelatihan (Nathiya & Suseendran, 2018). Naïve Bayes adalah pengklasifikasi berdasarkan Teorema Bayes (Makhtar et al., 2017).

Penggunaan klasifikasi membutuhkan pembagian data latih dan data uji, pembagian data latih dan data uji sangat mempengaruhi hasil klasifikasi. *K-fold cross validation* secara acak akan membagi data latih dan data uji (Almoammar, Alhenaki and Kurdi, 2019). Performa suatu algoritma akan sangat mempengaruhi hasil klasifikasi. Performa dari klasifikasi dapat dihitung dengan mencari nilai akurasi, presisi, dan *recall* (Ul Hassan,

Khan and Shah, 2018). Waktu pembuatan model juga menjadi indikator dalam melihat performa algoritma.

Confusion matrix akan membantu menghitung performa algoritma (Krstinić et al., 2020). *Software* yang dipakai yaitu Rapidminer, *software* data mining yang bersifat *open source*, mempunyai bahasa pemrograman java dan bersifat *multiplatform* (Altalhi et al., 2017). Penelitian ini pembagian data latih dan data uji dilakukan dengan *K-fold cross validation* serta perhitungan performa algoritma dilakukan dengan *confusion matrix* dan penelitian menggunakan bantuan *software* Rapidminer.

Tujuan dalam penelitian ini adalah mengetahui performa algoritma *Decision Tree*, *Naïve Bayes* dan *K-Nearest Neighbor* menggunakan *software* Rapidminer. Dari nilai performa yang ada tentunya akan didapatkan nilai performa yang terbaik sehingga akan menemukan satu algoritma terbaik untuk mengklasifikasi zona daerah risiko Covid-19. Algoritma terbaik ini akan menghasilkan pola klasifikasi terbaik dalam klasifikasi zona risiko Covid-19.

2. METODE PENELITIAN

Pada penelitian ini akan mencari performa algoritma klasifikasi dalam klasifikasi zona daerah risiko Covid-19. Metode yang dipakai adalah klasifikasi dengan menggunakan algoritma *Decision Tree*, *Naïve Bayes* dan *K-Nearest Neighbor* dengan *software* Rapidminer. Langkah pertama dan paling utama dalam penelitian ini adalah mengumpulkan data dan memproses data.

2.1 Pengumpulan Data

Data diambil dari *website* resmi pemerintah pusat untuk label dan *website* pemerintah Kabupaten/Kota untuk data lainnya. Pengambilan sampel data menggunakan *purposive sampling,* yaitu pemilihan sampel berdasarkan kriteria tertentu (Nahda and Harjito, 2021). Data yang tidak sesuai kriteria tidak masuk ke dalam data penelitian. *Website* pemerintah diambil beberapa data yaitu, zona risiko, data jumlah orang terkonfirmasi Covid-19, jumlah pasien meninggal, jumlah pasien sembuh, dan jumlah pasien yang masih aktif. Data lain yang menunjang 15 indikator seperti CR (*cure rate*), CFR (*case fatality rate*), dan laju insidensi . Persamaan (1) yaitu CR atau tingkat kesembuhan suatu daerah. Persamaan (2) merupakan rumus CFR atau tingkat kematian suatu daerah karena Covid-19. Dan persamaan (3) merupakan laju insidensi yaitu jumlah orang terkena Covid-19 setiap 100.000 orang.

$$CR = \frac{jumlah\ kesembuhan}{jumlah\ kasus} \text{ x } 100 \quad (1)$$

(Diao et al., 2020),

$$CFR = \frac{jumlah\ kematian}{jumlah\ kasus} \text{ x } 100 \quad (2)$$

$$laju\ insidensi = \frac{Jumlah\ pasien\ positif}{jumlah\ populasi} \text{ x } 100.000 \quad (3)$$

(Kazemi-Karyani et al., 2020).

Waktu pengambilan data dilakukan dari tanggal 5 Juli 2020 sampai 8 Agustus 2021. Pengumpulan data menghasilkan data penelitian berjumlah 21.190 *record* data sekunder dengan 7 atribut (aktif, positif, meninggal, sembuh, CFR, CR, dan laju insidensi), 1 atribut kelas yaitu (hijau, rendah sedang, tinggi) dan 3 ID (tanggal, provinsi dan kabupaten/kota). Pembagian dari 21.190 data menjadi beberapa kelas yaitu pada Tabel 1. Setelah pengumpulan data dilanjutkan dengan input data, proses selanjutnya adalah *preprocessing data.* Pada tahap *preprocessing* data, 3 ID yaitu tanggal, provinsi dan kabupaten/kota dihapuskan sehingga data hanya tersisa 8 atribut. Keterangan 8 atribut yang terpakai ada pada Tabel 2.

Tabel 1. Jumlah Data Setiap Kelas

kelas data	jumlah data
zona rendah	7.166
zona sedang	11.250
zona tinggi	2.774

Tabel 2. Keterangan atribut

nama atribut	Keterangan
Aktif	jumlah pasien terkonfirmasi covid-19 yang masih menjalani pengobatan dan isolasi mandiri
Positif	jumlah pasien keseluruhan yang terkonfirmasi covid-19
Meninggal	jumlah pasien terkonfirmasi covid-19 dan sudah meninggal
Sembuh	jumlah pasien terkonfirmasi covid-19 dan sudah dinyatakan sembuh
Cfr	tingkat kematian suatu daerah yaitu dengan cara jumlah pasien meninggal dibagi jumlah pasien keseluruhan dikali 100
Cr	tingkat kesembuhan suatu daerah yaitu dengan cara jumlah pasien sembuh dibagi jumlah pasien keseluruhan dikali 100
laju insidensi	tingkat laju penyebaran virus yaitu dengan cara jumlah pasien positif dibagi jumlah populasi daerah dikali 100.000
risiko penyebaran	pembagian zona untuk melihat risiko covi-19 pada setiap daerah

2.2 Pemrosesan Data

Proses selanjutnya yaitu pemrosesan data. Pemrosesan data diawali dengan memilih algoritma. Setiap algoritma akan memproses data menggunakan 10-*fold cross validation* untuk membagi data latih dan data uji secara acak, sehingga perbandingan data uji dan data latih yaitu 9:1. Setelah proses pembagian data uji dan data latih, setiap algoritma akan memproses data hingga selesai sampai menghasilkan *output*.

Output yang dihasilkan berisi *confusion matrix,* waktu pembuatan model, pola klasifikasi dan sebagainnya. Hasil *output* setiap algoritma akan dibandingkan, dan jika dari hasil keseluruhan

algoritma belum didapatkan hasil yang sesuai maka proses pengolahan data akan berulang.

Pada pemrosesan hasil performa pada penelitian ini menggunakan akurasi, presisi, recall akan dihitung secara manual. Jumlah label pada data ada tiga sehingga *confusion matrix* yang digunakan pada penelitian ini adalah *confusion matrix multiclass.* Persamaan (4), (5) dan (6) merupakan rumus perhitungan akurasi, presisi dan recall untuk *confusion matrix multiclass.*

$$average\ accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{TP}_i+\mathrm{TN}_i}{\mathrm{TP}_i+\mathrm{TN}_i+\mathrm{FP}_i+\mathrm{FN}_i}\right) \quad (4)$$

(Truica and Leordeanu, 2017)

$$Weighted\ Precision = \frac{\sum_{i=1}^{N} Precision_i \text{ x } \mathrm{Number}_i}{\sum_{i=1}^{N} \mathrm{Number}_i} \quad (5)$$

$$Weighted\ Recall = \frac{\sum_{i=1}^{N} Recall_i \text{ x } \mathrm{Number}_i}{\sum_{i=1}^{N} \mathrm{Number}_i} \quad (6)$$

$$Precision_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i} \quad (7)$$

$$Recall_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i} \quad (8)$$

(Fu et al., 2021)

Keterangan :
N : jumlah *class*/jumlah label
i : kelas ke-i
Number_i : Jumlah data pada kelas ke –i

Perhitungan dari persamaan (7) akan dipakai untuk persamaan (5) begitupun pada persamaan (8) akan dipakai untuk mendukung perhitungan pada persamaan (6). Perhitungan manual dari persamaan (4), (5), dan (6) merupakan hasil akhir pemrosesan data sehingga secara keseluruhan akan didapatkan nilai akurasi, presisi, *recall,* waktu dan pola klasifikasi.

2.3 Analisis Hasil

Tahap analisis hasil dimaksudkan untuk menghasilkan tujuan dari penelitian ini. Analisis hasil dilakukan dengan melihat hasil dari pemrosesan data. Analisis ini akan melihat dari indikator nilai performa dalam melihat keseluruhan performa.

Indikator nilai performa algoritma yang dipakai yaitu akurasi, presisi, *recall*, dan waktu. Setiap indikator ini memiliki nilai tujuan yang berbeda. Akurasi bertujuan melihat keefektifan, presisi bertujuan mengukur hasil label prediksi yang sesuai dengan label data, dan *recall* melihat sensitivitas saat identifikasi klasifikasi (Asdaghi and Soleimani, 2019).

Waktu bertujuan untuk melihat seberapa lama waktu yang digunakan algoritma dalam membuat model yaitu dari algoritma dimulai sampai *output* keluar (Fibrianda and Bhawiyuga, 2018). Jika algoritma memerlukan waktu lebih lama tetapi nilai indikator yang lain seperti akurasi, presisi, *recall* untuk performa lebih besar, maka algoritma itu yang terpilih (Baliarsingh et al., 2019). Waktu bukan menjadi acuan utama dalam menentukan performa terbaik, tetapi waktu masih menjadi indikator untuk melihat seberapa lama proses pengolahan data berjalan.

Analisis hasil berupa akurasi, presisi, *recall* dan waktu akan menghasilkan performa setiap algoritma. Semua performa algoritma yang ada akan didapatkan nilai performa terbaik. Algoritma dengan nilai performa terbaik ini yang nantinya akan dilihat pola klasifikasi.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Penelitian

3.1.1 Hasil Pengambilan Data

Penelitian ini memerlukan beberapa data yang terkait dengan klasifikasi zona daerah risiko Covid-19 di Indonesia. Data – data yang digunakan adalah data perkembangan Covid-19 yang diambil dari *website* satgas Covid-19 pemerintah provinsi maupun daerah yang bersifat publik. Data perkembangan Covid-19 yang diambil untuk penelitian ini adalah jumlah positif, jumlah meninggal, jumlah sembuh, dan jumlah pasien aktif. Pada penelitian ini data zona dari peta risiko diperlukan sebagai label/kelas. Data yang diperoleh kemudian dikumpulkan pada *excel.*

Setelah mengumpulkan data, dilakukan pemilihan data atau *preprocessing* awal untuk meminimalkan kesalahan dan mengoptimalkan hasil dari proses klasifikasi yang digunakan. Yang pertama dilakukan adalah penanganan *missing value* yaitu jika suatu kabupaten/kota tidak memiliki data yang lengkap untuk setiap atribut, maka data kabupaten/kota tersebut tidak dimasukkan kedalam data penelitian. Selain itu, data yang memiliki label hijau dengan semua data pada atribut bernilai 0 akan dihapuskan.

Pengumpulan data menghasilkan data penelitian berjumlah 21.190 *record* data dengan 7 atribut (aktif, positif, meninggal, sembuh, CFR, CR, laju insidensi), 1 atribut kelas yaitu (rendah, sedang, tinggi) dan 3 ID (tanggal, provinsi dan kabupaten/kota). Jumlah pembagian data yaitu 2.774 data zona tinggi, 11.250 data zona sedang, dan 7.166 data zona rendah. Data yang dikumpulkan untuk penelitian ada pada Tabel 3.

Tabel 3. Data Penelitian

tanggal	05/07/2020	05/07/2020	...	08/08/2021
provinsi	maluku	sumut	…	Yogyakarta
kabupaten /kota	ambon	Asahan	…	Yogyakarta
risiko	sedang	rendah	…	Tinggi
penyebaran				
positif	577	21	…	21.315
meninggal	13	4	…	148
sembuh	272	12	…	3.014
aktif	292	5	…	18.153
CFR	2,25	19,05	…	0,69
CR	47,14	57,14	…	14,14
laju insidensi	166,14	2,73	…	5.705,47

3.2 Hasil Pengolahan Data

Pengolahan data menggunakan *software* yaitu Rapidminer dengan algoritma klasifikasi yang digunakan yaitu *Decision Tree, Naïve Bayes* dan *K-Nearest Neighbor* menghasilkan *output. Output* yang dihasilkan berupa waktu, *confusion matrix* serta pola klasifikasi. *Confusion matrix* yang ada dapat digunakan untuk menghitung nilai akurasi, presisi dan *recall.* Salah satu contoh tabel *confusion matrix* yang didapatkan ada pada Tabel 4.

Salah satu contoh perhitungan nilai akurasi, *recall* dan presisi secara manual yaitu dengan melihat dari Tabel 4 sebagai berikut:

$$average\ accuracy = \frac{1}{3}\left(\left(\frac{6.890+8.547+904}{21.190}\right) + \left(\frac{8.547+904+6.890}{21.190}\right) + \left(\frac{904+6.890+8.547}{21.190}\right)\right)$$
$$= \frac{1}{3}\left(\left(\frac{16.341}{21.190} + \frac{16.341}{21.190} + \frac{16.341}{21.190}\right)\right) = 0{,}7712$$

$$\text{Number}_1 = 7.166, \text{Number}_2 = 11.250,\ \text{Number}_3 = 2.774$$
$$Precision_1 = \frac{6.890}{6.890+2.193+26} = \frac{6.890}{9.109} = 0{,}756$$
$$Precision_2 = \frac{8.547}{8.547+273+1.844} = \frac{8.547}{10.664} = 0{,}801$$
$$Precision_3 = \frac{904}{904+3+510} = \frac{904}{1.417} = 0{,}638$$

$$Weighted\ Precision = \frac{\sum_{i=1}^{N} Precision_i \text{ x Number}_i}{\sum_{i=1}^{N} \text{Number}_i}$$
$$= \frac{(0{,}756 \text{ x } 7.166) + (0{,}801 \text{ x } 11.250) + (0{,}638 \text{ x } 2.774)}{7.166 + 11.250 + 2.774}$$
$$= \frac{16.207{,}04}{21.190} = 0{,}7648$$

$$Recall_1 = \frac{\text{TP}_1}{\text{TP}_1+\text{FN}_1} = \frac{6.890}{6.890+3+273} = \frac{6.890}{7.166} = 0{,}9615$$
$$Recall_2 = \frac{8.547}{8.547+2.193+510} = \frac{8.547}{11.250} = 0{,}7597$$
$$Recall_3 = \frac{904}{904+1.844+26} = \frac{904}{2.774} = 0{,}3259$$
$$Weighted\ Recall = \frac{\sum_{i=1}^{N} Recall_i \text{ x Number}_i}{\sum_{i=1}^{N} \text{Number}_i}$$
$$= \frac{(0{,}9615 \text{ x } 7.166) + (0{,}7597 \text{ x } 11.250) + (0{,}3259 \text{ x } 2.774)}{7.166 + 11.250 + 2.774}$$
$$= \frac{16.340{,}7806}{21.190} = 0{,}7712$$

Evaluasi performa algoritma *Naïve Bayes* pada *software* Rapidminer didapatkan nilai performa yaitu akurasi 77,12%, presisi 76,48%, *recall* 77,12%, dengan waktu pembuatan model 2 detik.

3.2.1 *Output* Algoritma *Naïve Bayes*

Algoritma *Naïve Bayes* pada *software* Rapidminer membutuhkan waktu untuk pembuatan model selama 2 detik. *Output confusion matrix* pada algoritma *Naïve Bayes* dengan *software* Rapidminer bisa di lihat pada Tabel 4.

Tabel 4 *Confusion Matrix Naïve Bayes* Rapidminer

	True **Rendah**	***True*** **Sedang**	***True*** **Tinggi**
Pred Rendah	6.890	2.193	26
Pred Sedang	273	8.574	1.844
Pred Tinggi	3	510	904

Pada Tabel 4 mengenai *confusion matrix* dari data yang berjumlah 21.190 data dengan 7.166 kelas rendah, 11.250 kelas sedang, dan 2.774 kelas tinggi. Jumlah data yang sesuai dengan prediksi setelah diklasifikasikan dengan algoritma *Naïve Bayes* ada 6.890 data kelas rendah, 8.574 data kelas sedang, dan 904 data kelas tinggi.

Klasifikasi zona risiko daerah Covid-19 di Indonesia menggunakan *Naïve Bayes* dengan *software* Rapidminer mendapatkan nilai performa yaitu akurasi 77,12%, presisi 76,48%, *recall* 77,12%, dengan waktu pembuatan model 2 detik.

3.2.2 *Output* Algoritma *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* pada *software* Rapidminer membutuhkan waktu untuk pembuatan model selama 53 detik. *Output confusion matrix* pada algoritma *K-Nearest Neighbor* dengan *software* Rapidminer bisa di lihat pada Tabel 5.

Tabel 5 *Confusion Matrix K-Nearest Neighbor* Rapidminer

	True **Rendah**	***True*** **Sedang**	***True*** **Tinggi**
Pred Rendah	6.730	511	3
Pred Sedang	435	10.091	1.593
Pred Tinggi	1	648	1.178

Pada Tabel 5 mengenai *confusion matrix K-Nearest Neighbor* dari data yang berjumlah 21.190 data dengan 7.166 kelas rendah, 11.250 kelas sedang, dan 2.774 kelas tinggi. Jumlah data yang sesuai dengan prediksi setelah diklasifikasikan dengan algoritma *K-Nearest Neighbor* ada 6.730 data kelas rendah, 10.091 data kelas sedang, dan 1.178 data kelas tinggi.

K-Nearest Neighbor dengan *software* Rapidminer mendapatkan nilai performa yaitu akurasi 84,94%, presisi 84,06%, *recall* 83,25%, dengan waktu pembuatan model 0,01 detik.

3.2.3 Algoritma *Decision Tree*

Algoritma *Decision Tree* pada *software* Raidminer membutuhkan waktu untuk pembuatan model selama 2 detik. *Output confusion matrix* pada algoritma *Decision Tree* dengan *software* Rapidminer bisa di lihat pada Tabel 6.

Tabel 6 *Confusion Matrix Decision Tree* Rapidminer

	True **Rendah**	***True*** **Sedang**	***True*** **Tinggi**
Pred Rendah	6.848	42	0
Pred Sedang	318	11.204	2.221
Pred Tinggi	0	64	553

Tabel 6 mengenai *confusion matrix Decision Tree* dari data yang berjumlah 21.190 data dengan 7.166 kelas rendah, 11.250 kelas sedang, dan 2.774 kelas tinggi. Jumlah data yang sesuai dengan prediksi setelah diklasifikasikan dengan algoritma *Decision*

Tree ada 6.848 data kelas rendah, 11.204 data kelas sedang, dan 553 data kelas tinggi.

Decision Tree dengan *software* Rapidminer mendapatkan nilai performa yaitu akurasi 87,8%, presisi 87,6%, *recall* 87,8%, dengan waktu pembuatan model 2 detik.

3.3 Pembahasan

Hasil yang didapatkan dari *output* yaitu *confusion matrix. Confusion matrix* yang ada diolah menjadi nilai performa untuk masing – masing algoritma. Nilai performa ini berupa waktu, akurasi, *recall* dan presisi. Secara keseluruhan nilai performa algoritma *Decision Tree, Naïve Bayes dan K-Nearest Neighbor* pada Tabel 7.

Tabel 7 Keseluruhan nilai performa

performa	Algoritma		
	decision tree	***k-nearest neighbor***	***naïve bayes***
waktu	2	53	2
akurasi	87%	85%	77%
presisi	87%	84%	76%
recall	88%	83%	77%

Hasil dari *Output* secara keseluruhan. Nilai performa algoritma *Naïve Bayes* memerlukan waktu 2 detik dalam membuat model klasifikasi. Dengan persentase akurasi sebesar 77,12%, persentase kesesuaian label prediksi dengan label data sebesar 76,48% dan sensitivitas sebesar 77,12%.

Performa algoritma *K-Nearest Neighbor* memerlukan waktu 53 detik dalam membuat model klasifikasi. Persentase akurasi sebesar 84,94%, persentase kesesuaian hasil label prediksi dengan label data sebesar 84,06%, dan sensitivitas sebesar 83,25%.

Algoritma *Decision Tree* memerlukan waktu 2 detik dalam membuat model klasifikasi. Persentase akurasi sebesar 87,8%, persentase kesesuaian hasil dari label prediksi dengan label data sebesar 87,6%, dan persentase sensitivitas sebesar 87,8%.

Penilaian akurasi, presisi dan *recall* terendah yaitu algoritma *Naïve Bayes* dengan nilai performa rata – rata 77%. Sehingga algoritma dengan nilai terendah yaitu *Naïve Bayes*. Dengan Penilaian akurasi, presisi dan *recall* tertinggi yaitu algoritma *Decision Tree* dengan nilai rata – rata performa 87%.

Waktu tercepat dalam penelitian ini yaitu *Decision Tree* dan *Naïve Bayes* yaitu 2 detik. Tetapi, nilai performa selain waktu untuk algoritma *Naïve Bayes* menggunakan tidak terlalu tinggi. Ada algoritma lain yang memiliki nilai akurasi, presisi dan *recall* yang lebih tinggi yaitu algoritma *Decision Tree*. Sehingga algoritma terbaik pada penelitian ini yaitu *Decision Tree*.

Pemilihan pola klasifikasi zona risiko Covid-19 di Indonesia memilih pada pola klasifikasi dengan algoritma terbaik yaitu algoritma *Decision Tree* dengan melihat pohon keputusan yang dihasilkan.

Pola klasifikasi dari algoritma *Decision Tree* menggunakan Rapidminer memiliki 20 aturan. Pola klasifikasi ini membentuk 3 zona klasifikasi yaitu rendah, sedang dan tinggi. Jumlah aturan pada masing – masing zona yaitu rendah 7 aturan, sedang 10 aturan dan tinggi 3 aturan. Atribut atau data yang paling berpengaruh adalah atribut aktif atau jumlah pasien Covid-19 yang masih dirawat. Sedangkan atribut yang tidak berpengaruh yaitu atribut sembuh.

Aturan terpanjang yang dihasilkan terdiri dari 6 atribut dan terpendek terdiri dari 1 atribut. Aturan yang diperoleh merupakan aturan untuk mengetahui zona risiko, setiap aturan akan memiliki perbedaan jumlah atribut sesuai dengan bagian dari pohon keputusan. Tabel 12 merupakan tabel berisi 20 aturan yang ditemukan oleh *Decision Tree* Rapidmininer

Atribut yang berpengaruh dalam klasifikasi zona risiko Covid-19 yaitu aktif, CR, CFR, laju insidensi, positif, meninggal. Sedangkan atribut sembuh tidak terlihat memiliki pengaruh terhadap zona risiko Covid-19. Atribut yang paling berpengaruh dan menjadi *node* pertama yaitu atribut aktif atau jumlah pasien Covid-19 yang dirawat. Dengan demikian atribut aktif menjadi atribut yang berperan penting dalam menentukan zona risiko Covid-19.

Berdasarkan hasil performa yang didapatkan semua algoritma nilai performanya diatas 70%. Semua algoritma tidak memerlukan waktu yang lama dalam pembuatan model. Semua algoritma dengan kedua *software* sudah baik dalam melakukan klasifikasi. Klasifikasi zona menggunakan data mining dapat menjadi alternatif dalam penentuan zona risiko Covid-19. Nilai akurasi yang didapat tidak dapat 100% dikarenakan dari 15 indikator yang ada tidak semua indikator menjadi data publik. Sehingga tidak semua indikator menjadi data dalam penelitian ini.

4. KESIMPULAN

Hasil dari pengklasifikasian zona daerah risiko Covid-19 di Indonesia. Algoritma *Naïve Bayes* sebesar 77% dengan waktu 2 detik, *K-Nearest Neighbor* sebesar 84% dengan waktu 32 detik, dan *Decision Tree* sebesar 87% dengan waktu 2 detik.

Waktu tercepat dalam pembuatan model klasifikasi dengan menggunakan *Decision Tree* dan *Naïve Bayes.* Tetapi nilai performa selain waktu untuk algoritma *Naïve Bayes* tidak terlalu tinggi. Ada algoritma lain yang memiliki nilai akurasi, presisi dan *recall* yang lebih tinggi yaitu algoritma *Decision Tree.* Algoritma terbaik pada penelitian ini adalah algoritma *Decision Tree.*

Pola klasifikasi zona daerah risiko Covid-19 di Indonesia menghasilkan 20 aturan. Aturan ini membagi 3 zona klasifikasi yaitu rendah, sedang, dan tinggi. Dari pola klasifikasi didapatkan atribut yang berpengaruh dalam klasifikasi zona risiko Covid-19 yaitu aktif, CR, CFR, laju insidensi, Positif, meninggal. Sedangkan atribut sembuh tidak terlihat memiliki pengaruh terhadap klasifikasi zona risiko Covid-19.

Tabel 8. Aturan Rapidminer

Aktif	Positif	Meninggal	Aturan CFR	CR	Laju Insidensi	Zona
> 2.021	-	-	-	-	-	Tinggi
542,5 – 2.021	-	-	-	> 93,265%	> 2,595	Sedang
542,5 – 2.021	-	-	> 0,045%	≤ 93,265%	> 2,595	Sedang
542,5 – 2.021	-	-	0,025% – 0,045%	≤ 93,265%	> 2,595	Tinggi
542,5 – 2.021	-	-	≤ 0,025%	≤ 93,265%	> 2,595	Sedang
39,5 – 542,5	-	>94	>17,605%	-	> 5,085	Sedang
39,5 – 542,5	-	≤94	>17,605%	-	> 5,085	Tinggi
39,5 –542,5	-	-	≤ 17,605%	-	> 5,085	Sedang
39,5 – 542,5	-	-	-	-	2,59 – 5,085	Sedang
28,5 – 39,5	> 6.864	-	-	-	> 5,7	Rendah
28,5 – 39,5	≤ 6.864	-	-	-	>1.100	Sedang
28,5 - 39,5	≤ 6.864	-	> 0,045%	-	5,7 –1.100	Sedang
28,5 - 39,5	≤ 6.864	-	≤ 0,045%	-	5,7 –1.100	Rendah
28,5 – 39,5	-	-	-	-	2,59 – 5,7	Rendah
≤ 28,5	>59,5	-	> 0,89%	-	> 29,47	Rendah
≤ 28,5	≤ 59,5	-	> 0,89%	-	> 29,47	Sedang
≤ 28,5	-	-	≤ 0,89%	-	> 29,47	Sedang
≤ 28,5	-	-	-	-	≤ 29,47	Rendah
27,5 - 2.021	-	-	-	-	≤ 2,59	Rendah
≤ 27,5	-	-	-	-	-	Rendah

Keterangan : Tanda (-) yang berisi pada aturan berarti atribut tersebut tidak masuk ke dalam cabang yang diambil pada pohon keputusan.

DAFTAR PUSTAKA

ALMOAMMAR, A., ALHENAKI, L. & KURDI, H., 2019. Selecting Accurate Classifier Models for a MERS-CoV Dataset. *Proceedings of SAI Intelligent Systems Conference*, pp.1070–1084.

ALTALHI, A.H., LUNA, J.M., VALLEJO, M.A. & VENTURA, S., 2017. Evaluation and Comparison of Open Source Software Suites for Data Mining and Knowledge Discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3).

ASDAGHI, F. & SOLEIMANI, A., 2019. An Effective Feature Selection Method for Web Spam Detection. *Knowledge-Based Systems*, 166, pp.198–206.

AZIS, H., TANGGUH ADMOJO, F. & SUSANTI, E., 2020. Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah. *Techno.Com*, 19(3), pp.286–294.

BALIARSINGH, S.K., DING, W., VIPSITA, S. & BAKSHI, S., 2019. A Memetic Algorithm Using Emperor Penguin and Social Engineering Optimization for Medical Data Classification. *Applied Soft Computing Journal*, 85, p.105773.

BIRD, J.J., BARNES, C.M., PREMEBIDA, C., EKÁRT, A. & FARIA, D.R., 2020. Country-Level Pandemic Risk and Preparedness Classification Based on COVID-19 Data: A Machine Learning Approach. *Plos One*, 15(10), p.e0241332.

DIAO, Y., dkk., 2020. Estimating the Cure Rate and Case Fatality Rate of the Ongoing Epidemic COVID-19. *medRxiv*.

FIBRIANDA, M.F. & BHAWIYUGA, A., 2018. Analisis Perbandingan Akurasi Deteksi Serangan Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(9), pp.3112–3123.

FU, L., LIANG, P., LI, X. & YANG, C., 2021. A Machine Learning Based Ensemble Method for Automatic Multiclass Classification of Decisions. In: *International Conference on Evaluation and Assessment in Software Engineering*.

IKBAL, M., ANDRYANA, S. & SARI, R.T.K., 2021. Visualisasi dan Analisa Data Penyebaran Covid-19 dengan Metode Klasifikasi Naïve Bayes. *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, 5(4), pp.389–394.

KAZEMI-KARYANI, A., dkk, 2020. World One-Hundred Days After COVID-19 Outbreak: Incidence, Case Fatality Rate, and Trend. *Journal of Education and Health Promotion*, pp.1–10.

KRSTINIĆ, D., dkk 2020. Multi-label Classifier Performance Evaluation with Confusion Matrix. *Computer Science & Information Technology*, pp.01–14.

LISHANIA, I., GOEJANTORO, R. & NASUTION, Y.N., 2019. Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab

Sjahranie Samarinda. *Jurnal Eksponensial*, 10(2), pp.135–142.

MUHAMMAD, L.J., ISLAM, M.M., USMAN, S.S. & AYON, S.I., 2020. Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery. *SN Computer Science*, 1(4), pp.1–7.

NAHDA, K. & HARJITO, D.A., 2021. Pengaruh Corporate Social Responsibility Terhadap Nilai Perusahaan Dengan Corporate Governance Sebagai Variabel Moderasi. *Jurnal Siasat Bisnis*, 15(1), pp.1–12.

NOVIANTI, D., 2019. Implementasi Algoritma Naïve Bayes pada Data Set Hepatitis Menggunakan Rapid Miner. *Paradigma*, 21(2), pp.143–148.

QURROHMAN, T., HENDERI, WARNARS, H.L.H.S. & MAULANA, M., 2020. Covid-19 Series : Determines the Status Regional Zones of Covid-19 in Jakarta using Decision Tree and C4.5 Algorithm. *Solid State Technology*, 63, pp.1–8.

ROMLI, I. & ZY, A.T., 2020. Penentuan Jadwal Overtime dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 4(2), pp.694–702.

SAPUTRA, M.F.A., WIDIYANINGTYAS, T. & WIBAWA, A.P., 2018. Illiteracy Classification Using K Means-Naïve Bayes Algorithm. *International Journal on Informatics Visualization*, 2(3), pp.153–158.

SYARIFUDIN, 2020. Model Baru Kepemimpinan dan Pengelolaan Nusantara Modal Atasi Bencana, Gangguan dan Sukseskan Pembangunan (Sebuah Gagasan). In: *Covid19 & Disrupsi Tatanan Sosial Budaya, Ekonomi, Politik dan Multi (Catatan Akademisi, Jurnalis, Aktifis dan Diaspora)*. Bandarlampung: Pustaka Media.pp.335–342.

TRUICA, C.-O. & LEORDEANU, C.A., 2017. Classification of an Imbalanced Data Set Using Decision Tree Algorithms. *U.P.B. Sci. Bull*, 79(C).

UL HASSAN, C.A., KHAN, M.S. & SHAH, M.A., 2018. Comparison of machine learning algorithms in data classification. *ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing*, (September), pp.1–6.

ÜNAL, Y. & DUDAK, M.N., 2020. Classification of Covid-19 Dataset with Some Machine Learning Methods. *Journal of Amasya University The Institute of Science and Technology (JAUIST)*, 1, pp.30–37.